1 Introdução

1.1 Importância da Bioestatística

Os indivíduos variam em relação as suas características biológicas, psicológicas e sociais na saúde e na doença. Esta variabilidade gera uma grande quantidade de incertezas.

A Bioestatística, estatística aplicada às ciências biológicas e da saúde, é a ferramenta utilizada pelos pesquisadores para trabalhar com essas incertezas advindas da variabilidade. Várias definições foram escritas para a estatística, uma delas é a seguinte (1):

> *Estatística é a disciplina interessada com o tratamento dos dados numéricos obtidos a partir de grupos de indivíduos*

A Bioestatística lida com a variabilidade humana utilizando técnicas estatísticas quantitativas (2) que ajudam a diminuir a ignorância em relação a esta diversidade. A compreensão da variabilidade humana torna a medicina mais ciência, diminuindo as incertezas, na tentativa de verificar se os resultados encontrados de fato existem ou são apenas obra do acaso.

Na década de 1990, houve um acesso maior aos computadores. Os profissionais da saúde não estatísticos passaram a ter mais interesse no campo da bioestatística. Isto gerou uma onda que facilitou o aparecimento de novas ferramentas estatísticas de ponta. Apesar disso, o conhecimento da Bioestatística permanece restrito aos especialistas na área. Nos últimos anos, os pacotes de softwares foram aprimorados, tornando-se mais amigáveis e diminuindo significativamente o pânico ao se defrontar com uma série de números uma vez que a maioria deles exige apenas conhecimento básico de matemática.

Para a tomada de decisão em saúde é fundamental o acúmulo de conhecimento adquirido através da prática clínica, geradora da experiência do profissional, do intercâmbio com os pares e da análise adequada das evidências científicas publicadas em periódicos de qualidade. Para atingir este objetivo, é fundamental o conhecimento de bioestatística, incluindo aqui que o pensamento que deve nortear os profissionais da saúde ao lidar com o ser humano é o *pensamento probabilístico*.

1.2 Pílulas históricas da Estatística

> A história deve começar em algum lugar, mas a história não tem começo (3)

Entretanto, é natural, que se trace as raízes voltando ao passado, tanto quanto possível. Alguns referem-se à curiosidade em relação ao registro de dados à dinastia Shank, na China, possivelmente no século XIII a.c, com a realização de censos populacionais. Há relatos bíblicos de possíveis censos realizados por Moisés (1491 a.C.) e por Davi (1017 a.C.).

Os romanos e os gregos já realizavam censos por volta do século VIII a IV a.C. Em 578-534 a.C., o imperador *Servo Túlio* mandou realizar um censo de população masculina adulta e suas propriedades que serviu para estabelecer o recrutamento para o exército, para o exercício dos direitos políticos e para o pagamento de impostos. Os romanos fizeram 72 censos entre 555 a.C. e 72 d.C. A punição para quem não respondia, geralmente era a morte! Na Idade Média, na Europa, existem registros de diversos censos: durante o domínio muçulmano, na Península Ibérica, nos séculos VII a XV; no reinado de Carlos Magno (712-814) e ainda o maior registro estatístico feito na época, o *Domesday Book* (**Figura 1-1**), realizado na Inglaterra, por Guilherme I (3) , o Conquistador, onde registravam nascimentos, mortes, batismos e casamentos. Houve, também, recenseamentos nas repúblicas italianas no século XII ao XIII (4).

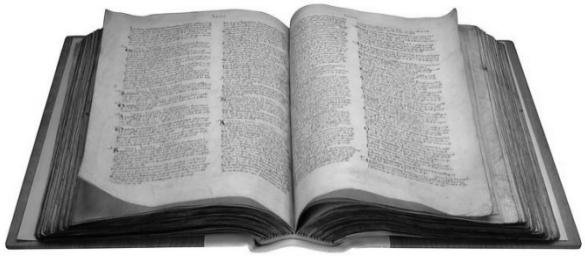

Figura 1-1 Domesday Book

John Graunt (24/04/1620 - 18/04/1674) foi um cientista britânico a quem se deve vários estudos demográficos ingleses. Foi o precursor da construção de Tábuas de Mortalidade. Realizou estudos com William Petty (1623 - 1687), economista britânico que propôs a *aritmética política*.

Em 1791, *Sir John Sinclair* (1754 - 1835) concebeu um plano de uma pesquisa empírica na Escócia para fornecer informações estatísticas. Foi a primeira vez que o termo estatística foi usado em inglês.

Girolamo Cardano (24/09/1501 - 21/09/1576) foi um médico, matemático, físico e filósofo italiano. É tido como o primeiro a introduzir ideias gerais da teoria das equações algébricas e as primeiras regras da probabilidade, descritas no livro *Liber de Ludo Aleae*, publicado em 1663. Descreveu pela primeira vez a clínica da febre tifoide. Foi amigo de Leonardo da Vinci.

Pierre-Simon Laplace, Marquês Laplace (23/03/1749 - 05/03/1927) foi um matemático, astrônomo e físico francês. Embora conduzisse pesquisas substanciais sobre física, outro tema principal dos esforços de sua vida foi a teoria das probabilidades. Em seu *Essai philosophique sur les probabilités*, Laplace projetou um sistema matemático de raciocínio indutivo baseado em probabilidades, que hoje coincidem com as ideias bayesianas.

Antoine Gombaud, conhecido como Chevalier de Méré (1607 - 1684) foi um nobre e jogador. Como não tinha mais sucesso nos jogos de azar, buscou ajuda de Blaise Pascal (19/06/1623 – 19/08/1662), matemático, físico francês, que se correspondeu com Pierre Fermat (matemático e cientista francês), nascendo desta colaboração a teoria matemática das probabilidades (1812). Blaise Pascal foi mais tarde chamado de o Pai da Teoria das Probabilidades.

A moderna teoria das probabilidades foi atribuída a *Abraham De Moivre* (25/05/1667 – 27/11/1754), matemático francês, que adquiriu fama por seus estudos na trigonometria, teoria das probabilidades e pela equação da curva normal. Em 1742, Thomas Bayes (1701 – 07/04/1761, matemático e pastor presbiteriano, inglês,

desenvolveu o Teorema de Bayes que descreve a probabilidade de um evento ocorrer, baseado em um conhecimento *a priori*.

Adrien-Marie Legendre (18/09/1752 - 10/01/1833) foi um matemático francês. Em 1783, tornou-se membro adjunto da *Academie des Sciences*, instituição que esteve na vanguarda dos desenvolvimentos científicos dos séculos XVII e XVIII. Fez importantes contribuições à estatística, à teoria dos números e à álgebra abstrata.

Figura 1-2 Johann Carl F. Gauss

Johann Carl Friedrich Gauss (30/04/1777 - 23/02/1855) foi um matemático, astrônomo e físico alemão (**Figura 1-2**) que contribuiu em diversas áreas das ciências como teoria dos números, estatística, geometria diferencial, eletrostática, astronomia e ótica. Muitos referem-se a ele como o Príncipe da Matemática, o mais notável dos matemáticos. Descobriu o método dos mínimos quadrados e a lei de Gauss da distribuição normal de erros e sua curva em formato de sino, hoje tão familiar para todos que trabalham com estatística.

Lambert Adolphe Jacques Quételet (22/02/1796 - 17/02/1874) foi um astrônomo, matemático, demógrafo e estatístico francês. Seu trabalho se concentrou em estatística social, criando regras de determinação de propensão ao crime

Francis Galton (16/02/1822 – 17/01/1911) foi um antropólogo, matemático e estatístico inglês. Entre muitos artigos e livros, criou o conceito estatístico de correlação e da regressão à média. Ele foi o primeiro a aplicar métodos estatísticos para o estudo das diferenças e herança humanas de inteligência. Criou o conceito de eugenia e afirmava que era possível a melhoria da espécie por seleção artificial. Acreditava que a raça humana poderia ser melhorada caso fossem evitados relacionamentos indesejáveis. Isto acompanhava o pensamento burguês europeu da época. Criou a psicometria, onde desenvolveu testes de inteligência para selecionar homens e mulheres brilhantes. Esta teoria teve papel importante na formação do fascismo e nazismo (5).

William Farr (30/11/1807 - 14/04/1883) foi um médico sanitarista e estatístico inglês, nascido na vila de Kenley, Shropshire. Foi o primeiro investigador a examinar séries temporais de morbimortalidade para longos períodos e, assim, considerado o criador da Estatística da Saúde Pública Moderna. Seus relatórios foram fundamentais para o desencadeamento das reformas sanitárias britânicas, em meados e final do século XIX (6).

Figura 1-3 Florence Nightingale

Florence Nightingale (12/05/1820 – 13/08/1910) foi uma enfermeira (**Figura 1-3**) que ficou famosa por ser pioneira no tratamento de feridos, durante a Guerra da Criméia (7). Ficou conhecida na história pelo apelido de “A dama da lâmpada”, pelo fato de servir-se de uma lamparina para auxiliar no cuidado aos feridos durante a noite. Também contribuiu no campo da Estatística, sendo pioneira na utilização de métodos de representação visual de informações, como por exemplo gráfico de setores (habitualmente conhecido como gráfico do tipo “pizza”).

John Snow (York, 15/03/1813 - Londres, 15/03/1858) foi um médico inglês (**Figura 1-4**), considerado pai da Epidemiologia Moderna. Recebeu, em 1853, o título de Sir após ter anestesiado a rainha Vitória no parto sem dor de seu oitavo filho, Leopoldo

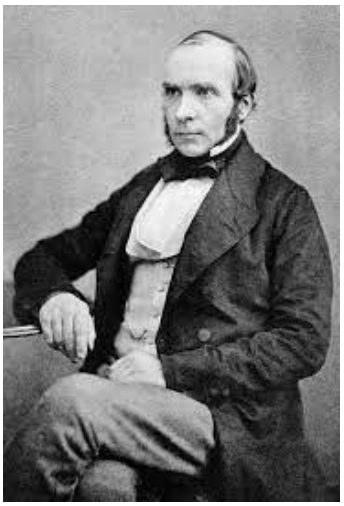

Figura 1-4 John Snow

de Albany. Este fato ajudou a divulgar a técnica entre os médicos da época. Demonstrou que a cólera era causada pelo consumo de águas contaminadas com matérias fecais, ao comprovar que os casos dessa doença se agrupavam em determinados locais da cidade de Londres, em 1854, onde havia fontes dessas águas (6).

Karl Pearson (27/03/1857 - 27/04/1936) foi um importante estatístico inglês, fundador do Departamento de Estatística Aplicada da *University College London* em 1911. Juntamente com Weldon e Galton fundou, em 1901, a revista *Biometrika* com o objetivo era desenvolver as teorias estatísticas, editada até os dias de hoje. O trabalho de Pearson como estatístico fundamentou muitos métodos estatísticos de uso comum, nos dias atuais: regressão linear e o coeficiente de correlação, teste do qui-quadrado de Pearson, classificação das distribuições (8).

Charles Edward Spearman (10/09/1863 - 17/09/1945) foi um psicólogo inglês conhecido pelo seu trabalho na área da estatística, como um pioneiro da análise fatorial e pelo coeficiente de correlação de postos de Spearman. Ele também fez bons trabalhos de modelos da inteligência humana.

William Sealy Gosset (13/07/1876 - 16/10/1937) foi um químico e estatístico inglês (**Figura 1-5**). Em 1907, enquanto trabalhava químico da cervejaria experimental de Arthur Guinness & Son, criou a distribuição t que usou para identificar a melhor variedade de cevada, trabalhando com pequenas amostras. A cervejaria Guinness tinha uma política que proibia que seus empregados publicassem suas descobertas em seu próprio nome. Ele, então, usou o pseudônimo "Student" e o teste é chamado "t de Student" em sua homenagem (9).

Figura 1-6 William Gosset

Ronald Aylmer Fisher (17/02/1890 - 29/07/1962) foi um estatístico, biólogo e geneticista inglês. Em 1919, Fisher se envolveu com pesquisa agrícola no centro de experimentos de *Rothamsted Research*, em Harpenden, Inglaterra, e desenvolveu novas metodologias e teoria no ramo de experimentos (10). Durante sua vida, Fisher (**Figura 1-6**) escreveu 7 livros e publicou cerca de 400 artigos acadêmicos em estatística e genética . Em um dos seus livros, *The design of Experiments* (1935), Fisher relata um experimento que surgiu de uma pergunta curiosa: o gosto do chá muda de acordo com a ordem em que as ervas e o leite são colocados? Essa simples questão resultou em um estudo pioneiro na área e serviu de sustentação para análise da aleatorização de dados experimentais (9). Ronald A. Fisher foi descrito (11) como "um gênio que criou praticamente sozinho os fundamentos para o moderno pensamento estatístico". Era muito temperamental. Seus atritos com outros estatísticos ficaram famosos, entre eles encontra-se ninguém menos do que Karl Pearson, outro notável estatístico.

Figura 1-5 Ronald A. Fisher

*Austin Bradford Hill (*08/07/1897 - 18 /04/1991) foi um epidemiologista e estatístico inglês (**Figura 1-7**), pioneiro no estudo do acaso nos ensaios clínicos e,

juntamente com Richard Doll, foi o primeiro a demonstrar a ligação entre o uso do cigarro e o câncer de pulmão. Hill é amplamente conhecido pelos Critérios de Hill, conjunto de critérios para a determinação de uma associação causal (12).

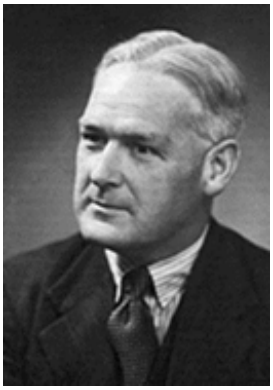

Figura 1-7 Bradford Hill

John Wilder Tukey (16/06/1915 - 26/07/2000) foi um estatístico norte-americano. Desenvolveu uma filosofia para a análise de dados que mudou a maneira de pensar dos estatísticos, sugerindo que se faça uma visualização dos dados, interpretando o formato, centro, dispersão, presença de valores atípicos, sumarizar numericamente e por fim escolher um modelo matemático. Foi o criador do boxplot e introduziu a palavra "bit" como uma contração do termo *binary digit*.

Douglas G. Altman (12 /07/1948 - 03/06/2018) foi um estatístico inglês (**Figura 1-8**), conhecido por seu trabalho em melhorar a confiabilidade dos artigos de pesquisa médica (13) e por artigos altamente citados sobre metodologia estatística. Ele foi professor de estatística em medicina na Universidade de Oxford. Há praticamente 30 anos, Altman (14) escreveu um artigo sobre problema da qualidade da pesquisa em medicina que causou um grande impacto e permanece válido até hoje. Nesta publicação ele afirma:

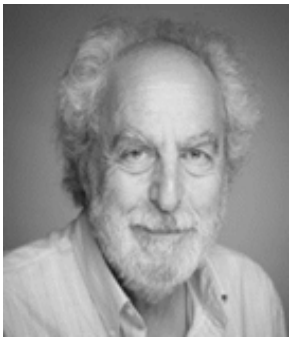

Figura 1-8 Douglas G. Altman

> A má qualidade de muitas pesquisas médicas é amplamente reconhecida, mas, de forma perturbadora, os líderes da profissão médica parecem apenas minimamente preocupados com o problema e não fazem nenhum esforço aparente para encontrar uma solução."

1.3 História resumida do R

O R é uma linguagem e um ambiente de desenvolvimento voltado fundamentalmente para a computação estatística. Foi inspirado em duas linguagens: S (John Chambers, do Bell Labs) que forneceu a sintaxe e Scheme (Hal Abelson e Gerald Sussman) implementou e forneceu a semântica.

O nome R provém em parte das iniciais dos criadores, *George Ross Ihaka* e *Robert Gentleman* (**Figura 1-9**), e também de um jogo figurado com a linguagem S. Em 29 de Fevereiro de 2000, o software foi considerado com funcionalidades e estável o suficiente para a versão 1.0.

Figura 1-9 Robert Gentlemen (E) e George Ross (D)

O R é um projeto GNU[1] Software Livre significa que os usuários têm liberdade para executar, copiar, distribuir, estudar, alterar e melhorar o software. Foi desenvolvido em um esforço colaborativo de pessoas em vários locais do mundo (15).

O projeto R fornece uma grande variedade de técnicas estatísticas e gráficas. É uma linguagem e um ambiente similar ao S. A linguagem do S que também é uma

[1] Esta sigla está associada ao animal gnu africano, símbolo de software de distribuição livre, quer dizer is Not Unix, sigla recursiva muito comum entre nerds!

linguagem de computador voltada para cálculos estatísticos. Um dos pontos fortes de R é a facilidade com que produções gráficas de qualidade podem ser produzidas. O R é também altamente expansível com o uso dos pacotes, que são bibliotecas para sub-rotinas específicas ou áreas de estudo específicas. Um conjunto de pacotes é incluído com a instalação de R e muito outros estão disponíveis na rede de distribuição do R - *Comprehensive R Archive Network* (CRAN) (16).

1.4 Sobre o autor

Petrônio Fagundes de Oliveira Filho nasceu em 04/10/1947, em Porto Alegre, Rio Grande do Sul, Brasil. Estudou no Ensino Médio do Colégio do Rosário, em Porto Alegre. Possui graduação em Medicina pela Universidade de Caxias do Sul (UCS), em 1973, residência em Pediatria no Hospital da Criança Conceição, Porto Alegre (1975). Em 1980, obteve o Título de Especialista em Pediatria (TEP); em 1998, mestre em Saúde Pública Materno Infantil, Universidade de São Paulo e, em 2009, o título de especialista em Estatística Aplicada (UCS).

Trabalhou como Pediatra no INAMPS até 2002 e em consultório privado até hoje. Aposentou-se como professor da Universidade de Caxias do Sul (UCS), em 2019, onde atuou desde 1975, nas áreas de Pediatria, Epidemiologia e Bioestatística, foi coordenador do Serviço e da Residência Médica em Pediatria, chefe de Departamento, coordenador do curso de Medicina e diretor de Ensino do Hospital Geral de Caxias do Sul (Hospital de Ensino da Universidade de Caxias do Sul) e membro de Conselho de Ética em Pesquisa da Universidade de Caxias do Sul, ligado ao CONEP (Conselho Nacional de pesquisa).

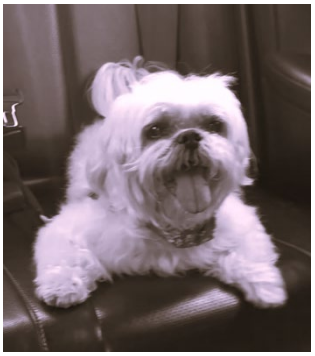

Figura 1-10 Floquinho

Durante mais de 20 anos fez parte do Núcleo de Consultoria e Epidemiologia do Centro de Ciência da Saúde (UCS). É autor de dois livros: Epidemiologia e Bioestatística: Fundamentos para a leitura crítica (Editora Rubio, 2015 e 2018) e SPSS - Análise de Dados Biomédicos, em coautoria com Valter Motta (MedBook, 2009). Além disso, participou de dezenas publicações e de capítulos de outros livros.

Desde 1976, é casado com Lena Maria Cantergiani Fagundes de Oliveira. Tem duas filhas, Nathalia e Andressa, e dois lindos netos, Gabriel e Felix. Ah, tem um cão shitzu branco, marrom claro e com algumas manchas pretas, Floquinho (**Figura 1-10**), que ao acompanhar seus estudos e análises estatísticas, late toda vez que ele menciona o nome de Ronald Fisher.

e-mail: petronioliveira@gmail.com

2 Natureza dos Dados

2.1 Variáveis e Dados

As pesquisas manuseiam dados referentes às variáveis que estão sendo estudadas. *Variável* é toda característica ou condição de interesse que pode de ser mensurada ou observada em cada elemento de uma amostra ou população. Como o próprio nome diz, seus valores são passíveis variar de um indivíduo a outro ou no mesmo indivíduo. Em contraste com a variável, o valor de uma constante é fixo. As variáveis podem ter valores numéricos ou não numéricos. O resultado da mensuração ou observação de uma variável é denominado *dado*.

A **Tabela 2-1** Variáveis e dados. mostra um conjunto de variáveis e suas medidas (dados) de um grupo de pacientes internados em uma determinada UTI. O termo medida deve ser entendido num sentido amplo, pois não é possível "medir" o sexo (observação) ou o estado geral (critérios) de alguém, ao contrário do peso e da pressão arterial que podem ser mensurados com instrumentos.

Tabela 2-1 Variáveis e dados.

Id	Nome	Idade	Sexo	PAS	PAD	Estado Geral
1	45	João	masculino	140	90	bom
2	32	Maria	feminino	110	70	regular
3	27	Pedro	masculino	120	80	grave
4	18	Teresa	feminino	100	60	bom

2.2 População e Amostra

Na pesquisa em saúde, a não ser quando se realiza um censo, coleta-se dados de um subconjunto de indivíduos denominado de amostra, pertencente a um grupo maior, conhecido como população. A população de interesse é, geralmente, chamada de população-alvo. A amostra para ser representativa da população deve ter as mesmas características desta. A partir dos dados encontrados na amostra, presume-se o resultado é condizente com a população. Este processo é denominado de inferência estatística. O interesse na amostra não está propriamente nela, mas na informação que ela fornece ao investigador sobre a população de onde ela provém. A amostra fornece estimativas (estatísticas) da população (**Figura 2-1**).

> **População** ou **população-alvo** consiste em todos os elementos (indivíduos, itens, objetos) cujas características estão sendo estudadas. **Amostra** é a parte, subconjunto, da população selecionada para estudo.

Em decorrência do acaso, diferentes amostras de uma mesma população fornecem resultados diferentes. Este fato deve ser levado em consideração ao usar uma amostra para fazer inferência sobre uma população. Este fenômeno é denominado

de **variação amostral** ou **erro amostral** e é a essência da estatística. O grau de certeza na inferência estatística depende da representatividade da amostra.

O processo de obtenção da amostra é chamado de **amostragem**. Mesmo que este processo seja adequado, a amostra nunca será uma cópia perfeita da população de onde ela foi extraída. Desta forma, em qualquer conclusão baseada em dados de uma amostra, sempre haverá o que é conhecido como erro amostral. Este erro deve ser tratado estatisticamente tendo em mente a teoria da amostragem, baseada em probabilidades.

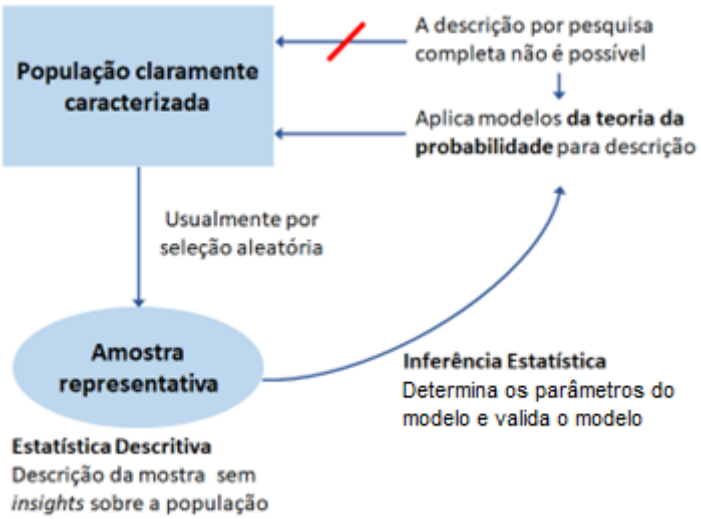

Figura 2-1 População, amostra e inferência estatística

2.3 Estimativas e Parâmetros

Estimativa é uma característica que resume os dados de uma amostra (estatística amostral) e o **parâmetro** é uma característica estabelecida para toda a população. Os valores dos parâmetros são normalmente desconhecidos, porque é inviável medir uma população inteira. A estimativa é um valor aproximado do parâmetro. As estimativas são representadas por letras romanas e os parâmetros por letras gregas. Por exemplo, a média da população é representada por μ e a média da amostra por $\bar{x}$; o desvio padrão da população é denotado σ e o desvio padrão da amostra por s.

Na maioria dos estudos, são utilizadas amostras que fornecem estimativas que, para serem representativas da população, devem ser probabilísticas. Ou seja, a amostra deve ser recrutada de forma aleatória, permitindo que cada um dos membros da população tenha a mesma probabilidade de ser incluído na amostra. Além disso, uma amostra deve ter um tamanho adequado para permitir inferências válidas.

2.4 Escalas de medição

2.4.1 Escala Nominal

As escalas nominais são meramente classificativas, permitindo descrever as variáveis ou designar os sujeitos, sem recurso à quantificação. É o nível mais elementar de representação. São usados nomes, números ou outros símbolos para designar a

variável. Os números, quando usados, representam códigos e como tal não permitem operações matemáticas. As variáveis nominais não podem ser ordenadas. Podem apenas ser comparadas utilizando as relações de igualdade ou de diferença, através de **contagens**. Os números atribuídos às variáveis servem como identificação, ou para associá-la a uma dada categoria. As categorias de uma escala nominal são exaustivas e mutuamente exclusivas. Quando existem duas categorias, a variável é dita **dicotômica** e com três ou mais categorias, **politômicas**.

Os nomes e símbolos que designam as categorias podem ser intercambiáveis sem alterar a informação essencial.

Exemplos: Tipos sanguíneos: A, B, AB, O; variáveis dicotômicas: morto/vivo, homem/mulher, sim/não; cor dos olhos, etc.

2.4.2 Escala Ordinal

As variáveis são medidas em uma escala ordinal quando ocorre uma ordem, crescente ou decrescente, inerente entre as categorias, estabelecida sob determinado critério. A diferença entre as categorias não é necessariamente igual e nem sempre mensuráveis. Geralmente, designam-se os valores de uma escala ordinal em termos de numerais ou postos (*ranks*), sendo estes apenas modos diferentes de expressar o mesmo tipo de dados. Também não faz sentido realizar operações matemática com variáveis ordinais. Pode-se continuar a usar contagem.

Exemplos: classe social (baixa, média, alta); estado geral do paciente: bom, regular, mau; estágios do câncer: 0, 1, 2, 3 e 4; escore de Apgar: 0, 1, 2… 10.

2.4.3 Escala Intervalar

Uma escala intervalar contém todas as características das escalas ordinais com a diferença de que se conhece as distâncias entre quaisquer números. Em outras palavras, existe um espectro ordenado com intervalos quantificáveis. Este tipo de escala permite que se verifique a ordem e a diferença entre as variáveis, porém não tem um zero verdadeiro, o zero é arbitrário.

O exemplo clássico é a mensuração da temperatura, usando as escalas de: Celsius ou Fahrenheit. Aqui é legítimo ordenar, fazer soma ou médias. No entanto, 0ºC não significa ausência de temperatura, portanto a operação divisão não é possível. Uma temperatura de 40ºC não é o dobro de 20ºC. Se 40ºC e 20ºC forem transformados para a escala Fahrenheit, passarão, respectivamente, para 104ºF e 68ºF e, sem dúvida, 104 não é o dobro de 68!

2.4.4 Escala de Razão

Há um espectro ordenado com intervalos quantificáveis como na escala intervalar. Entretanto, as medidas iniciam a partir de um zero verdadeiro e a escala tem intervalos iguais, permitindo as comparações de magnitude entre os valores. Refletem a quantidade real de uma variável, permitindo qualquer operação matemática.

Os dados tanto na escala intervalar como na de razão, podem ser contínuos ou discretos. Dados contínuos necessitam de instrumentos para a sua mensuração e assumem qualquer valor em um certo intervalo. Por exemplo, o tempo para terminar qualquer tarefa pode assumir qualquer valor, 10 min, 20 min, 35 min, etc., de acordo com o tipo de tarefa. Outros exemplos: peso, dosagem de colesterol, glicemia.

Dados discretos possuem valores iguais a números inteiros, não existindo valores intermediários. A mensuração é feita através da contagem. Por exemplo: número de filhos, número de fraturas, número de pessoas.

2.5 Tipos de Variáveis

A primeira etapa na descrição e análise dos dados é classificar as variáveis, pois a apresentação dos dados e os métodos estatísticos variam de acordo com os seus tipos. As variáveis, primariamente, podem ser divididas em dois tipos: numéricas ou quantitativas e categóricas ou qualitativas (19).

2.5.1 Variáveis Numéricas

As variáveis numéricas são classificadas em dois tipos de acordo com a escala de mensuração: continuas e discretas.

As **variáveis contínuas** são aquelas cujos dados foram mensurados em uma escala intervalar ou de razão, podendo assumir, como visto, qualquer valor dentro de um intervalo de números reais, dependendo da precisão do instrumento de medição. O tratamento estatístico tanto para variável intervalar como de a razão é o mesmo. A diferença entre elas está na presença do zero absoluto. As variáveis numéricas contínuas têm unidade de medida. Por exemplo, um menino de 4 anos tem 104 cm.

Uma variável numérica é considerada **discreta** quando é apenas possível quantificar os resultados possíveis através do processo de contagem. Também têm unidade de medida – *número de elementos*. Por exemplo, o número de fraturas, o número de acidentes, etc.

2.5.2 Variáveis Categóricas

As variáveis categóricas ou qualitativas são de dois tipos: nominal e ordinal, de acordo com a escala de mensuração. Um tipo particularmente comum é uma variável binária (ou variável dicotômica), que tem apenas dois valores possíveis. Por exemplo, o sexo é masculino ou feminino. Este tipo de variável é bastante utilizado na área da saúde, em Epidemiologia. As variáveis nominais não têm quaisquer unidades de medida e a nominação das categorias é completamente arbitrária e pertencer a uma categoria não significa ter maior importância do que pertencer à outra. Uma variável ordinal tem uma ordem inerente ou hierarquia entre as categorias. Do mesmo modo que as variáveis nominais, as variáveis ordinais não têm unidades de medida. Entretanto, a ordenação das categorias não é arbitrária. Assim, é possível ordená-las de modo lógico. Um exemplo comum de uma variável categórica ordinal é a classe social, que tem um ordenamento natural da maioria dos mais desfavorecidos para os mais ricos. As escalas, como a escore de Apgar e a escala de coma de Glasgow (20), também são variáveis ordinais. Mesmo que pareçam numéricas, elas apenas mostram uma ordem no estado dos pacientes. O escore de Apgar (21) é uma escala, desenvolvida para a avaliação clínica do recém-nascido imediatamente após o nascimento. Originalmente, a escala foi usada para avaliar a adaptação imediata do recém-nascido à vida extrauterina. A pontuação pode variar de zero a 10. Uma pontuação igual ou maior do que oito, indica um recém-nascido normal. Uma pontuação de sete ou menos pode significar depressão do sistema nervoso e abaixo de quatro, depressão grave.

As variáveis ordinais, da mesma forma que as nominais, não são números reais e não convém aplicar as regras da aritmética básica para estes tipos de dados. Este fato gera uma limitação na análise dos dados.

2.5.3 Como identificar o tipo de variável?

A maneira mais fácil de dizer se os dados são numéricos é verificar se eles têm unidades ligadas a eles, tais como: g, mm, ºC, ml, número de úlceras de pressão, número de mortes e assim por diante. Se não, podem ser ordinais ou nominais – ordinais se os valores podem ser colocados em ordem. A **Figura 2-2** é uma ajuda para o reconhecimento do tipo de variável (22).

Figura 2-2 Caminho para identificar o tipo da variável

2.5.4 Variáveis Dependentes e Independentes

De um modo geral as pesquisas são realizadas para testar as hipóteses dos pesquisadores e, para isso, eles medem variáveis com a finalidade de compará-las. A maioria das hipóteses podem ser expressas por duas variáveis: uma variável explicativa ou preditora e uma variável desfecho (19).

A **variável preditora** ou explanatória é a que se acredita ser a causa e também é conhecida como variável independente, porque o seu valor não depende de outras variáveis. Em Epidemiologia, é com frequência referida como exposição ou fator de risco.

A **variável desfecho** é aquela que é o efeito, consequência ou resultado da ação de outra variável, por isso, também chamada de variável dependente. Em um estudo que tenta verificar se o tabagismo, durante a gestação, pode interferir no peso do recém-nascido, tem o fumo (variável categórica) como variável preditora (exposição ou fator de risco) e o peso do recém-nascido (variável numérica contínua) como variável desfecho.

3 Produção dos Dados

3.1 Processo de Pesquisa

A pesquisa é um processo de construção do conhecimento. O objetivo deste processo é gerar um novo conhecimento e/ou confirmar ou refutar algum conhecimento prévio. A pesquisa é um processo de aprendizagem tanto do pesquisador quanto da sociedade que se beneficiará deste novo conhecimento. Para ser chamada de científica, a pesquisa deve obedecer aos princípios consagrados pela ciência (23).

A pesquisa nasce de uma dúvida do pesquisador, de algum questionamento que ele considerou interessante sobre o mundo, ou seja, de algo que se costuma chamar de pergunta ou questão da pesquisa. Existem vários motivos que geram questões de pesquisa:

- Avaliação crítica de pesquisas realizadas por outros pesquisadores.
- Condução de uma pesquisa primária com a finalidade de responder uma questão (ou questões), gerando um novo conhecimento ou ampliação do conhecimento existente.
- Para obter habilidades de pesquisa ou experiência, com frequência como parte de um programa educacional.
- Testar a viabilidade de um projeto ou técnica de pesquisa.

3.1.1 Questão de Pesquisa

A pesquisa visa estabelecer novos conhecimentos em torno de um tema específico. O tema de pesquisa pode surgir do próprio interesse ou experiência do pesquisador, ou partir da encomenda de alguma instituição financiadora. Algumas vezes, a pesquisa se origina de outros estudos realizados pelo próprio pesquisador ou outros pesquisadores.

À medida que a ideia da pesquisa cresce, o pesquisador estabelece uma pergunta de pesquisa específica ou um conjunto de questões que ele deseja responder. Algumas vezes, o tema da pesquisa é tão amplo que o pesquisador tem que ter cuidado para não se perder do seu objetivo. Este objetivo é que vai guiá-lo no estabelecimento da pergunta ou perguntas a serem respondidas no estudo. Estes questionamentos são conhecidos como **questão de pesquisa** ou **pergunta de partida**.

O foco da questão de pesquisa pode ser na descrição de um fenômeno clínico. Neste caso a pergunta é dita **descritiva**, por exemplo, pesquisa de prevalência de uma enfermidade, proporção de utilização de um serviço de saúde, características de um teste, etc. Quando a pergunta busca a explicação para um fenômeno, ela é dita **analítica**, por exemplo, comparação entre dois fenômenos. Em geral, perguntas analíticas são mais interessantes. Entretanto, as perguntas descritivas são fundamentais no início de um estudo analítico.

Uma boa pergunta de pesquisa deve ter as seguintes características (24):

- **Factível**: o pesquisador deve conhecer desde o início os limites e problemas práticos que podem interferir na pesquisa. A viabilidade está relacionada com o tamanho amostral, com o domínio técnico adequado, com o tempo e custos envolvidos e com um foco dirigido estritamente aos objetivos mais importantes.

- **Interessante**: a questão de pesquisa deve despertar o interesse não apenas do pesquisador, mas também de seus pares e agentes financiadores.
- **Nova**: a pesquisa deve ser inovadora, original, em algum sentido, para que o estudo seja uma contribuição ao conhecimento ou amplie um conhecimento existente;
- **Ética**: se o estudo impõe riscos físicos ou invasão de privacidade ou não traz nenhuma informação nova, o pesquisador deve suspendê-lo. É importante discutir previamente com pesquisadores mais experientes ou com algum representante do Comitê de Ética em Pesquisa da instituição.
- **Relevante**: nenhuma das características da questão de pesquisa é mais importante do que a sua relevância. Para isto basta pensar nos benefícios que os resultados da pesquisa trarão à Medicina atual.

Ou seja, antes de dedicar tempo e esforço para escrever um projeto de pesquisa deve-se avaliar se a questão de pesquisa é FINER (Factível, Interessante, Nova, Ética e Relevante).

3.1.2 Hipótese de Pesquisa

Uma vez estabelecida a(s) pergunta(s) de pesquisa adequada(s), os pesquisadores formulam hipóteses para serem testadas. Enquanto a pergunta de pesquisa possa ser um pouco vaga em sua natureza como: “existe uma relação entre o tipo psicológico e a capacidade de parar de usar drogas?” Uma hipótese de pesquisa, necessita ser precisa. Há necessidade de especificar qual o tipo psicológico está relacionado à habilidade de parar de usar drogas.

A precisão da hipótese é fundamental em um projeto de pesquisa, pois ela determinará o delineamento de pesquisa a ser seguido pelo pesquisador e as técnicas estatísticas apropriadas para a análise dos dados. A fonte e o tipo de dados são determinados pela característica do delineamento recomendado pela hipótese de pesquisa.

O objetivo da pesquisa, usando o método científico, é refutar ou não as hipóteses de pesquisa. Se a hipótese do pesquisador não for rejeitada, houve a geração de um novo conhecimento.

3.2 Processo de Amostragem

Após o estabelecimento das hipóteses a serem testadas, há necessidade de coletar os dados. Uma vez que é praticamente impossível analisar toda a população que constitui a **população-alvo**, extrai-se uma **amostra** desta população. Este processo é denominado de **amostragem** (25).

Uma amostra deve ser representativa da população, ou seja, deve ter características semelhantes às da população e ser fidedigna. A fidedignidade está relacionada à precisão dos dados que sofrem influência dos instrumentos de aferição, questionários não validados e falhas humanas. Uma amostra inadequada ameaça a validade da pesquisa. Os dados coletados de maneira não aleatória são chamados de **evidência anedótica**. O nível de confiança nos resultados de uma pesquisa está diretamente relacionado à qualidade da amostra. A amostra deve ser representativa. Uma amostra deve conter apenas dados úteis que permitam a resposta da pergunta de pesquisa, evitando desperdício e fuga dos objetivos traçados. A aleatoriedade provoca uma diferença entre o resultado da amostra e o verdadeiro valor da população que é denominada **erro amostral**. Não importa quão bem a amostra seja coletada, os erros

amostrais irão sempre ocorrer. Entretanto, não existe técnica estatística que salve amostras coletadas incorretamente, tendenciosas!

3.2.1 Amostras probabilísticas

Para evitar vieses, erros sistemáticos, que favorecem determinados desfechos, o ideal é coletar uma amostra probabilística. A amostra probabilística adota o princípio da **equiprobabilidade**, isto é, "todos os sujeitos da população têm a mesma probabilidade de fazerem parte da amostra". Esta probabilidade é conhecida e diferente de zero. As amostras probabilísticas têm o potencial de ser possível a generalização para a população; ser imparcial e com menor erro amostral.

Amostra aleatória simples: é a mais utilizada pois garante representatividade da amostra junto à população. A amostra aleatória simples não emprega nenhum critério particular para a definição da amostra. O mecanismo mais comum de obter este tipo de amostra é por um simples sorteio, em geral, usando programas de computador.

Amostra aleatória estratificada: quando a população é constituída por subpopulações ou estratos e é razoável supor que a variável de interesse apresenta comportamento diferente nos diferentes estratos, pode-se usar este tipo de amostragem. Neste caso, a amostra deve ter a mesma estratificação da população para ser representativa. Um exemplo comum de estratificação é o nível socioeconômico. A partir do momento que os estratos estão definidos se procede uma amostra aleatória simples de cada estrato.

Amostra aleatória sistemática: as unidades amostrais são selecionadas a partir de um esquema rígido preestabelecido de sistematização que tem o propósito de abranger toda a população-alvo. Para isso, ordena-se os indivíduos da população (por exemplo, um grande arquivo com 20000 fichas) e calcula-se uma constante conveniente, $c=N/n$, onde N é tamanho da população e n é o tamanho da amostra. Se $n=500$, a constante será 40, ou seja, será selecionado aleatoriamente o primeiro membro da amostra (k), de maneira que k seja menor do que a constante e maior do que 1. A partir daí os sucessivos membros serão: $k + c$; $k + 2c$; $k + 3c$; … até atingir n.

Amostra aleatória por conglomerados (*clusters*): este tipo de amostra é utilizada quando dentro da população são identificados agrupamentos (*clusters*) naturais, por exemplo, espaços, vilas, etc. Neste tipo de amostragem o elemento focal não é o sujeito, mas o *cluster*. Identificados estes, sorteiam-se os conglomerados e se analisa todos os indivíduos dos conglomerados sorteados.

3.2.2 Amostras não probabilísticas

Na amostragem não aleatória ou intencionada há uma escolha deliberada da amostra, subordinada a objetivos específicos do pesquisador. Não há garantia de representatividade da população. É importante averiguar, neste tipo de amostragem, a presença de *conflitos de interesse*.

Amostra de conveniência: é uma técnica comum onde é selecionada uma mostra que esteja acessível. Em outras palavras, os indivíduos são recrutados porque eles estão prontamente disponíveis. Neste tipo de amostra há incapacidade de fazer afirmações gerais com rigor estatístico sobre a população.

Amostra por cotas: é uma versão não probabilística da amostra estratificada. Tem três etapas:

1. Segmentação, onde se divide em grupos, por exemplo, sexo, classe social, região, etc.;
2. Definição do tamanho das cotas;
3. Seleção por meio de amostras de conveniência.

Amostra de resposta voluntária: o pesquisador solicita aos membros de uma população-alvo para que eles participem da amostra e as pessoas decidem se entram ou não. Esses tipos de amostras são enviesados porque as pessoas podem ter interesses particulares ou opiniões negativas e tendem a querer participar.

3.2.3 Tamanho amostral

A determinação do tamanho de uma amostra é de suma importância, pois amostras desnecessariamente grandes acarretam desperdício de tempo e de dinheiro e amostras muito pequenas podem levar a resultados não confiáveis, ameaçando a validade da pesquisa.

Não existe um número estabelecido para o tamanho da amostra. Há uma solução para cada caso. O tamanho da amostra depende (26):

- do tipo de problema;
- do tipo de variável;
- da magnitude do erro estatístico aceito pelo pesquisador;
- da diferença minimamente importante entre os grupos;
- da probabilidade de que a amostra identifique uma diferença verdadeira: *Poder estatístico*;
- do tempo, dinheiro e pessoal disponível, bem como da dificuldade em se obterem dados e da complexidade da pesquisa.

O tamanho amostral mínimo é determinado por fórmulas estatísticas complexas. Os cálculos são muito pesados, mas agora, felizmente, existem programas de computador disponíveis que realizam este trabalho, por exemplo o **G-Power3** (27). Além disso, é possível acessar um site que fornece informações e ferramentas para o cálculo amostral em pesquisas da área da saúde[1]. Existem tabelas extensas para calcular o número de participantes (28) para um determinado nível de poder (e vice-versa).

3.3 Principais Delineamentos de Pesquisa

Em geral, a pesquisa clínica, é dividida em dois tipos de investigação. O primeiro é aquele em que o observador apenas observa o doente, as características da sua doença e sua evolução, sem atuar de modo a modificar qualquer aspecto que esteja estudando. Trata-se de **estudo observacional**.

O segundo corresponde aos **estudos experimentais**, onde o pesquisador não se limita a observar, mas promove uma intervenção com o objetivo de conhecer os efeitos dessa sobre os participantes da pesquisa. A intervenção pode ser a prescrição de um medicamento, uma dieta, atividade física ou repouso, ou simplesmente, o estabelecimento de um programa de atenção à saúde.

Os estudos podem ser também classificados em primários ou secundários ou integrativos (29). Estudos primários correspondem a pesquisas originais que constituem a maioria das publicações encontradas nas revistas médicas. Estudos secundários são aqueles que procuram sumarizar e extrair conclusões de estudos primários

[1] http://calculoamostral.bauru.usp.br/calculoamostral/index.php

- *Estudos Primários*
  - Estudos Observacionais
    - Relato de Caso e Série de Casos
    - Estudo Transversal
    - Estudo Caso-controle
    - Estudo de Coorte
  - Estudos Experimentais
    - Experimento laboratorial
    - Ensaio Clínico
- *Estudos Secundários*
  - Revisões não sistemáticas
  - Revisões Sistemáticas
  - Diretrizes (*Guidelines*)
  - Análise de decisão
  - Análise Econômica

3.3.1 Elementos básicos de um delineamento de pesquisa

Os estudos contêm três elementos básicos:

1. Variáveis componentes: Nas investigações das relações entre as variáveis identificam-se pelo menos duas variáveis nos estudos epidemiológicos.
   a. *Desfecho*: Aquilo que vai acontecer durante uma investigação na mensuração da condição de saúde-doença. Sinônimo: variável dependente.
   b. *Exposição*: O fator que precede o desfecho. Sinônimos: fator em estudo, variável preditora, variável independente.
2. Temporalidade: Quanto ao tempo os estudos podem ser contemporâneos, retrospectivos e prospectivos, de acordo como os dados são obtidos em relação ao momento atual.
3. Enfoque: Um estudo pode ter vários enfoques. Na maioria deles, na área médica, eles relacionam-se à prevenção, ao diagnóstico, à terapêutica e ao prognóstico.

3.4 Estudos Observacionais

3.4.1 Relato de Caso ou Série de casos

No relato de caso, descrevem-se casos raros, eventos não comuns ou inesperados, doenças desconhecidas ou raras. Um evento notável deve ser identificado. Um relato de caso tem a descrição de até dez casos. Acima deste número tem-se uma série de casos (30).

Metodologicamente, faz-se um relato descritivo simples de características interessantes observadas em um paciente ou grupo de pacientes. Os indivíduos são acompanhados em um espaço de tempo curto e não possuem participantes-controles. A coleta dos dados é, na maioria das vezes, retrospectiva.

Uma série de casos não é planejada e não envolve quaisquer hipóteses investigativas. Pode ser empregada como precursor de outros estudos.

3.4.2 Estudos Transversais ou Seccionais

Os estudos transversais são também conhecidos como estudos seccionais. Este tipo de estudo fornece a informação sobre a prevalência, ou seja, a proporção dos indivíduos que tem a doença ou condição clínica em um determinado momento. Por este motivo são também conhecidos como estudos de prevalência (31).

Observam dados coletados em um grupo de indivíduos em um único momento, sem um período de seguimento. O desfecho e exposição são avaliados no mesmo momento no tempo. Os dados são coletados apenas uma vez para cada indivíduo, podendo ser em dias diferentes em diferentes sujeitos. As informações são, em geral, obtidas em um curto espaço de tempo.

É um estudo estático, representa a "fotografia" de um momento. Entretanto, se as variáveis preditora e de desfecho são definidas apenas com base nas hipóteses causa-efeito do investigador e não no delineamento do estudo, é possível também examinar associações.

Os estudos de corte transversal, de um modo geral, são desenhados para determinar *"O que está acontecendo?"*. São usados para:

- Determinar a prevalência de uma doença, como a prevalência de HIV em gestantes.
- Pesquisar atitudes ou opiniões em relação a um determinado assunto (pesquisa de satisfação)
- Verificar interrelações entre variáveis, como observação das características de fumantes pesados em relação ao sexo, idade, etc.
- Enquetes

3.4.2.1 Cuidados na interpretação de dados de estudos transversais

1. *Efeito temporal*
   Como os dados (exposição e desfecho) são coletados no mesmo momento, fica difícil estabelecer qualquer relação temporal entre eles (dilema ovo/galinha). Por exemplo, não é possível estabelecer uma relação de causalidade entre hipertensão e doença cardíaca se os dados são coletados de forma a ficar impossível saber que surgiu em primeiro lugar.
2. *Estudos transversais repetidos*
   Os estudos transversais, algumas vezes, são repetidos em outro momento ou em outros locais com a finalidade de verificar variabilidade nos achados. Por exemplo, medir a prevalência de uma doença em momentos diferentes ou em diferentes locais. Os indivíduos serão um pouco diferentes, devendo-se interpretar as diferenças destes resultados com cautela.
3. *Estudos transversais que parecem longitudinais*
   Uma armadilha comum é confundir um estudo seccional com um longitudinal porque os dados foram coletados através do tempo até completar o tamanho amostral previsto. O importante é que os dados (variável preditora e desfecho) foram coletados somente uma vez para cada indivíduo e no mesmo momento. Isto gera uma interpretação errônea se analisarmos como um estudo longitudinal.

3.4.2.2 Análise dos Estudos Transversais

Quando se compara a prevalência de doença em expostos e não expostos, a medida de associação usada é a *Razão de Prevalência Pontual* (RPP).

3.4.3 Estudos Caso-Controle

Para examinar a possível associação de uma exposição a uma determinada doença, identifica-se um grupo de doentes (casos) e, com a finalidade de comparação, um grupo de pessoas sem a doença (controles) e determina-se a chance (*odds*) de exposição e não exposição entre casos e entre controles.

Os estudos caso-controle, portanto, partem da presença ou ausência de um desfecho e após olham para trás no tempo (retrospectivamente) para detectar possíveis fatores de risco (**Figura 3-1**)(32). Analisam o que aconteceu e são usados para investigar fatores de risco de doenças raras onde um estudo prospectivo seria muito longo para identificar uma quantidade suficiente de casos.

É útil também para investigar surtos agudos (infecção alimentar) para identificar se existe ou não associação entre a exposição e o desfecho investigado. Com frequência, os estudos caso-controle são o primeiro passo na busca de uma etiologia quando há suspeita de que alguma de várias exposições esteja associada a uma determinada doença.

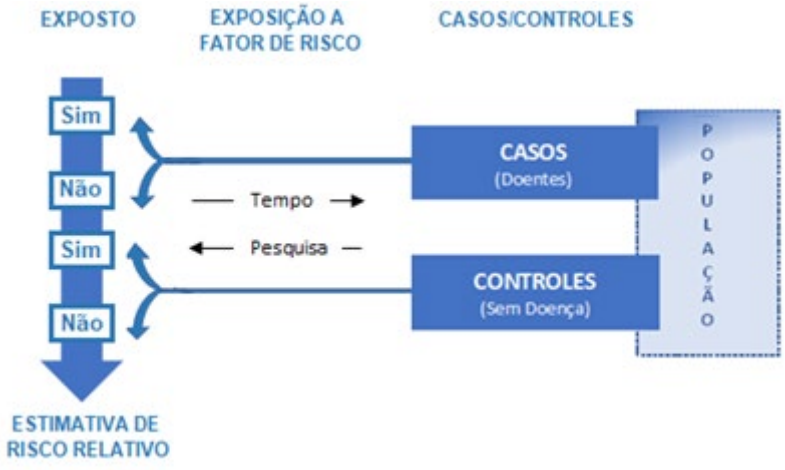

Figura 3-1 Desenho de um estudo caso-controle

3.4.3.1 Seleção dos casos

Os casos podem ser selecionados de várias fontes, incluindo indivíduos hospitalizados, de consultórios ou clínicas, principalmente quando registros adequados são mantidos.

Muitos problemas podem ocorrer na seleção de casos, neste tipo de estudo. Se os casos forem selecionados de um único hospital, quaisquer fatores de risco identificados podem ser apenas daquele hospital, em decorrência do padrão de referência e nível de atendimento (um hospital terciário que apenas atende um determinado convênio, por exemplo, o Sistema Único de Saúde). Por isso, devem ser utilizados casos procedentes de vários hospitais da comunidade, pois aí os casos pertenceriam a diferentes grupos sociais e diferentes graus de gravidade da doença.

3.4.3.1.1 *Casos incidentes ou prevalentes*

Os casos usados nos estudos caso-controle podem ser casos incidentes (recém-diagnosticados) ou casos prevalentes da doença (pessoas que apresentaram a doença em algum período).

O problema do uso de casos incidentes é que há necessidade de se esperar que novos casos sejam diagnosticados e isto pode requerer muito tempo. Enquanto os casos prevalentes já estão disponíveis havendo um maior número disponível para o estudo. Em ambos os modelos existem problemas, pois nos casos prevalentes algumas pessoas podem morrer logo após o diagnóstico e estarem pouco representadas no estudo. Por outro lado, nos casos incidentes, serão excluídos os pacientes que morreram antes do diagnóstico ser feito. Não existe uma solução fácil para este problema, mas é importante lembrar-se destas questões ao interpretar os resultados e tirar conclusões do estudo.

3.4.3.2 Seleção dos controles

Da mesma forma do que nos estudos experimentais, a escolha dos controles afeta a comparação com os casos (33). A escolha dos controles inclui:

- Pacientes do mesmo hospital, mas com condições ou doenças não relacionadas;
- Pacientes pareados um a um em relação a fatores prognósticos, tais como sexo e idade;
- Uma amostra aleatória originária da mesma população de onde provêm os casos.

Sem dúvida, o melhor grupo controle é a terceira opção, mas esta é raramente possível. Por este motivo, alguns estudos caso-controle incluem mais de um grupo controle para tornar o estudo mais robusto

3.4.3.2.1 *Controles pareados*

O emparelhamento é definido como processo de seleção dos controles para que sejam semelhantes aos casos em algumas características como, por exemplo, idade, gênero, raça, condição socioeconômica e ocupação.

Controles emparelhados são bastante comuns. O autor deve ter o cuidado de especificar cuidadosamente o modo como houve o pareamento. Por exemplo, “emparelhado por idade dentro de dois anos” mostra a amplitude do pareamento. É difícil realizar o emparelhamento para muitos fatores, pois um pareamento seguro não existe. Em um delineamento pareado, a análise estatística deve levar em conta o emparelhamento e os fatores usados por ele. Onde um indivíduo em um par tiver um dado perdido, ambos devem ser omitidos da análise estatística.

3.4.3.3 Estudos caso-controle aninhados

Um delineamento do tipo caso-controle aninhado é um estudo de caso-controle ”aninhado” em um estudo de coorte (34). É um excelente desenho para variáveis preditoras que são caras para medir e que podem ser avaliadas no final do estudo em indivíduos que desenvolvem o resultado durante o estudo (casos) e em uma amostra daqueles que não o fazem (controles).

O investigador começa com uma coorte adequada (**Figura 3-2**) (35) com casos suficientes ao final do acompanhamento para fornecer poder adequado para responder à pergunta de pesquisa. No final do estudo, aplica critérios que definem o resultado de interesse para identificar todos aqueles que desenvolveram o resultado (casos). Em seguida, seleciona uma amostra aleatória dos indivíduos que não desenvolveram o resultado (controles).

A principal razão para usar delineamentos caso-controle aninhado é reduzir o trabalho e o custo na coleta de dados. A principal desvantagem desse projeto é que muitas questões e circunstâncias da pesquisa não são passíveis de armazenamento para posterior análise.

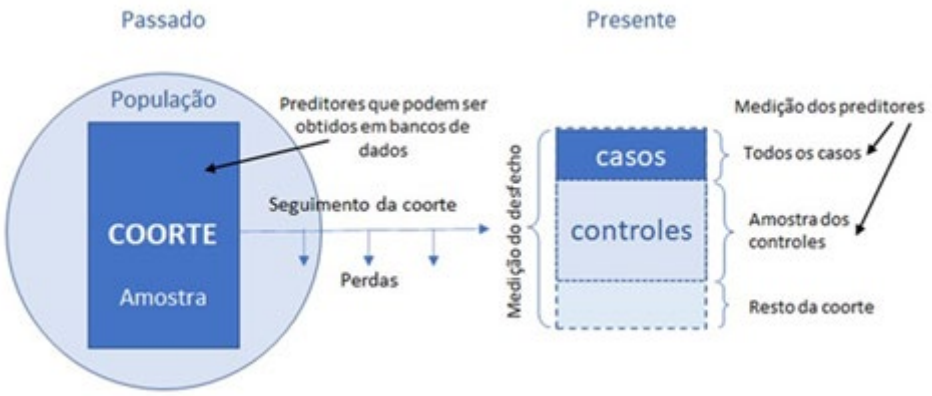

Figura 3-2 Desenho de um estudo caso-controle aninhado

3.4.3.4 Estudo caso-controle de base populacional

São os estudos caso-controle onde os casos e controles são uma amostra completa ou probabilística de uma população definida

3.4.3.5 Limitações dos estudos caso-controle

Várias limitações podem afetar os estudos caso-controle:

- A escolha do grupo controle afeta as comparações entre casos e controles;
- Os dados da exposição ao fator de risco são coletados retrospectivamente e dependem da memória dos participantes, registros médicos e, portanto, podem ser incompletos, sem acurácia ou enviesados (viés de memória);
- Se o processo que conduz à identificação dos casos está relacionado a um possível fator de risco, a interpretação dos resultados será difícil (viés averiguação).
    - Por exemplo: suponha que os casos sejam mulheres jovens com hipertensão selecionadas de uma clínica de contracepção. Nesta situação, um possível fator de risco, o anticoncepcional oral (ACO), estará vinculado à seleção dos casos e, desta forma, o uso de ACO será mais comum entre os casos do que entre os controles populacionais.

3.4.3.6 Análise dos Estudos Caso-Controle

A principal estratégia de análise é o cálculo da *odds ratio* (Razão de Chances), que pode ser interpretado como uma estimativa do Risco Relativo.

O Risco Relativo somente pode ser calculado quando é possível o cálculo da incidência (ver seção 18.5.2). Nos estudos caso-controle, isso não é possível, pois aqui o estudo começa com casos e controles em vez de indivíduos expostos e não expostos ao fator de risco. Desta maneira, se comparam as *odds* (chance) de uma exposição passada a um fator de risco suspeitado em indivíduos doentes e em controles não doentes. Esta relação é denominada de *odds ratio* (ver seção 18.5.1).

3.4.4 Estudos de Coorte

Os estudos de coorte são considerados o padrão-ouro dos estudos observacionais. Seu nome se originou das coortes dos soldados romanos, cada uma delas constituída por 480 a 600 legionários. As coortes romanas eram distintas entre si e tinham sua identidade determinada por, ao menos, uma característica comum entre os indivíduos de cada grupo. Podia ser por características estratégicas no campo de batalha, por uma cor presente na indumentária, ou outras. Em Epidemiologia, o termo coorte permaneceu com significado semelhante.

Em um estudo de coorte, um grupo de pacientes sadios (coorte), expostos ou não a um suspeitado fator de risco, é seguido através do tempo para determinar a incidência da doença em questão em cada um dos grupos (36).

Neste modelo de estudo, a característica comum aos dois grupos é a exposição. Tem-se uma coorte de expostos e uma coorte de não expostos que são acompanhadas por um período de tempo que permita o aparecimento do desfecho. No final do estudo, compara-se a incidência do desfecho (doença) entre os expostos com a incidência do desfecho entre os não expostos. Se existe uma associação positiva entre a exposição e o desfecho, se espera que a incidência do desfecho entre os expostos seja maior do que a incidência de desfecho entre não expostos.

Um esquema simplificado de um estudo de coorte é mostrado na **Figura 3-3** (37).

Figura 3-3 Desenho de um estudo de coorte

Observar que como se identifica novos casos (incidência) à medida que eles ocorrem, é possível determinar uma relação temporal entre a exposição e a doença, isto

é, se a exposição precedeu o início da doença. Isto é *fundamental para estabelecer uma relação causal entre a exposição e a doença*.

Os estudos de coorte têm semelhança com os ensaios clínicos randomizados. Ambos os estudos comparam grupos expostos a grupos não expostos. Não havendo possibilidade de realizar a randomização, por exemplo, por motivos éticos quando a exposição é sabidamente prejudicial, é indicado um estudo de coorte. A diferença fundamental, portanto, é a ausência de randomização nos estudos de coorte.

Existem duas maneiras básicas para formar os grupos:

1. Seleciona-se a população-alvo baseado no fato dos indivíduos estarem expostos ou não ao fator em estudo (Figura 3-3);
2. Ou seleciona-se a população-alvo antes que qualquer um dos seus membros se torne exposto, ou antes, que a exposição seja identificada (Figura 3-4). Um exemplo típico deste modelo é o clássico Estudo de Framingham (38).

Figura 3-4 Desenho de uma coorte com grupos expostos e não expostos. (39).

3.4.4.1 Tipos de estudos de coorte

De acordo com as características do seguimento, as coortes podem ser:

1. **Estudo de Coorte Prospectivo** (Coorte Concorrente ou Longitudinal), onde os grupos são montados no presente, coletados os dados basais deles e continua-se a coletar dados com o passar do tempo até a doença se desenvolver ou não.
2. **Estudo de Coorte Retrospectivo ou Histórico** (Coorte não concorrente), onde a exposição é avaliada em dados passados e o desfecho (doença ou não) é verificado no momento do início do estudo. O problema aqui é que a averiguação da exposição depende dos registros pregressos.
3. **Estudo de Coorte Misto** (Prospectivo e Retrospectivo), onde a exposição é verificada em registros objetivos no passado (como em uma coorte histórica) e o seguimento e a medida do desfecho se fazem no futuro.

3.4.4.2 Vieses em estudos de coorte

Os potenciais vieses nos estudos de coorte são os seguintes:

1. **Viés de confusão** – é a grande ameaça dos estudos observacionais. O confundimento causa um erro sistemático na inferência, podendo aumentar ou diminuir uma associação observada entre exposição e doença. Uma variável funciona como fator de confusão quando ela está associada com a exposição e ao mesmo tempo com a doença. Ela não deve fazer parte da cadeia causal da exposição à doença. Por exemplo, num estudo sobre fatores de risco, uma associação entre o hábito de beber café e a doença coronária é detectada. Porém, se não for considerado o fato de que os fumantes bebem mais café do que os não-fumantes, pode-se chegar à errônea conclusão de que o café é um fator de risco independente para doença coronária, o que não corresponde à realidade. Neste caso, o café é um fator de confusão e não um fator causal independente para a doença coronária (40).
2. **Viés na avaliação dos desfechos** – este viés pode ocorrer quando o pesquisador que avalia o desfecho também sabe sobre o *status* de exposição dos sujeitos da pesquisa. Evita-se este problema "cegando" a pessoa que faz a avaliação da doença.
3. **Viés de informação** – ocorrem principalmente em estudos históricos onde as informações dependem de registros passados e podem ser diferentes entre as pessoas expostas e não expostas.
4. **Viés de não resposta e perdas de acompanhamento** – a não participação e as perdas podem introduzir um grande viés, alterando o cálculo da incidência nos expostos e entre os não expostos.
5. **Viés de análise** – se os estatísticos tiverem alguma hipótese em relação aos dados que estão analisando, eles podem introduzir vieses em suas análises.

3.4.4.3 Análise dos estudos de coorte

Para verificar se existe associação entre certo desfecho (doença) e uma determinada exposição calcula-se o **Risco Relativo** (RR). Este é definido como a razão entre a incidência (risco) em expostos e a incidência (risco) em não expostos (ver seção 18.5.2).

3.4.4.4 Vantagens e desvantagens dos estudos de coorte

1. Vantagens
    a. Adequado para exposições raras
    b. Bom poder para testar hipóteses
    c. Importante em estudos etiológicos e prognósticos
    d. Salienta os múltiplos desfechos de uma exposição
2. Desvantagens
    a. Inadequado em desfechos raros
    b. Perdas no seguimento levam a viés de seleção
    c. Demorado/elevado custo

3.5 Ensaios Clínicos

Experimentos são estudos nos quais o pesquisador *manipula a variável preditora* (intervenção) e observa o efeito no desfecho que está sendo avaliado ao longo

do tempo. A abordagem experimental, especificamente, o ensaio clínico randomizado controlado é a ferramenta de escolha para comparar terapêuticas ou intervenções.

Os estudos experimentais podem também comparar os cuidados prestados por serviços de saúde, programas de educação em saúde e estratégias administrativas. Os estudos experimentais realizados com seres humanos são denominados de **ensaios clínicos**.

Nos ensaios clínicos não controlados os indivíduos servem como seus próprios controles (antes-e-depois). Os resultados destes estudos estão sujeitos vários problemas:

- **Melhora previsível**. Paciente melhora espontaneamente e não pelo tratamento.
- **Flutuação na gravidade da doença**.
- **Efeito Hawthorne**: o indivíduo melhora pela atenção e não pela terapêutica (41).
- **Regressão à média**: uma limitação importante surge quando se quer avaliar a evolução de um grupo que tenha sido selecionado por estar no extremo de uma distribuição sem que haja um grupo controle. Empiricamente, observa-se que indivíduos que se encontrem num determinado momento, em um dos extremos de uma distribuição, tendem a estarem menos distantes da média em um momento posterior, sem que qualquer intervenção tenha sido desenvolvida. Este fenômeno é conhecido como efeito de *regressão à média*. Por exemplo: uma pessoa com uma doença crônica tem dias piores e outros melhores. Se ela é medicada com gotas homeopáticas ou faz uso de florais nos dias em que se sente excepcionalmente mal vai notar que é frequente uma melhora, seguindo estes "tratamentos". Não que eles funcionem, mas pela regressão à média (42).

3.5.1 Características de um ensaio clínico

Um ensaio clínico deve ter algumas características fundamentais (**Figura 3-5**) (43):

1. Os indivíduos devem ser designados por randomização para os grupos de comparação.
    a. A randomização é a melhor abordagem no delineamento de um ensaio clínico (44).
    b. Randomizar significa sortear (por meio de computadores, tábua de números aleatórios) os indivíduos para decidir a alocação dos mesmos em um dos grupos de estudo. O elemento decisivo da randomização é a imprevisibilidade da próxima alocação.
2. O pesquisador compara o grupo de estudo com um grupo controle apropriado.
3. O investigador manipula a variável independente (preditora).

Figura 3-5 Estrutura de um ensaio clínico randomizado.

3.5.2 Elementos básicos de um ensaio clínico

3.5.2.1 Seleção dos participantes

Os pesquisadores devem determinar e explicar detalhadamente os critérios de inclusão e de exclusão:

- Objetivos dos critérios de inclusão e exclusão
  - Restringir a heterogeneidade da amostra
  - Diminuir o número de variáveis independentes
  - Fazer com que exista uma chance maior de que as diferenças nos desfechos estejam relacionadas aos tratamentos
  - Melhorar a *validade interna*, ou seja, o grau em que os resultados do estudo são consistentes para aquela amostra particular de indivíduos. Esta validade depende basicamente do rigor metodológico usado para delinear o ensaio clínico, podendo ser ameaçada por dois tipos de erros: sistemático ou aleatório.
  - Tornar a generalização (validade externa) mais precisa. Entretanto deve-se ter cuidado com critérios de inclusão e exclusão muito rígidos, pois podem diminuir esta capacidade de generalização

O grau de detalhamento deve ser suficientemente preciso para permitir que outros reproduzam o estudo. O tamanho da amostra deve ser claramente determinado pelo poder do teste estatístico. Poder é a habilidade de o teste estatístico detectar diferenças entre os grupos, dado que tais diferenças existam na população em estudo. Lembrar que resultados não significativos podem ser apenas uma evidência para um inadequado tamanho amostral.

O grupo controle deve ser selecionado utilizando-se os mesmos critérios do grupo experimental. Prestar atenção em possíveis armadilhas que podem gerar vieses:

- Uso de grupo controle histórico (não concorrente);
- Grupo controle selecionado de outros locais (outras clínicas, outros hospitais).

O grupo controle adequado é um grupo controle concorrente, tratado no mesmo momento e no mesmo local do grupo experimental. O característico é o grupo controle não receber tratamento. Mais comumente recebem um placebo, indistinguível do tratamento experimental, mas sem componente ativo. Mesmo assim, pode haver melhora dos participantes do grupo controle (Efeito Placebo) (45). Quando não for ético

suspender o tratamento e administrar placebo, o grupo controle pode ser constituído por indivíduos que recebem o tratamento padrão.

3.5.2.2 Alocação

A alocação deve ser aleatória. A randomização é a principal técnica para reduzir o viés, criando grupos homogêneos. Como foi visto, é uma das características fundamentais dos ensaios clínicos. O poder da randomização depende da ocultação da sequência de alocação.

A randomização pode ser:

- **Completa**: os indivíduos que obedecem ao critério de inclusão e exclusão são randomizados de modo que todos têm a mesma probabilidade de pertencer a cada um dos grupos. Isto maximiza o poder. Pode ser feita por blocos para assegurar a igualdade numérica dos grupos (estudos multicêntricos).
- **Estratificada**: os participantes são estratificados de acordo com possíveis variáveis de confusão (gravidade da doença, idade, sexo, etc.) e a randomização é realizada dentro de cada estrato.
- **Randomização e alocação desigual**: os sujeitos têm uma maior probabilidade de ser randomizados em um grupo (em geral, grupo experimental) do que o outro (comparação). Este tipo de estudo tem menor poder.

3.5.2.3 Condução/Seguimento/Avaliação

Em um ensaio clínico deve estar assegurado de que o estudo tenha um tempo de seguimento adequado, pois nem todos os indivíduos participam conforme o plano original. Podem ocorrer perdas de alguns pacientes durante o acompanhamento, seja porque com o tempo se constata que eles não têm a doença em estudo ou porque não aderiram ao tratamento ou intervenção e abandonaram o estudo. Quanto maior o número de pacientes perdidos e menos informações sobre eles, menos confiança pode ser colocada nos resultados do estudo. De um modo geral, não se deve tolerar perdas que sejam maiores que a incidência do desfecho no estudo. Uma regra simples é que perdas menores que 5% produzem pouco viés e perdas maiores que 20% são uma ameaça importante à validade do estudo. As perdas entre 5 e 20% devem ser avaliadas com cuidado, se possível utilizando-se uma análise de sensibilidade (pior cenário), principalmente se as perdas forem diferentes nos grupos pelo maior risco de viés.

Neste tipo de análise, nos estudos com resultado positivo, todos os pacientes perdidos no grupo experimental, inicialmente, são considerados como tendo o desfecho. Posteriormente, analisa-se como se nenhum dos indivíduos perdidos no grupo controle atingiu o desfecho. Se o resultado permanecer positivo, as perdas não afetaram a validade do estudo. Estudos sem relato adequado ou nenhum relato de perdas ou exclusões devem ser avaliados com muito cuidado.

Outro aspecto importante, no seguimento dos sujeitos da pesquisa, é o tratamento igual de todos os grupos. Para garantir este princípio, utiliza-se da técnica de **cegamento** ou **mascaramento** (46). Esta técnica impede que os participantes da pesquisa (pesquisadores, avaliadores e participantes) tomem conhecimento de qual grupo de tratamento o participante se encontra. Este conhecimento antecipado pode influenciar as expectativas, as opiniões e as crenças em relação aos resultados do estudo. O cegamento tem como principal finalidade a eliminação do viés de aferição, além de melhorar a adesão ao tratamento, reduzir as perdas de seguimento e diminuir

o viés causado por co-intervenções (assistência suplementar maior para um dos grupos).

Quando o cegamento ocorre nos pacientes e nos pesquisadores, diz-se que o estudo é **duplo-cego**. Se ele também incluir os avaliadores do estudo, ele é **triplo cego**. Um ensaio clínico em que não há cegamento é dito aberto (*open label*, no caso de estudos com fármacos).

A avaliação dos desfechos também pode afetar os resultados. É importante garantir-se que aqueles que registram os desfechos estejam cegados em relação a que grupo o sujeito da pesquisa pertence. Os autores devem estabelecer regras cuidadosas para decidir se um desfecho ocorreu ou não e despender esforços iguais para identificar desfechos para todos os pacientes no estudo.

3.5.2.3.1 Intenção de tratar

Os pesquisadores violam a randomização se omitirem da análise os pacientes que não receberam a intervenção designada ou, pior ainda, contarem eventos que ocorreram nos sujeitos não aderentes que foram designados para a intervenção contra o grupo controle. Os sujeitos de uma pesquisa, para evitar tal viés, devem ser analisados dentro do grupo para o qual eles foram alocados pela randomização (47). Este princípio é denominado **intenção de tratar**.

3.5.2.3.2 Análise e magnitude do efeito

Calcula-se uma série de estimativas quantitativas para analisar a magnitude do efeito da intervenção em um ensaio clínico. Entre elas, destacam-se o **Risco Relativo**, **Redução Relativa do Risco**, **Número Necessário para Tratar** que serão estudados no capítulo 18.

Outro método para avaliar resultados de um ensaio clínico para dados de tempo até o evento é a **análise de sobrevida**. Esta fornece informação sobre a rapidez com que os eventos ocorrem. A curva de sobrevida pode utilizar dados de pacientes acompanhados por diferentes períodos de tempo.

3.5.3 Ensaios clínicos de equivalência e não inferioridade

Ensaios clínicos controlados com placebo são ideais para avaliar a eficácia de um tratamento. Eles permitem o controle do efeito placebo e são mais eficientes, exigindo um menor número de pacientes para detectar um efeito do tratamento. Um ensaio clínico placebo controlado é eticamente justificado se não existe tratamento padrão, se o tratamento padrão não se mostrou eficaz, não há riscos associados com o retardo no tratamento e se a possiblidade de se retirar do estudo está incluída no protocolo. Sempre que possível e justificado, os ensaios clínicos placebo controlados devem ser a primeira escolha para avaliação de um tratamento.

Dado que um grande número de tratamentos eficazes comprovados está disponível, ensaios clínicos controlados por placebo são, muitas vezes, antiéticos. Nestas situações, ensaios clínicos com controle ativo são geralmente apropriados.

Se o objetivo do ensaio clínico é testar se um novo tratamento é similar em eficácia a um tratamento já existente, ele é denominado de **Estudo de Equivalência**. O Ensaio Clínico é delineado de maneira que possa demonstrar que, dentro limites aceitáveis, os dois tratamentos são igualmente eficazes. Existe equivalência quando a diferença observada entre os dois tratamentos for menor que a máxima diferença aceitável, determinada previamente. Estes limites devem ser clinicamente apropriados.

Se condição em investigação for muito grave, os limites para a equivalência devem ser estreitados. Quanto menor forem os limites de equivalência, maior o tamanho amostral. Este delineamento é útil se o novo tratamento trouxer benefícios, tais como menores efeitos colaterais, facilidade no uso e ser mais barato.

Em muitos estudos com controle ativo, os pesquisadores desejam comprovar que o tratamento em estudo, no mínimo, não é substancialmente pior que o tratamento controle. Estes estudos são chamados de **Estudos de Não Inferioridade**. Um aspecto importante do delineamento e da interpretação desses estudos é a determinação da margem de não inferioridade. Os estudos de não inferioridade devem demonstrar, pelo menos, que o tratamento em estudo tem alguma eficácia, não inferior ao tratamento padrão. A análise dos estudos de não inferioridade é, por natureza, unidirecional.

Quando um ensaio clínico busca evidenciar que um tratamento é melhor do que outro ele é denominado **Estudos de Superioridade**. Quando o ensaio clínico é delineado, ele deve ter uma hipótese bilateral e o tamanho da amostra definido de maneira que haja alto poder estatístico para detectar uma diferença clinicamente significativa entre os dois tratamentos. Os ensaios clínicos clássicos têm esta característica. Entretanto, nos dias atuais, este desenho de estudo pode não ser eticamente possível, uma vez que é pouco provável que não exista um tratamento com algum benefício comprovado. A comparação, portanto, deverá ser feita com o tratamento já existente, provando que o tratamento em estudo é similar ou, pelo menos, não seja inferior (48).

3.5.4 Outros tipos de ensaios clínicos

3.5.4.1 Ensaio clínico com delineamento cruzado

No delineamento cruzado (*crossover design*), os sujeitos da pesquisa são randomizados para um grupo e depois mudados para o outro grupo (**Figura 3-6**). Cada sujeito serve como seu próprio controle, diminuindo a variabilidade intragrupo, aumentando o poder e consequentemente, reduzindo o erro β (erro que ocorre quando a análise estatística dos dados não consegue rejeitar uma hipótese, no caso desta hipótese ser falsa). É um tipo de delineamento bastante atrativo e útil (49).

A maior desvantagem é o efeito residual (*carryover*), por isso os estudos cruzados devem ter um período de *washout*, período sem nenhum tratamento. Este período de tempo deve ser suficiente para a eliminação da droga para se ter certeza de que nenhum efeito da terapia permaneceu. Também pode haver um viés de acordo com a ordem de administração das terapias, pois os pacientes podem reagir de modo diferente como resultado do entusiasmo no início do tratamento que pode diminuir com o tempo.

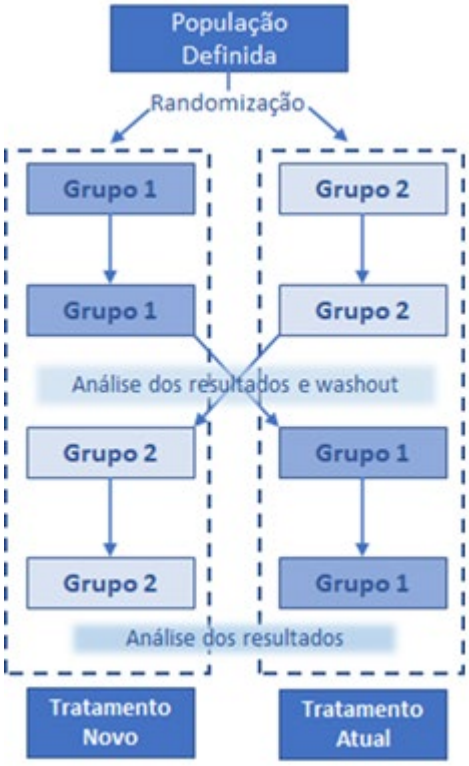

Figura 3-6 Ensaio clínico randomizado com delineamento cruzado.

3.5.4.2 Delineamento Fatorial

Uma variação interessante de ensaio clínico é o *delineamento fatorial*. Este tipo de estudo permite que sejam testadas duas drogas em apenas um estudo, assumindo que os desfechos antecipados para as duas são diferentes e que seus modos de ação são independentes. Este desenho de estudo gera economia.

Um exemplo de delineamento fatorial é observado no *Physician's Health Study* onde usando um delineamento fatorial 2 x 2 foi testada a aspirina para a prevenção primária de doença cardiovascular (50), e betacaroteno para a prevenção primária de câncer.

No estudo da prevenção primária do câncer, os autores concluíram, após 12 anos de suplementação de betacaroteno, que o mesmo não produziu nem benefícios e nem prejuízos em termos de incidência de câncer (51).

3.5.5 Fases de um ensaio clínico

Para a realização de um ensaio clínico, a intervenção deve passar por várias fases (52).

Fase Não Clínica

Antes de começar a testar novos tratamentos em seres humanos, os cientistas testam as substâncias em laboratórios (in vitro) e em animais de experimentação. O objetivo principal desta fase é verificar como esta substância se comporta em um organismo. Assim, após esta fase se pode verificar se o medicamento é seguro para ser testado em seres humanos. Todo este processo é regido por leis da bioética em pesquisa em animais.

Fase Clínica

A fase clínica é a fase de testes em seres humanos. Esta etapa é constituída por quatro fases consecutivas e somente depois de finalizadas todas as fases, a droga poderá ser autorizada para comercialização e disponibilizada para uso em seres humanos. As sucessivas fases dentro da fase clínica são:

Fase I - Um estudo de fase I testa a droga pela primeira vez. O objetivo principal é avaliar a segurança do produto investigado. Nesta fase, o medicamento é testado em pequenos grupos (10 – 30 pessoas), geralmente, de voluntários sadios. Podemos ter exceções se estivermos avaliando medicamentos para câncer ou portadores de HIV-AIDS. Se a droga se mostrar segura, é possível ir para a Fase II.

Fase II - Nesta fase, o número de pacientes é maior (70 - 100). O objetivo é avaliar a eficácia da medicação, isto é, se ela funciona para tratar determinada doença, e também conseguir informações mais detalhadas sobre a segurança (toxicidade). Somente se os resultados forem bons é que o medicamento será estudado como um estudo clínico fase III.

Fase III - Nesta fase, o novo tratamento é comparado com o tratamento padrão existente. São os ensaios clínicos. O número de pacientes aumenta e depende da hipótese (em geral, 100 a 1.000). Devem de preferência utilizar desfechos clínicos, grupo controle, além de serem randomizados e duplo-cegos.

Fase IV - Estes estudos são realizados para se confirmar que os resultados obtidos na fase III são aplicáveis a grande parte dos doentes. Nesta fase, o medicamento já foi aprovado para ser comercializado. A vantagem dos estudos fase IV é que eles permitem acompanhar os efeitos dos medicamentos em longo prazo. É uma fase de vigilância pós-comercialização.

4 Ambiente do R

4.1 Instalação do R básico

Para usar o *R*, há necessidade de carregar o programa básico que contém a sua linguagem de programação. O sistema é formado por um programa básico, *Graphical User Interface* (R-Gui) e muitos pacotes com procedimentos adicionais.

O **site** oficial do *R* fornece as versões atualizadas do software e informações sobre este sofisticado projeto de computação estatística.

Para baixar o R, usa-se um "CRAN Mirror", clicando em CRAN (*Comprehensive R Archive Network*) na margem esquerda, abaixo de *Download*. O CRAN é central no uso do R: é o local de onde se carrega o software e todos os pacotes necessários para instalar e para expandir o *R*.

Em vez de ter um único local, o CRAN é "espelhado" em diferentes locais do mundo. "Espelhado" significa simplesmente que existem versões idênticas do CRAN distribuídas por todo o mundo. É possível baixar o *R* diretamente da **nuvem** ou escolher uma origem mais próxima do seu local de atuação. No Brasil, encontram-se várias opções, como a **Universidade Federal do Paraná**, **Fundação Oswaldo Cruz, RJ**, **Universidade de São Paulo, São Paulo** e **Universidade de São Paulo, Piracicaba**

Após escolher uma das alternativas acima (pode ser qualquer uma delas) surgirá a página *The Comprehensive R Archive Network* com as opções para escolher o sistema operacional. Escolha o sistema de acordo com o seu computador (Windows, macOS ou Linux). Ao clicar em uma dessas opções, se o sistema operacional escolhido é o Windows, aparecerá a página *R for Windows*. Nesta, deve-se clicar em *base*. No caso de outros sistemas operacionais, seguir as orientações mostradas no site do *R*.

Clicando em *base*, haverá um redirecionamento para a página onde aparece a versão do *R* para o Windows mais atual. Clique no link que diz *Download R-…for Window* para baixar o instalador em um diretório do computador, em geral *Downloads*.

Para instalar o programa básico, basta executar o instalador *R-…-win.exe* baixado no diretório. Ao fazer isso, aparece na tela do computador, no canto esquerdo, em baixo, o arquivo salvo. Execute este arquivo com um clique sobre ele. Aparecerá uma janela perguntando *"Deseja permitir que este aplicativo faça alterações no seu dispositivo?"*. Clique em *Sim.* A seguir o instalador pedirá para escolher o Idioma. Selecione Português Brasileiro.

Em sequência aparecerão informações sobre o diretório no qual o *R* será instalado em seu computador. Recomenda-se aceitar a configuração padrão sugerida pelo instalador do software.

A próxima janela pedirá para personalizar os componentes que serão instalados. Recomenda-se usar as configurações sugeridas pelo instalador que irá reconhecer automaticamente a arquitetura do seu sistema Windows (32 e/ou 64 bits).

A partir daqui, siga as recomendações padrão propostas pelo instalador até completar a instalação, clicando em *Concluir*.

O *R* não precisa ser iniciado, pois o software que será usado, neste livro, é o *RStudio*. Este, para ser executado, necessita ter o *R* instalado no computador. Ou seja, o *R* é o programa "cérebro" necessário para as análises de dados que serão realizadas. Ele precisa estar instalado para permitir o funcionamento do *RStudio*.

4.2 RStudio

O *RStudio* é um membro ativo da comunidade *R*. Foi fundado em 2009 por Joseph J. Allaire, engenheiro de software americano. O *RStudio*, inspirado pelas inovações dos usuários de *R* em ciência, educação e indústria, desenvolveu ferramentas gratuitas e abertas para facilitar o uso do *R*.

O *RStudio* é um projeto filiado à *Foundation for Open Access Statistics* (FOAS). A FOAS trabalha para garantir o sucesso do projeto *R*. Eles promovem o uso e o desenvolvimento de software livre para estatísticas, como a linguagem *R* e o ambiente para estatísticas computacionais. Junto está o *R Consortium* que é uma colaboração entre a Fundação R, *RStudio* , Microsoft, TIBCO, Google, Oracle, HP e outros.

O *RStudio* é patrocinado para financiar e inspirar ideias que permitirão que o *R* se torne uma plataforma ainda melhor para a ciência.

4.2.1 Instalação do Rstudio

Para instalar o *RStudio* , acessar o **site** e clicar em *Download* para obter a versão desejada. Recomenda-se a versão *RStudio Desktop – Open Source License* que é gratuita. Esta versão entrega as ferramentas integradas para o *R*.

A seguir, aparecerão os instaladores disponíveis, conforme a plataforma suportada pelo seu computador. As mais utilizadas são Windows e Mac OS X. Neste livro, como base, serão mostrados os passos para a plataforma Windows[1].

Em sequência, executar o instalador baixado *RStudio-2023.06.2-561.exe*[2] e seguir as suas instruções.

4.2.2 Iniciando o Rstudio

Para iniciar o *RStudio* basta clicar no ícone indicativo (**Figura 4-1**) que se encontra no menu *Iniciar* do Windows.

Figura 4-1 Ícone do RStudio

O *RStudio* abre como mostrado na **Figura 4-2**. O *RStudio* é uma interface mais funcional e amigável para o *R*. Contém um conjunto de ferramentas integradas projetadas para ajudá-lo a ser mais produtivo com o R.

Inclui o *Console* , editor que suporta execução direta de códigos e uma variedade de ferramentas robustas para plotagem, exibição de histórico, depuração e gerenciamento de seu espaço de trabalho incluídos em uma interface que está, inicialmente, dividida em 3 painéis:

[1] A instalação para Mac OS X pode ser facilmente obtida em busca do Google. Depois de instalado, o uso do *RStudio* não difere do Windows.

[2] Versão disponível em 10/09/2023.

1. *Console*
2. *Environment, History, Connections, Tutorial*
3. *Files, Plots, Packages, Help*

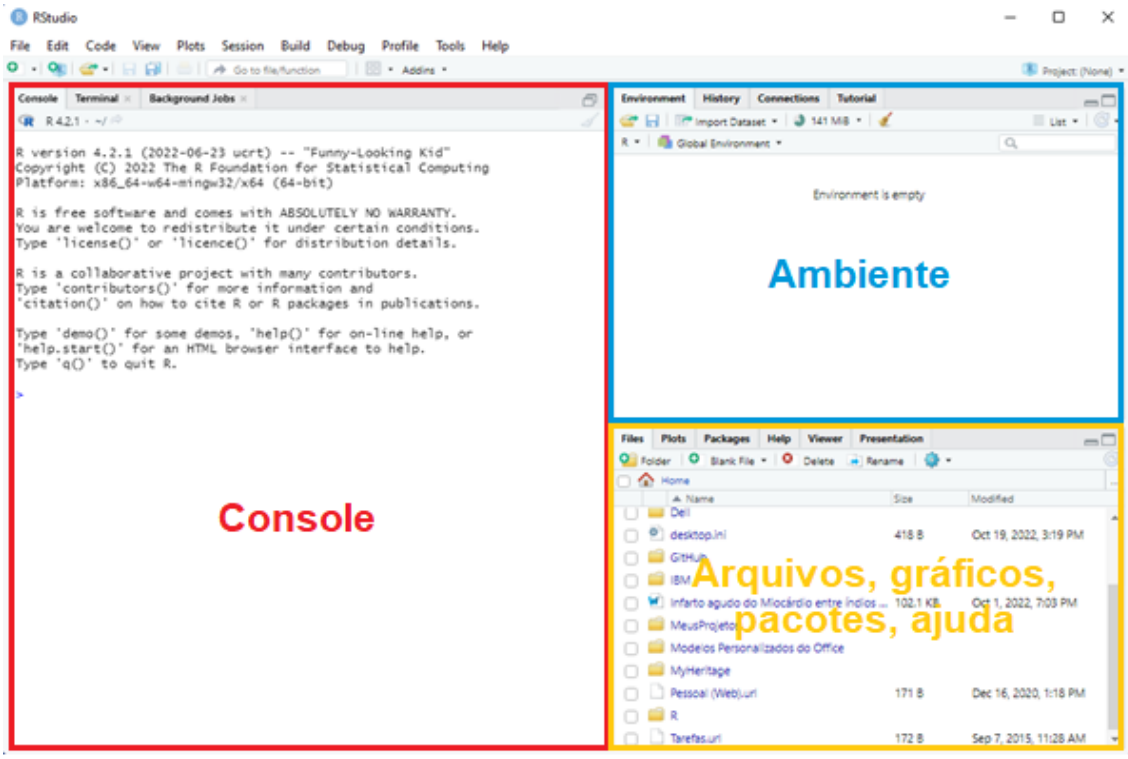

Figura 4-2 Tela inicial do RStudio

4.2.2.1 Console e R Script

Do lado esquerdo fica o **Console** (**Figura 4-2**, em vermelho), onde os comandos podem ser digitados e onde aparecem os resultados da execução destes. Ao abrir o *RStudio* , aparece no *Console* uma série de informações sobre o R, como versão em uso e, por último, o diretório onde está armazenado o espaço de trabalho (*workspace*). Estas informações podem ser facilmente apagadas, clicando na barra de ferramentas, no menu *Edit*, e após em *Clear Console* ou, usando as teclas *Ctrl+L*.

O *Console* é a principal parte do *R*. Aqui é onde o R realmente executa o comando. No início do *Console*, existe um caractere (>). Este é um *prompt* que informa que o R está pronto para receber um novo código. Pode-se digitar o código diretamente no *Console* após o *prompt* e obter uma resposta imediata. Por exemplo, se for digitado 1 + 1 e pressionado *Enter*, o R imediatamente gera uma saída de 2 (**Figura 4-3**).

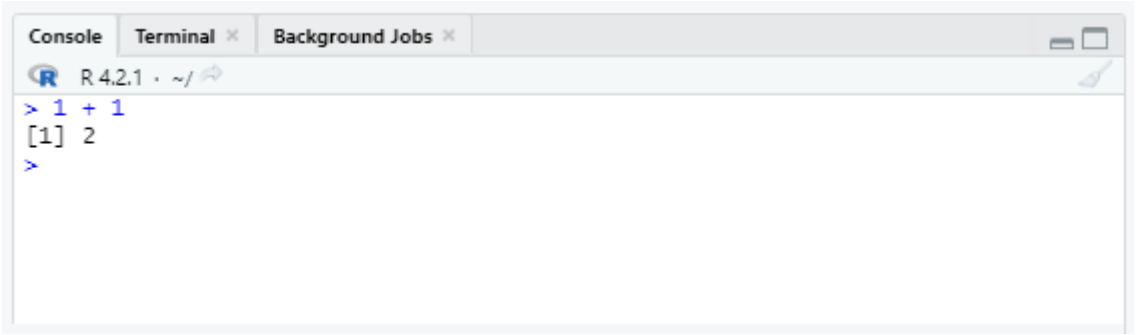

Figura 4-3 Console

Recomenda-se que a maior parte dos comandos sejam digitados no bloco de notas do *RStudio* , o ***R Script***. Reservar o *Console* apenas para depurar ou fazer análises e cálculos rápidos. A razão para isso é simples: se o comando for digitado diretamente no *Console*, ele não será salvo e se for cometido um erro na digitação, haverá necessidade de digitar tudo novamente. Portanto, é melhor escrever os comandos no *R Script* e, quando estiver pronto para executar, enviar para o *Console*.

O *R Script* é o quarto painel do *RStudio* e seu bloco de notas. Ele é criado através do menu *File* > *New File* > *R Script* ou clicando no botão verde com o sinal (+), na barra de ferramentas de acesso rápido, na parte superior à esquerda. Ao criar um novo *R Script* será aberto o painel do bloco de notas (**Figura 4-4**, em verde).

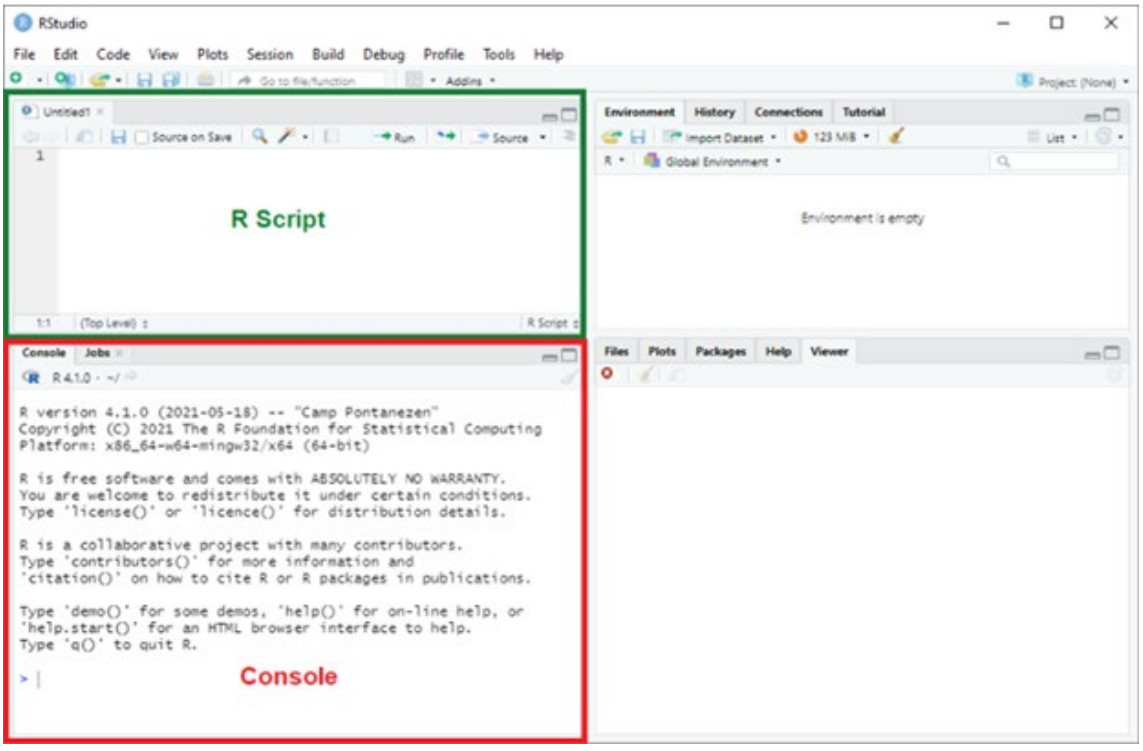

Figura 4-4 R Script

Um diferencial do *RStudio* é que os comandos são autocompletáveis. Basta começar a escrever o comando, inserindo 3 ou mais caracteres, por exemplo, `summ` referente a função `summary()`, usada para sumarizar um conjunto de dados, e surge um menu de opções, facilitando a digitação (**Figura 4-5**).

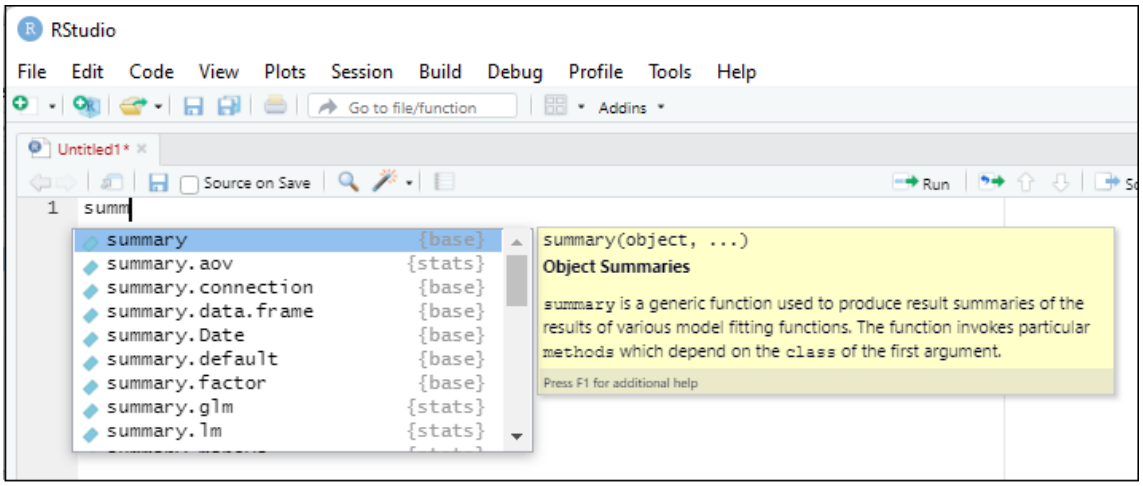

Figura 4-5 Menu autocompletável

Após digitar no *Console*, para que seja executado o comando há necessidade de clicar na tecla *Enter*. No *RScript*, clicar em *Run*, acima, na barra, no lado direito, ou usar o atalho *Ctrl + Enter*. Textos podem ser copiados e colados no *script* e linhas em branco podem ser inseridas. Além disso, no final da sua sessão, é possível salvar o arquivo, que poderá ser recarregado no futuro, se precisar refazer a análise.

Os *scripts* do *R* são apenas arquivos de texto com a extensão (.R). Quando se cria um *R Script*, aparece como *Sem título* (*Untitled*). Antes de começar a digitar um novo *script* no *R Sem título*, recomenda-se salvar o arquivo com um novo nome de arquivo. Dessa forma, se algo no computador falhar durante o trabalho, o *R* terá o código protegido.

Ao digitar o código em um *script*, o R não executa o código enquanto se digita. Para que o R realmente avalie o código digitado, há necessidade de primeiro enviar o código para o *Console*, clicando no botão *Run* ou usando a tecla de atalho *Crtl+Enter*. Cada linha é marcada no início por um número em sequência.

Além da digitação de comandos, o *R Script* permite fazer comentários onde tudo que for escrito após o símbolo # não é considerado, é apenas uma explicação, um esclarecimento. Os comentários são literais, escritos diretamente para explicar o comando executado. São repetidos na saída do *Console* sem não aparecer nos resultados.

4.2.2.2 Ambiente, História, Conexão e Tutorial

No lado superior direito há um painel com quatro abas (**Figura 4-2**, em azul):

1. **Ambiente** (*Environment*) - onde ficam armazenados os objetos criados, as bases de dados importadas, etc., na sessão ativa. É possível visualizar informações como o número de observações e linhas dos bancos de dados ativos. A guia também tem algumas ações clicáveis, como *Import Dataset*, que permite importar arquivos csv, Excel, SPSS, etc.
2. **História** (*History*) - onde fica o histórico dos comandos executados no *Console*. Estes comandos podem ser pesquisados nesta guia. Os comandos são exibidos em ordem (mais recentes na parte inferior) e agrupados por bloco de tempo.
3. **Conexões** (*Connections*) - mostra todas as conexões feitas com fontes de dados suportadas e permite saber quais conexões estão ativas no momento. O *RStudio* suporta múltiplas conexões de banco de dados simultâneas.

4. **Tutorial** - a partir da versão 1.3, o *R Script* ganhou um painel Tutorial dedicado, usado para executar tutoriais que ajudarão você a aprender e dominar a linguagem de programação *R*. Na primeira vez que se abre o programa, clicando nesta aba, o *RStudio* solicita que seja instalado o pacote `learnr` (**Figura 4-6**). Isto permite acesso a vários tutoriais úteis que merecem ser explorados

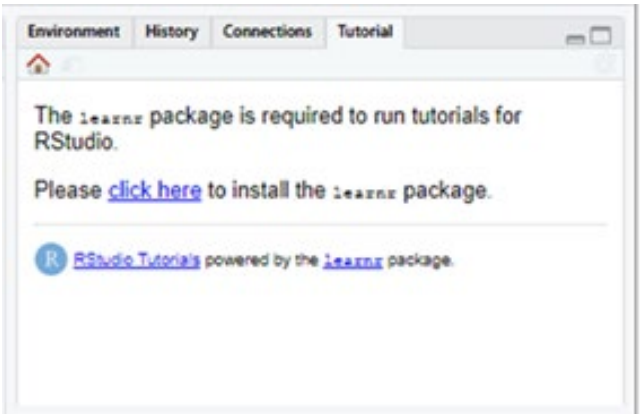

Figura 4-6 Tutoriais no RStudio

Arquivos, Gráficos, Pacotes, Ajuda e Apresentação

No lado direito, abaixo, existem outras abas muito úteis (**Figura 4-2**, em amarelo):

1. **Arquivos (Files)** – esta guia dá acesso ao diretório onde se encontram os seus arquivos. Um bom recurso do painel *Files* é que se pode usá-lo para definir seu diretório de trabalho. Para isso, clique em *More* e depois em *Set As Working Directory*.
2. **Gráficos (Plots)** – local onde ficam os gráficos gerados. Existem botões para abrir o gráfico em uma janela separada e exportar o gráfico como um *.pdf* ou *.jpeg*.
3. **Pacotes (Packages)** – mostra uma lista de todos os pacotes *R* instalados no seu computador e indica se eles estão atualmente carregados ou não. Pacotes que estão sendo executados na sessão atual, estão marcados, enquanto aqueles que estão instalados, mas ainda inativos, estão desmarcados.
4. **Ajuda (Help)** – menu de ajuda para as funções *R*. Você pode digitar o nome de uma função na janela de pesquisa (por exemplo, `histogram` ou usar o `?hist`), no *Console* ou no *R Script*, para procurar ajuda sobre uma função (**Figura 4-7**). A Ajuda no *R Studio* pode também ser acessada no menu *Help* da barra de ferramentas onde existem várias opções. Para complementar, alguns livros são muito uteis, como o *R Cookbook* (53) ou *Using R for Introductory Statistics* (54). No entanto, na maioria das vezes a forma mais prática de conseguir ajuda com uma dúvida específica é a busca em fóruns na internet, como o *Stack Overflow*: https://stackoverflow.com/ .
5. **Apresentação (Presentation)** – é visualizador de apresentações. Nas últimas versões do *Rstudio*, é possível com o *Quarto*, editar um código em *R Markdown* para construir uma apresentação. Não faz parte do objetivo deste livro desenvolver este assunto. É possível encontrar um tutorial em https://quarto.org/docs/get-started/hello/rstudio.html.

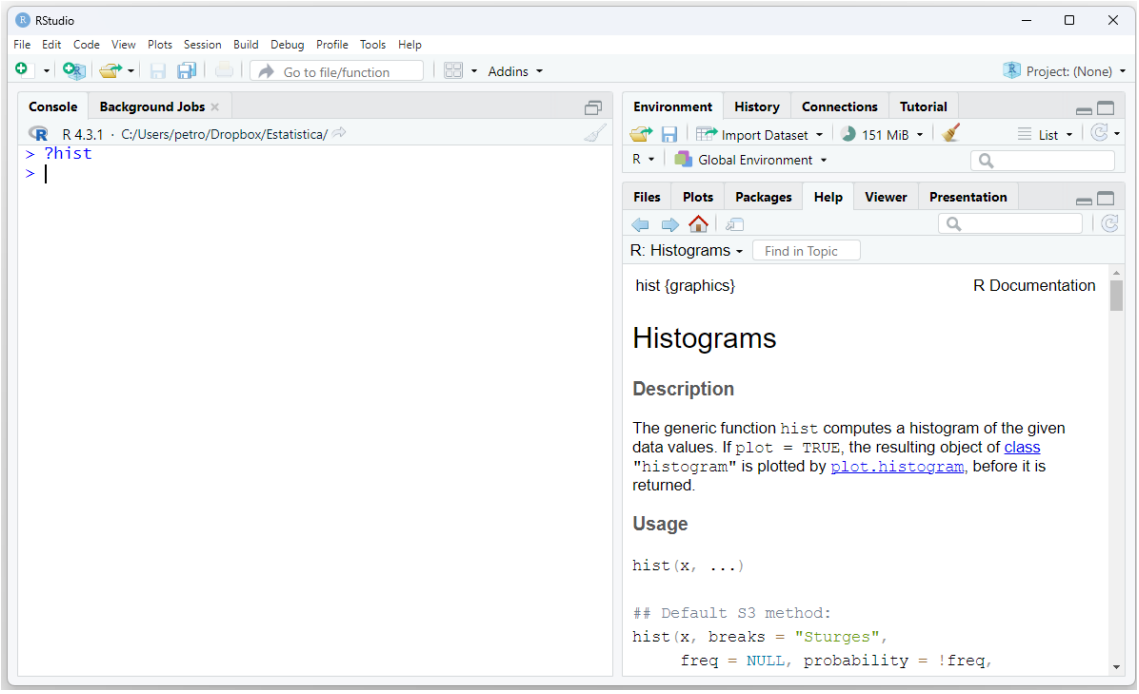

Figura 4-7 Ajuda no RStudio

4.3 Pacotes

Para que o R cumpra a sua função de dialogar com o usuário para realizar análises estatística e construir gráficos, ele necessita ter instalado pacotes.

Quando se instala o R básico, ele vem com vários pacotes que permitem uma grande quantidade de análises. Entretanto, à medida que se utiliza o R, torna-se necessário instalar novos pacotes criados pela comunidade do R. Esses novos pacotes contêm novas funções e novos comandos que aumentarão a funcionalidade do R.

Um pacote é uma coleção de funções, dados e documentação que expande os recursos do R base. O uso dos pacotes é a chave para o uso bem-sucedido do *R*. Eles são instalados à medida que o trabalho com o R exigir.

4.3.1 Repositório de pacotes

Quando se identifica a necessidade de um novo pacote, é fundamental saber onde ele se encontra. O principal repositório de pacotes é o CRAN (*Comprehensible R Archive Network),* já comentado anteriormente. Para acessar este repositório, use o **link** e escolha um espelho (*0-Cloud* ou o mais próximo geograficamente). Depois que o pacote for instalado, ele será mantido em sua biblioteca (*library*) R associada à sua versão principal atual do R. Haverá necessidade de atualizar e reinstalar os pacotes sempre que atualizar uma versão principal do R.

Estando na página do CRAN, no menu, à esquerda, clique em *Packages* . Isto o colocará na página dos *Contributed Packages*, onde a maioria dos pacotes podem ser encontrados em *Table of available packages, sorted by name* . Também é possível clicar em CRAN *Task Views* , onde estão os pacotes separados por tópicos.

4.3.2 Instalação de um novo pacote

Instalar um pacote significa simplesmente baixar o código do pacote em um computador pessoal. Existem duas maneiras principais de instalar novos pacotes. O método mais comum é baixá-los do CRAN, usando a função `install.packages()`. Dentro dos parênteses, como argumento, coloca-se entre aspas (duplas ou simples) o nome do pacote. Como visto, deve-se, de preferência, digitar o comando no *R Script*. Por exemplo, será instalado o pacote `ggplot2` que contém múltiplas funções gráficas como abaixo:

```
install.packages("ggplot2")

library(ggplot2)
```

Para carregar o pacote, isto é, para fazer com que suas funções se tornem ativas para uso na sessão, deve-se usar a função `library()`, como mostrado no comando acima. Se o *RStudio* for fechado e reaberto, o pacote deverá ser novamente ativado. Observe que a função `library()` não requer que o nome do pacote seja digitado entre aspas. Isto acontece porque antes de o pacote ser instalado o R não o reconhece , portanto, há necessidade de indicar o nome (caracteres), para que o R procure na internet, por exemplo, o que ele deve baixar. Já, depois de instalado, o pacote é um objeto conhecido pelo R, logo as aspas não são mais necessárias.

Uma outra maneira de instalar pacotes no R, é usar o botão `Install`, localizado na aba *Packages*, no painel inferior, à direita. Clicando em `Install`, abre-se a caixa de diálogo da **Figura 4-8**. Digitar em *Packages* o nome do pacote `ggplot2` e o *RStudio* completará com opções para achar o pacote. Clicar em `ggplot2` e verifique se *Install dependencies* foi selecionado. A seguir clicar em *Install* e aguardar aparecer no *Console* a mensagem que o pacote foi instalado com sucesso.

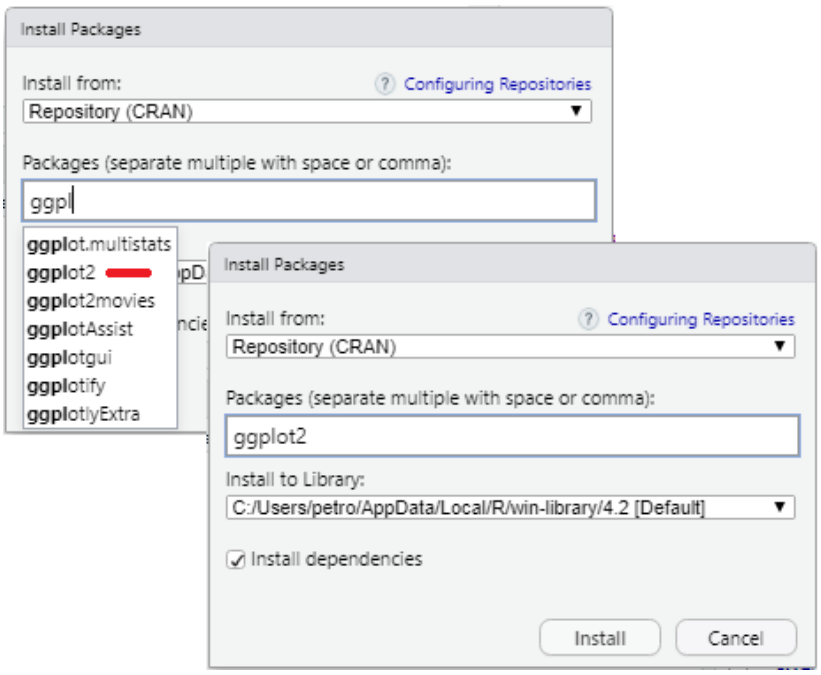

Figura 4-8 Instalação do pacote 'ggplot2' usando a caixa de diálogo 'Install Packages'

4.3.3 Atualização dos pacotes

Periodicamente, há necessidade de atualizar os pacotes instalados. Essa necessidade advém do fato que, com o tempo, os autores de pacotes lançarão novas versões com correções de defeitos e novos recursos e, geralmente, é uma boa ideia manter-se atualizado. Para realizar a atualização proceda da seguinte maneira:

```
# atualiza todos os pacotes disponíveis, solicitando permissão

update.packages()

# atualiza, sem solicitações de permissão/esclarecimento

update.packages(ask = FALSE)

# atualiza um pacote específico

update.packages("ggplot2")
```

4.3.4 Instalando e carregando mais de um pacote

Para carregar mais de um pacote simultaneamente, pode-se usar uma das funções: `libraries()` ou `packages()` do pacote `easypackages`. Em primeiro lugar, instalar e carregar o pacote:

```
install.packages("easypackages")

library(easypackages)
```

Posteriormente, basta usar uma das funções do `easypackages`:

```
libraries("readxl", "dplyr", "ggplot2", "car")
```

Outro pacote que gerencia pacotes do R é o `pacman`. Este pacote tem a função `p_load()` que instala e carrega um ou mais pacotes. Usar esta função, escrevendo o nome dos pacotes sem necessidade de aspas:

```
install.packages("pacman")

library(pacman)

p_load(readxl, dplyr, ggplot2, car)
```

Ou, escrever diretamente:

```
pacman::p_load(readxl, dplyr, ggplot2, car)
```

O pacote `pacman` tem outas funções, entre elas a função `p_update()` que atualiza o pacote e , se usada sem especificar o pacote , atualiza todos. Para saber mais sobre o pacote `pacman`, use a ajuda.

```
p_update(readxl, dplyr, ggplot2, car)
```

4.3.5 Citação de pacotes em publicações

No R existe um comando que mostra como citar o R ou um de seus pacotes. Basta digitar a função `citation()` no *Console* ou no *R Script* e observar a saída. Para um pacote específico, basta colocar o nome do pacote entre aspas, na função.

```
citation()
```

```
## To cite R in publications use:
##
##   R Core Team (2023). _R: A Language and Environment for Statistical
##   Computing_. R Foundation for Statistical Computing, Vienna, Austria.
##   <https://www.R-project.org/>.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2023},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

```
citation ("ggplot2")
```

```
## To cite ggplot2 in publications, please use
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```

4.4 Diretório de trabalho

O diretório de trabalho (**Working Directory**) é uma pasta onde o R lê e salva arquivos. Deve-se criar um diretório de trabalho para a sessão . Para isso, no *RStudio* siga o caminho: *Session* > *Set Working Directory* > *Choose Directory* ou use o atalho *Ctrl* + *Shift* + *H* e escolha o diretório desejado ou crie um novo.

Ao finalizar, aparecerá no *Console* (**Figura 4-9**):

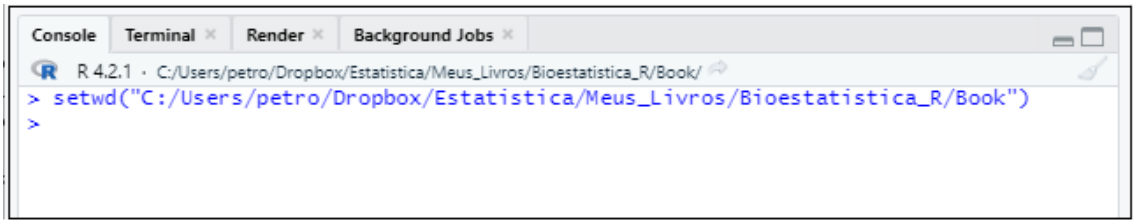

Figura 4-9 Diretório de trabalho

Note que o R usou a função `setwd()` que significa “definir diretório de trabalho”. Também é possível usar esta função diretamente no *R Script* ou no *Console*, digitando conforme o caminho do diretório.

Para saber qual é o diretório de trabalho que está sendo usado pelo R pode-se executar a função `getwd()`. A saída no *Console* mostrará o diretório de trabalho usado, portanto é recomendado que se faça isso no início da sessão para verificar se há ou não necessidade de modificar o diretório.

4.5 Projeto

Uma funcionalidade importante do *RStudio* é a possibilidade de se criar projetos. Um projeto nada mais é do que uma pasta no seu computador. Nessa pasta, estarão todos os arquivos que serão usados ou criados na sua análise.

A principal razão de se utilizar projetos é simplesmente *organização*. Com eles, fica muito mais fácil importar conjunto de dados para dentro do R, criar análises reprodutíveis e compartilhar o trabalho realizado.

Ao se começar uma nova análise, é interessante criar um **Novo Projeto**. Para isso, clicar *File* > *New Project* ou clicar no menu que está na parte superior, à direita, *Project (none)* > *New Project…*. Abrirá a janela da **Figura 4-10**.

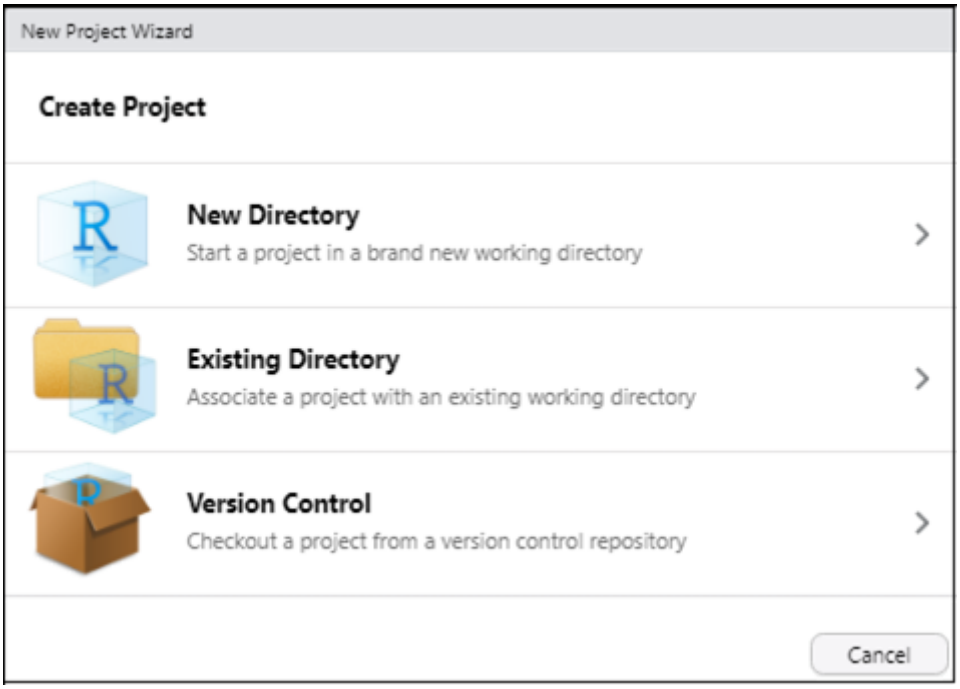

Figura 4-10 Assistente de novo projeto

Clique em *New Directory* para criar um novo diretório. Por exemplo, para as aulas de Bioestatística, pode-se criar um diretório com o nome de *bioestatistica* (evite usar acentos, maiúsculas ou caracteres especiais) ou qualquer outro nome.

Quaisquer documentos Excel ou arquivos de texto associados podem ser salvos nesta nova pasta e facilmente acessados R, indo ao menu *Project (none)* > *Open Project…*. A partir daí, é possível realizar análises de dados ou produzir visualizações com seus dados importados.

Quando um projeto estiver aberto no *RStudio*, o seu nome aparecerá no canto superior direito da tela. Na aba *Files*, aparecerão todos os arquivos contidos no projeto. Quando se clica no nome do projeto, abre um menu que torna muito fácil a navegação pelos projetos existentes. Basta clicar em qualquer um deles para trocar de projeto, isto é, deixar de trabalhar em uma análise e começar a trabalhar em outra.

4.6 O R como calculadora

O R pode ser utilizado para uma série de operações matemáticas desde as mais simples às mais complexas. Para isso, basta digitar no *Console* ou no *R Script*, usando os operadores.

4.6.1 Operadores

Operadores são usados para realizar operações com variáveis e valores

4.6.1.1 Operadores aritméticos

No R, você pode usar operadores aritméticos para realizar operações matemáticas comuns.

```
10 + 5        # Adição
10 - 5        # Subtração
10 * 5        # Multiplicação
10 / 5        # Divisão
10 ^ 5        # Potência
10 %% 3       # Divisão modular (divisão com resto)
10 %/% 3      # Divisão inteiro
```

Ao executar os comandos acima, aparecerá no *Console* o resultado das operações da seguinte maneira:

```
> 10 + 5        # Adição
[1] 15
> 10 - 5        # Subtração
[1] 5
> 10 * 5        # Multiplicação
[1] 50
> 10 / 5        # Divisão
[1] 2
> 10 ^ 5        # Potência
[1] 1e+05
> 10 %% 3       # Divisão modular (divisão com resto)
[1] 1
> 10 %/% 3      # Divisão inteiro
[1] 3
```

Observe que o R repete a operação e coloca em baixo o resultado precedido por [1]. O resultado da operação de exponenciação é exibido como notação científica, onde e+05 significa 10^5.

4.6.1.2 Operadores de atribuição

São usados para comparar dois valores.

```
3 == 3        # Igualdade
3 == 4
3 != 4        # Não igual (diferente)
```

```
> 3 == 3        # Igualdade
[1] TRUE
> 3 == 4
```

```
[1] FALSE
> 3 != 4         # Não igual (diferente)
[1] TRUE
[1] FALSE
```

```
6 > 3          # Maior
3 < 4          # Menor
5 >= 3         # Maior ou igual
4 <= 4         # Menor ou igual
```

```
> 6 > 3           # Maior
[1] TRUE
> 3 < 4           # Menor
[1] TRUE
> 5 >= 3          # Maior ou igual
[1] TRUE
> 4 <= 4          # Menor ou igual
[1] TRUE
```

Observe que, na linguagem R, o sinal de igualdade é escrito com duplo =.

4.6.1.3 Operadores lógicos

Operadores lógicos são usados para combinar declarações condicionais:

```
# Conjunção lógica E, retorna TRUE se ambos os elementos são  verdadeiros

6 == 6 & 7 == 8

2 * 3 && 1 * 6
```

```
> 6 == 6 & 7 == 8
[1] FALSE
>
> 2 * 3 && 1 * 6
[1] TRUE
```

```
# Conjunção lógica OU, retorna TRUE se um dos elementos é verdadeiro

(2 * 2) | sqrt(16)

6 == 6 | 7 == 8
```

```
> (2 * 2) | sqrt(16)
[1] TRUE
>
> 6 == 6 | 7 == 8
[1] TRUE
```

```
# Conjunção lógica NÃO, retorna FALSE se o  elemento é verdadeiro

!6==6

!2==4
```

```
> (2 * 2) | sqrt(16)
[1] TRUE
>
> 6 == 6 | 7 == 8
[1] TRUE
```

4.6.1.4 Outros operadores

```
log (10)           # Logaritmo natural (base e)

log10 (10)         # Logaritmo base 10

sqrt (81)          # Raiz quadrada

abs (3 - 6)        # Resultado absoluto
```

```
> log (10)           # Logaritmo natural (base e)
[1] 2.302585
>
> log10 (10)         # Logaritmo base 10
[1] 1
>
> sqrt (81)          # Raiz quadrada
[1] 9
>
> abs (3 - 6)        # Resultado absoluto
[1] 3
```

4.7 Objetos

O *R* permite salvar valores dentro de um *objeto*. Os objetos são criados utilizando o *operador de atribuição* (<–). Para digitar este operador, basta teclar o sinal *menor que* (<), seguido de *hífen* (–) , sem espaços. Existe um atalho que é pressionar (`Alt + –`) . O símbolo = pode ser usado no lugar de <–, mas não é recomendado.

Objeto é um pequeno espaço na memória do computador onde o *R* armazenará um valor ou o resultado de um comando, utilizando um nome arbitrariamente definido. Tudo criado pelo *R* pode se constituir em um objeto, por exemplo: uma variável, uma operação aritmética, um gráfico, uma matriz ou um modelo estatístico. Através de um objeto torna-se simples acessar os dados armazenados na memória. Ao criar um objeto, se faz uma declaração. Isto significa que se está afirmando que uma determinada operação aritmética irá, agora, tornar-se um objeto que irá armazenar um determinado valor. As declarações são feitas uma em cada linha do *R Script*.

Os objetos devem receber um nome e é obrigatório que ele comece por uma letra (ou um ponto) e não é permitido o uso do hífen. Pode-se usar o ponto e *underlines* para separar palavras. Deve ser evitado o uso de nomes que sejam de objetos do sistema, ou outros objetos já criados, funções ou constantes. Por exemplo, não deve ser utilizado:

```
c, q, r, s, t, C, D, F, I, T, diff, exp, log, mean, pi
, range, rank, var, NA, NaN, NULL, FALSE, TRUE, break, else, if, break
, function, in, while
```

que devem ser reservados, pois têm significados especiais.

Quando se usa um objeto com o nome `pi`, ele assumirá outro valor diferente de 3,141593. Preservando este nome, toda vez que usarmos a palavra `pi`, o R assume o valor pré-estabelecido. Além disso, o R faz a diferença entre letras maiúsculas e minúsculas. Ou seja, `soma` é um objeto diferente de `Soma` e ambos são diferentes de `SOMA`.

Para exibir o conteúdo de um objeto, basta digitar seu nome no R *Script* ou no *Console* e executar. Em análises mais extensas, verificar se já há um objeto com o mesmo nome, pois seus valores serão substituídos ao executar o novo objeto. Para saber se já existe um objeto com o nome definido, digite as primeiras letras do objeto criado e o *R Studio* listará, usando a sua função de autocompletar, tudo que começar com essas letras no arquivo. Assim ficará fácil verificar se já existe um objeto com o nome desejado.

No comando abaixo, é criado um objeto que receberá a soma de dez números, utilizando a função `sum()`. O objeto foi denominado de `soma`. Para exibir o valor contido no objeto `soma`, é necessário digitar `soma` no *R Script* ou *Console* e executar:

```
soma <- sum (2, 3, 12, 15, 21, 4, 8, 7, 13, 21)

soma
```

```
> soma <- sum (2, 3, 12, 15, 21, 4, 8, 7, 13, 21)
> soma
[1] 106
```

4.8 Funções

A função é uma orientação ao R para que ele execute algum procedimento específico, por isso, em geral, têm nomes sugestivos do que elas realizam. Nas seções anteriores foram utilizadas algumas funções – sum(), setwd(), log10(), etc. Outro exemplo é a função `mean()` que entrega a média aritmética de uma série de números colocados entre parênteses. O resultado, como regra geral, deve ser colocado em um objeto que será armazenado na memória do computador.

Esta série de números pode antes ser armazenada por um objeto, nomeado `dados` e, posteriormente, se usa a função `mean()` com este objeto `dados`. O resultado, exibido no *Console*, será recebido por outro objeto `media_dados` que será colocado na memória do computador.

```
dados <- c(3, 5, 7, 9, 6, 7)
media_dados <- mean(dados)
media_dados
```

```
> media_dados
[1] 6.166667
```

As *funções* podem ser criadas pelo pesquisador, de acordo com as suas necessidades. Entretanto, na maioria das vezes, elas são encontradas prontas, fazendo parte de um pacote. Pacotes contêm muitas funções que para serem executadas necessitam que estes estejam instalados e carregados. As funções para exercerem a sua ação devem receber dentro delas (entre parênteses) os *argumentos* que elas exigem. Os argumentos de uma função são sempre separados por vírgulas.

Para se saber quais argumentos necessários para uma determinada função basta consultar a ajuda, onde se encontrará a documentação da mesma. Para isso basta digitar no *Console*, no caso da função `mean()`, `help(mean)` ou `?mean`:

```
help(mean)
```

O resultado deste comando aparecerá na aba *Help*, na parte inferior, à direita (**Figura 4-11**):

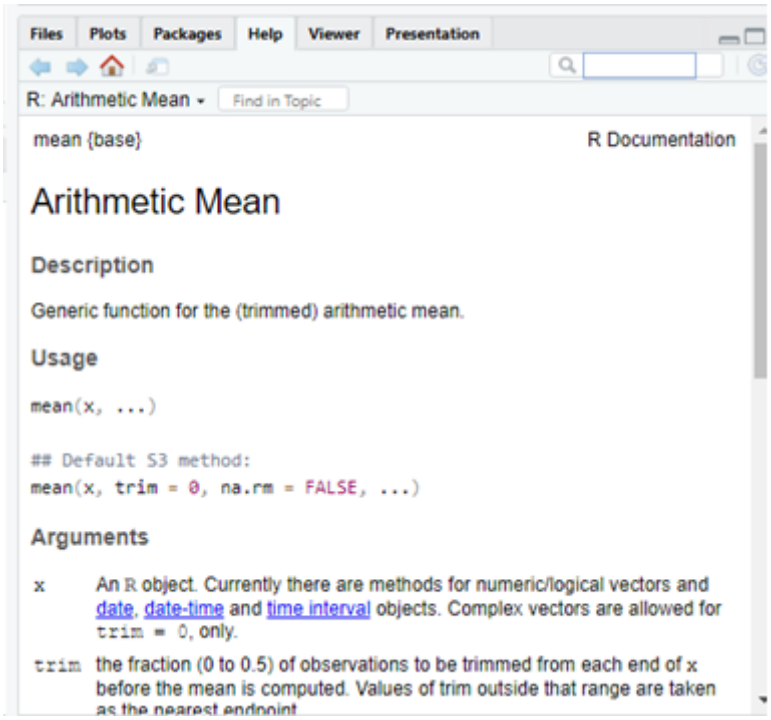

Figura 4-11 Ajuda para Média Aritmética.

Os principais argumentos da função `mean()` são:

- **x** ⟶ vetor numérico
- **trim** ⟶ fração das observações (varia de 0 a 0,5) extraída de cada extremidade de x para calcular a média aparada
- **na.rm** ⟶ valor lógico (TRUE ou FALSE) que indicam se os valores ausentes (NA) devem ser removidos antes que o cálculo continue

Este último argumento é muito importante quando, na sequência de valores existe algum valor não informado ou inexistente. No R, eles são denominados de valores ausentes (*missing values*) e denotados de **NA** (*Not Available*).

Por exemplo, em uma coleta de uma série de valores, correspondentes ao peso de 15 recém-nascidos, havendo a "falta" de um dos registros, ao calcular a média com a função `mean()`, ela retornará NA.

```
pesoRN <- c (3340,3345,3750,3650,3220,4070,NA,3970,3060,3180,2865,2815,
             3245,2051,2630)
mean (pesoRN)
```

```
> mean (pesoRN)
[1] NA
```

Colocando o argumento `na.rm = TRUE`, para remover os valores faltantes, a função retornará a média aritmética sem este valor:

```
mean (pesoRN, na.rm = TRUE)
```

```
> mean (pesoRN, na.rm = TRUE)
[1] 3227.929
```

4.8.1 Criando funções

No R, é possível criar funções pessoais que podem simplificar um código e, eventualmente, diminuir o tempo de execução das análises.

Fórmula geral: as funções têm uma fórmula geral:

> nome_da_função <– function (x) {transformar x}

Por exemplo, a área de um círculo é igual a $\pi \times r^2$, onde r é o raio do círculo. Para calcular a área do círculo, pode-se criar uma função que faça este trabalho:

```
area.circ <- function(r){
 area <- pi*r^2
 return(area)
}
```

Após executar a função, é possível calcular a área de um círculo, conhecendo-se o seu raio:

```
r = 5
area.circ(5)
```

```
> area.circ(5)
[1] 78.53982
```

Pode-se criar qualquer função. Por exemplo, O Indice de Massa Corporal é igual ao peso (kg) dividido pela altura^2, em metros. Uma função para fazer este cálculo é:

```
imc <- function(peso, altura){
  res <- peso/altura^2
  return(res)
}
```

Logo, o IMC de um indivíduo que tenha 67 kg e 1,7 m é:

```
imc(67, 1.70)
```

```
> imc(67, 1.70)
[1] 23.18339
```

4.8.1.1 Ativação de uma função criada

Recomenda-se salvar as funções criadas em uma pasta específica em seu computador. Para ativar esta função, usa-se a função nativa `source()`. O argumento desta função é o caminho (no exemplo, é o diretório do autor) onde se encontra a função buscada, por exemplo, a função `imc()` criada acima:

```
source('C:/Users/petro/Dropbox/Estatistica/Bioestatistica_usando_R/Funcoes/imc.R')
```

4.9 Classes

São os atributos de um objeto e o seu conhecimento é de suma importância. A partir do conhecimento do tipo de classe que as funções sabem o que exatamente fazer com um objeto. Por exemplo, não é possível somar duas letras e se for feita a tentativa de somar “a” e “b”, O R retorna um erro: Error in “a” + “b”: non-numeric argument to binary operator .

No R, os textos são escritos entre aspas simples ou duplas. As aspas servem para diferenciar nomes (objetos, funções, pacotes) de textos (letras e palavras). Os textos são muito comuns em variáveis categóricas e são popularmente chamados de *strings* ou *character*. Além desta classe, o R tem outras classes básicas que são a *numeric* e a *logical*. Um objeto de qualquer uma dessas classes é chamado de *objeto atômico*. Esse nome se deve ao fato de essas classes não se misturarem (55).

Para saber qual o tipo de classe que um objeto pertence, basta usar a função `class()`. Por exemplo:

```
idade <- c(3, 5, 7, 9, 6, 7)
class (idade)
```

```
> class (idade)
[1] "numeric"
```

```
nome <- c("Pedro", "Maria", "Margarida", "Alice", "João", "Luís")
class(nome)
```

```
> class(nome)
[1] "character"
```

4.10 Vetores

Um **vetor** é uma variável com um ou mais valores do mesmo tipo. Por exemplo, o número de filhos em 10 famílias foi 4, 5, 3, 2, 2, 1, 2, 1, 3 e 2. O vetor nomeado de `n.filhos` é um objeto numérico de comprimento = 10. A maneira mais fácil de criar um vetor no R é concatenar (ligar) os 10 valores, usando a função concatenar `c()` assim:

```
n.filhos <- c(4, 5, 3, 2, 2, 1, 2, 1, 3, 2)
n.filhos
```

```
> n.filhos
 [1] 4 5 3 2 2 1 2 1 3 2
```

Como os vetores são conjuntos *indexados*, pode-se dizer que cada valor dentro de um vetor tem uma **posição**. Essa posição é dada pela ordem em que os elementos

foram colocados no momento em que o vetor foi criado. Isso nos permite acessar individualmente cada valor de um vetor (55).

Para acessar um determinado valor, basta colocar a posição do mesmo entre colchetes [ ]. Se há interesse em conhecer o número de filhos da quinta família, procede-se da seguinte forma:

```
n.filhos[5]
```

```
> n.filhos[5]
[1] 2
```

Se houver tentativa de acessar um valor inexistente, o R retorna NA.

```
n.filhos[11]
```

```
> n.filhos[11]
[1] NA
```

Se houver necessidade de excluir um dos elementos, basta colocar entre colchetes a posição do mesmo com sinal negativo. Por exemplo, para excluir o valor correspondente a sexta família, usa-se:

```
n.filhos[-6]
```

```
> n.filhos[-6]
[1] 4 5 3 2 2 2 1 3 2
```

Observa-se que o valor 1 foi excluído da série de elementos.

Quando são colocados elementos em um vetor que pertençam a classes diferentes, o R promove o que se denomina de **coerção**, pois o vetor pode ter apenas uma classe de objeto. Dessa forma, as classes mais fortes reprimem as mais fracas. Por exemplo, sempre que for misturado números e texto em um vetor, os números serão considerados como texto:

```
vetor <- c(12, 15, 4, 6, "A", "D")
vetor
```

```
> vetor
[1] "12" "15" "4"  "6"  "A"  "D"
```

Observe que, agora, todos os elementos do vetor passaram a ser textos e, por isso, estão entre aspas.

4.10.1 Tipos de vetores

Dado um vetor, pode-se determinar seu tipo com `typeof()`, ou verificar se é um tipo específico com uma das funções: `is.character()`, `is.double()`, `is.integer()`, `is.logical( )`.

```
n.filhos <- c(4, 5, 3, 2, 2, 1, 2, 1, 3, 2)
typeof(n.filhos)
```

```
> typeof(n.filhos)
[1] "double"
```

A função retorna uma saída de “double” quando o argumento é um valor numérico.

```
is.numeric(n.filhos)
```

```
> is.numeric(n.filhos)
[1] TRUE
```

As expressões do tipo *character* devem aparecer entre aspas duplas ou simples. Os números no *R* são geralmente tratados como objetos numéricos (números reais de dupla precisão). Mesmo números inteiros são tratados como numéricos. Para fazer um número inteiro ser tratado como objeto inteiro, deve-se utilizar a letra L após o número.

```
n.filhos <- c(4L, 5L, 3L, 2L, 2L, 1L, 2L, 1L, 3L, 2L)
typeof(n.filhos)
```

```
> typeof(n.filhos)
[1] "integer"
```

Outros exemplos:

```
nomes <- c('Maria', 'João', 'Manuel', 'Petronio', 'José')
typeof(nomes)
```

```
> typeof(nomes)
[1] "character"
```

```
altura <- c(1.60, 1.78, 1.55, 1.67, 1.69)
typeof(altura)
```

```
> typeof(altura)
[1] "double"
```

4.11 Dataframes

Dataframes são objetos de dados genéricos de R, usados para armazenar os dados tabulares, onde os dados são organizados de maneira lógica em um formato de linha-e-coluna semelhante ao de uma planilha do Excel. O data frame é uma estrutura bidimensional. Estas dimensões podem ser encontradas com a função `dim()`. Os Dataframes podem ser formados com objetos criados previamente, desde que tenham o mesmo comprimento (56).

Abaixo serão criadas algumas variáveis, todas relacionadas ao nascimento de 15 bebês:

```
id <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
pesoRN <- c (3340,3345,3750,3650,3220,4070,3380,3970,3060,3180,
             2865,2815,3245,2051,2630)
compRN <- c (50,48,52,48,50,51,50,51,47,47,47,49,51,50,44)
sexo <- c (2,2,2,1,1,1,2,1,1,1,2,2,1,1,2)
tipoParto <- c (1,1,2,1,2,2,1,2,1,1,1,2,1,1,1)
idadeMae <- c (40,19,26,19,32,24,27,20,21,19,23,36,21,23,23)
```

Tem-se um grupo de variáveis (vetores) isoladas. Seria útil reuni-las em um só objeto, usando a função `data.frame()`. Este novo objeto receberá o nome de `dadosNeonatos`.

```
dadosNeonatos <- data.frame (id,
                             pesoRN,
                             compRN,
                             sexo,
                             tipoParto,
                             idadeMae)
```

Ao ser executado o comando retornará um novo objeto da classe `data.frame`:

```
class (dadosNeonatos)
```

```
> class (dadosNeonatos)
[1] "data.frame"
```

Para acrescentar outra variável no banco de dados `dadosNeonatos`, será criado a um vetor (utiNeo) [3], a situação (ida ou não para UTI) dos 15 recém-nascidos. O símbolo $ indica a união do nome do dataframe com o nome da nova variável `dadosNeonatos$utiNeo` .

```
dadosNeonatos$utiNeo <- c (2,2,2,2,1,2,1,2,2,2,2,1,2,2,2)
```

Para observar o novo banco de dados, pode-se usar a função `str()` do R base. Digitar no *R Script*:

```
str(dadosNeonatos)
```

```
> str(dadosNeonatos)
'data.frame':  15 obs. of  7 variables:
 $ id       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ pesoRN   : num  3340 3345 3750 3650 3220 ...
 $ compRN   : num  50 48 52 48 50 51 50 51 47 47 ...
 $ sexo     : num  2 2 2 1 1 1 2 1 1 1 ...
 $ tipoParto: num  1 1 2 1 2 2 1 2 1 1 ...
 $ idadeMae : num  40 19 26 19 32 24 27 20 21 19 ...
 $ utiNeo   : num  2 2 2 2 1 2 1 2 2 2 ...
```

Observando a saída da função, verifica-se que o dataframe contém 15 linhas e 7 colunas e que todas as variáveis estão como variáveis *numéricas* , mas as variáveis `sexo`, `tipoParto` são variáveis *categóricas*, bem como a variável `utiNeo`, acrescentada depois. Há necessidade de fazer uma transformação dessas variáveis, que será mostrado na próxima seção.

4.12 Fatores

Os fatores, no R, são usados para trabalhar com variáveis categóricas. São variáveis usadas para categorizar e armazenar os dados, tendo um número limitado de valores diferentes.

Um fator armazena os dados como um vetor de valores inteiros. O fator no R também é conhecido como uma variável categórica que armazena valores de dados de *string* e inteiros como níveis. O fator é usado principalmente em modelagem estatística e análise exploratória de dados com *R* (57).

4.12.1 Criando fatores

No dataframe `dadosNeonatos`, criado anteriormente, contém três variáveis (`sexo`, `tipoParto` e `utiNeo`) que estão como variáveis numéricas, mas são variáveis tipicamente categóricas. É possível, desta forma, realizar operações aritméticas com elas. Isto, obviamente, seria um absurdo. Assim, é necessário transformá-las em

[3] A variável criada, `utiNeo`, possui dois níveis: 1 = sim; 2 = não, referente se o bebê foi ou não para a UTI

fatores. Para isso, é usada a função `factor()`, nativa do R. Os principais argumentos desta função são:

x ⟶ vetor numérico

levels ⟶ vetor opcional dos valores que *x* pode assumir

labels ⟶ vetor de caracteres dos rótulos para os níveis, na mesma ordem

ordered ⟶ vetor lógico (TRUE ou FALSE). Se TRUE, os níveis dos fatores são assumidos como ordenados

No exemplo, as variáveis não têm uma ordem lógica, então, o argumento `ordered` não será usado.

```
dadosNeonatos$utiNeo <- factor (dadosNeonatos$utiNeo,
                                levels = c(1,2),
                                labels = c('sim','não'))
dadosNeonatos$tipoParto <- factor(dadosNeonatos$tipoParto,
                                  levels = c(1,2),
                                  labels = c("normal","cesareo"))
dadosNeonatos$sexo <- factor (dadosNeonatos$sexo,
                              levels = c(1,2),
                              labels = c("M","F"))
```

Após a transformação, executa-se novamente a função `str()` para ver como ficou o dataframe:

```
str(dadosNeonatos)
```

```
> str(dadosNeonatos)
'data.frame':  15 obs. of  7 variables:
 $ id       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ pesoRN   : num  3340 3345 3750 3650 3220 ...
 $ compRN   : num  50 48 52 48 50 51 50 51 47 47 ...
 $ sexo     : Factor w/ 2 levels "M","F": 2 2 2 1 1 1 2 1 1 1 ...
 $ tipoParto: Factor w/ 2 levels "normal","cesareo": 1 1 2 1 2 2 ...
 $ idadeMae : num  40 19 26 19 32 24 27 20 21 19 ...
 $ utiNeo   : Factor w/ 2 levels "sim", "não": 2 2 2 2 1 2 1 2 2 2 ...
```

Agora, as três varáveis passaram a ser fatores e as outras mantiveram-se numéricas.

Desta forma, é possível trabalhar com ela fazendo, por exemplo, uma contagem da frequência do tipo de parto, usando a função `table()`:

```
table(dadosNeonatos$tipoParto)
```

```
> table(dadosNeonatos$tipoParto)

 normal cesareo
     10       5
```

Ou seja, aproximadamente 70% dos partos desta amostra são normais.

4.13 Salvando o dataframe criado

O dataframe, criado e modificado anteriormente, pode ser salvo para uso posterior no diretório de trabalho.

Para isso existe a função `save ()`, fornecendo como argumentos o dataframe a ser salvo e o nome do arquivo (`file =`) entre aspas. Por convenção, esta função salva com a extensão `.RData` que deve ser digitada, pois o R não a adiciona automaticamente.

```
save(dadosNeonatos, file = "dadosNeonatos.RData")
```

Este comando colocará o arquivo no diretório de trabalho em uso. Portanto, se o objetivo é salvar em outro local, deve ser informado ao *R* qual o novo diretório.

Para carregar o objeto salvo anteriormente com o comando `save ()`, usa-se a função `load ()`. Se o arquivo a ser lido não estiver no diretório de trabalho da sessão, há necessidade de especificar o caminho até o arquivo:

```
load("dadosNeonatos.RData")
```

Ou, indicando o diretório onde está o arquivo:

```
load("C:/Users/petro/Dropbox/Estatistica/Meus_Livros/Bioestatistica_R/Book/
     dadosNeonatos.RData")
```

É possível salvar em outro tipo de extensão como Excel (`.xlsx`), Valores Separados por Vírgula (`.csv`), etc. O procedimento é o mesmo, mudando a função. Para salvar em uma extensão `.xlsx`, utiliza-se a função `write_xlsx()` do pacote `writexl` (58):

```
writexl::write_xlsx(dadosNeonatos, "dadosNeonatos.xlsx")
```

Para salvar com a extensão `.csv`, usar a função `write.csv()` ou `write.csv2()` que faz parte do pacote `utils`, incluído no R base. A primeira função, usa `"."` para a separação dos decimais e `","` para separar as variáveis; a segunda função usa `","` para os decimais e `";"` para separar as variáveis, convenção do Excel para algumas localidades, como o Brasil (59). Portanto, uma outra maneira de salvar o arquivo é:

```
write.csv2 (dadosNeonatos, "dadosNeonatos.csv")
```

5 Manipulando dados no RStudio

5.1 Importando dados de outros softwares

Quando se estudou os dataframes, observou-se que é possível inserir dados diretamente no R. Entretanto, se o conjunto de dados for muito extenso, torna-se complicado. Desta forma, é melhor importar os dados de outro software, como o Excel, SPSS, etc. A recomendação é que se construa o banco de dados, por exemplo, no Excel, e, depois, exporte o arquivo em um formato que o R reconheça como `.xlsx`, `.csv`, `.sav`.