



# Bioestatística quantitativa aplicada

Edison Capp  
Otto Henrique Nienov  
Organizadores

Caroline Darski  
Charles Francisco Ferreira  
Cristiana Palma Kuhl  
Fernanda Dapper Machado  
Fernanda Vargas Ferreira  
Hellen Meiry Grosskopf Werka  
Johanna Ovalle Diaz  
Marina Petter Rodrigues  
Michele Strelow Moreira  
Nadine de Souza Ziegler  
Paula Barros Terraciano  
Pedro Henrique Comerlato  
Sinara Santos

Universidade Federal do Rio Grande do Sul  
Faculdade de Medicina  
Programa de Pós-Graduação em Ciências da Saúde:  
Ginecologia e Obstetrícia

# Bioestatística Quantitativa Aplicada

Porto Alegre 2020  
UFRGS

U58b Universidade Federal do Rio Grande do Sul. Faculdade de Medicina. Programa de Pós-Graduação em Ciências da Saúde: Ginecologia e Obstetrícia Bioestatística quantitativa aplicada/ Universidade Federal do Rio Grande do Sul; organizadores: Edison Capp e Otto Henrique Nienov – Porto Alegre: UFRGS, 2020.

260p.

ISBN: 978-65-86232-43-1

E-Book: 978-65-86232-44-8

1. Epidemiologia e Bioestatística 2. Estatística 3. SPSS I. Capp, Edison, org. II. Nienov, Otto Henrique, org. III Título.

NLM: WA950

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)  
(Bibliotecária Shirlei Galarça Salort – CRB10/1929)

Endereço:

PPG em Ciências da Saúde: Ginecologia e Obstetrícia

FAMED – UFRGS

Rua Ramiro Barcellos, 2400/2º andar

CEP 900035-003 – Porto Alegre – RS

Telefone: +55 51 3308 5607

E-mail: ppggo@ufrgs.br

Editoração e diagramação: Edison Capp

Capa: Edison Capp, imagens: [www.freepik.com/starline](http://www.freepik.com/starline)

Edison Capp  
Otto Henrique Nienov  
Organizadores

Caroline Darski  
Charles Francisco Ferreira  
Cristiana Palma Kuhl  
Fernanda Dapper Machado  
Fernanda Vargas Ferreira  
Hellen Meiry Grosskopf Werka  
Johanna Ovalle Diaz  
Marina Petter Rodrigues  
Michele Strelow Moreira  
Nadine de Souza Ziegler  
Paula Barros Terraciano  
Pedro Henrique Comerlato  
Sinara Santos

## 2 Guia Básico do Microsoft Office Excel

*Fernanda Dapper Machado*

*Fernanda Vargas Ferreira*

*Edison Capp*

*Otto Henrique Nienov*

A primeira versão do Excel foi lançada em 1985 para o Mac e, em 1987, para o Windows. Posteriormente, com sua incorporação no Microsoft Office, suas versões gradativamente tornaram esta empresa a líder no desenvolvimento desse tipo de *software*. Atualmente, este programa de edição de planilhas eletrônicas tem sido empregado, por exemplo, para organização geral de planilhas para uso doméstico, acadêmico ou empresarial, realização de análises estatísticas ou financeiras e construção e manipulação de tabelas e gráficos.

Uma das primeiras dúvidas acerca da aplicabilidade desse programa se encontra em como “iniciar um banco de dados”. Entretanto, é fundamental identificar previamente como são as características da população ou amostra que são passíveis de representação numérica, obtidas como resultado de medições e contagens. Tais características são denominadas de variáveis e podem ser classificadas em qualitativas e quantitativas.

Variáveis qualitativas são resultado de uma classificação por atributos ou tipos, por exemplo, cor dos olhos. Adicionalmente, tais variáveis se subdividem em a) nominal, que ocorre por meio de “categorias” como sexo (masculino ou feminino) ou estado civil e b) ordinal, quando há uma ordenação, por exemplo, grau de satisfação e escolaridade.

Similarmente, variáveis quantitativas refletem números de uma contagem ou mensuração, como número de filhos e idade. Essas variáveis podem ser classificadas em a) contínuas, se os valores podem assumir qualquer valor em um determinado intervalo (ex.: massa, estatura), e b) discreta, cujos valores podem ser enumerados, por exemplo, número de ligações em um celular e número de discentes de uma disciplina ou em uma escola (página 30, Figura 8).

## Como montar um banco de dados

No banco de dados, atribua um título para cada variável, o qual ficará disposto na primeira linha das colunas. Os dados referentes a cada caso, paciente ou objeto de pesquisa ficarão dispostos nas linhas, na respectiva célula ou casela. Lembre-se, sempre, de destinar uma variável para a identificação do caso, a qual pode ser feita por uma sequência numérica, facilitando a localização de informações no caso de serem identificados equivocados de digitação e também a organização do banco de dados.

É importante montar um “dicionário” do banco de dados para facilitar a identificação das variáveis do estudo em caso de compartilhamento ou quando você for visitar o banco de dados futuramente. O dicionário deve identificar as variáveis, seu significado, a unidade de medida, a definição das siglas utilizadas e os rótulos dos valores das variáveis categóricas.

## Conhecendo um banco de dados

A aplicação prática da estatística está relacionada com o problema exposto pelo estudo, ou seja, pela pergunta da pesquisa. Por isso, é fundamental o planejamento das análises estatísticas que serão realizadas para responder a pergunta antes de iniciarmos as coletas dos dados. A estatística faz parte do planejamento e envolve questões fundamentais, como o cálculo do tamanho da amostra necessário para responder a pergunta de pesquisa. Neste capítulo, vamos aprender a utilizar as funções básicas do Excel para caracterizar uma amostra e/ou descrever os seus resultados. Para isso, utilizaremos um banco de dados existente, o “Banco de dados 1.xls” (disponível em <https://bit.ly/bancosdedados>), que nos fornece dados de alunos de uma disciplina.

O primeiro passo ao se trabalhar com um banco de dados é conhecê-lo. Sempre é necessário ter uma identificação das variáveis dispostas nas colunas. Cada linha corresponde a um caso, usualmente a primeira coluna é reservada para a variável de identificação, onde cada caso receberá um código próprio. O restante das colunas corresponderá às variáveis do estudo (respostas de um questionário ou medidas laboratoriais, por

exemplo). Verifique a natureza das variáveis: se qualitativa ou quantitativa. Para isso, consulte o dicionário, que está disponível em caso de dúvidas na identificação das variáveis.

## Acessando as funções estatísticas

As funções estatísticas do Excel podem ser acessadas no menu "Fórmulas", na aba "Inserir função", onde é possível selecionar a categoria "Estatística", que apresentará todas as funções estatísticas disponíveis no programa. Também é possível acessar através da ferramenta de acesso rápido "Auto soma", onde estão disponíveis ferramentas básicas de soma, média, valor mínimo e máximo. Na aba "Biblioteca de funções", em "Mais funções", pode-se selecionar "Estatística", que dará acesso às funções estatísticas do programa.

## Distribuição de frequência

A frequência corresponde simplesmente à contagem do número de ocorrências, podendo ser expressa pela frequência absoluta ( $n$ ) e frequência relativa ( $n\%$ ). Para exemplificarmos, no "Banco de dados 1", vamos verificar a frequência absoluta e relativa da variável "sexo". A variável "sexo" é uma variável qualitativa dicotômica, ou seja, apresenta duas categorias: sexo masculino e sexo feminino. Será necessário montar uma tabela que indique nas linhas as categorias da variável "sexo" e nas colunas as frequências absoluta e relativa. A tabela resultante será semelhante ao Quadro 1. Na célula ou casela reservada para a frequência absoluta do sexo masculino, utilizaremos a função CONT.SE. Para isso, clique em "Fórmulas", "Estatística" e selecione "CONT.SE". A janela resultante trará uma breve instrução de como utilizar a função.

Quadro 1 - Exemplo para criação de tabela de frequência.

Sexo	Frequência Absoluta ( $n$ )	Frequência relativa ( $n\%$ )
Masculino		
Feminino		
Total		

A função CONT.SE faz a contagem de elementos de acordo com um critério pré-estabelecido. Essa função permite a opção de contar as células vazias e apresenta outras opções, como contar número, valores e vazios. Neste caso, em "Intervalo", é possível selecionar o intervalo de células e contar aquelas que não estão em branco. Selecione todas as células correspondentes à coluna "sexo". Cuide para não selecionar o nome da variável na contagem. Em "Critérios", como queremos selecionar apenas os indivíduos do sexo masculino, inserimos "Masculino". Na própria janela, o programa indica a quantidade de indivíduos do sexo masculino que serão incluídos no intervalo selecionado. Ao clicar em "Ok", a célula selecionada será preenchida com a quantidade (valor absoluto) de indivíduos do sexo masculino no intervalo selecionado. É possível ainda adicionar mais critérios à função, como por exemplo, contar quantos indivíduos do sexo masculino estão no doutorado. Agora, basta repetir o processo para preencher os indivíduos do sexo feminino. Para calcular o total, utilize a função CONTAR.VALORES. Esta função conta o número de células que não estão vazias em um intervalo. Para isso, basta informar o intervalo desejado e o programa retorna com a quantidade de sujeitos daquele intervalo.

Para calcular o número relativo, divida o número absoluto de cada categoria pelo número total da categoria. Essa proporção multiplicada por 100 (%) corresponderá à frequência de cada categoria (sexo masculino e sexo feminino). Para montar a tabela do exercício, é necessário fazer essa etapa para os sexos masculino e feminino. Para obtermos o número relativo total, basta utilizar a função SOMA. Na "Página Inicial" é possível configurar o número de casas após a vírgula, clicando em " $\rightarrow$ .000" ou " $\leftarrow$ .000", conforme necessário. Na interpretação dos resultados, observamos que, dos 115 alunos da disciplina, 82,6% ( $n = 95$ ) são do sexo feminino.

Outra maneira de se obter esses resultados é utilizando a função FREQUÊNCIA, que calcula a frequência absoluta em que os valores ocorrem dentro de um intervalo de valores e, em seguida, retorna uma matriz vertical de números. Ou seja, exerce o mesmo que vimos na função CONT.SE. No entanto, pelo fato de FREQUÊNCIA retornar uma matriz, deve ser inserida como uma fórmula matricial.



Para utilizar essa função, precisamos atribuir códigos para as categorias da variável. Neste exemplo, vamos utilizar os códigos "1" para o sexo masculino e "2" para o sexo feminino. Primeiro, selecione a coluna correspondente à variável "Sexo", esse passo é importante para evitar que a próxima etapa modifique outras variáveis do banco de dados. Vamos utilizar a ferramenta "Localizar e Selecionar", disponível na página inicial. Selecionamos "Substituir" e, no campo "Localizar" escrevemos "Masculino" e, no campo "Substituir por", o código "1" correspondente. Em seguida, clique em "Substituir tudo". O programa sinaliza a quantidade de substituições realizadas. Em seguida, realize a mesma etapa substituindo a categoria "Feminino". Agora, na coluna da variável "Sexo", você verá apenas os códigos 1 e 2 atribuídos aos sexos masculino e feminino, respectivamente.

Ao realizar a substituição, na tabela em que utilizamos a função CONT.SE, os valores calculados serão alterados. Isso ocorre porque o critério apresentado na função já não existe mais. Podemos inserir na função CONT.SE os novos critérios numéricos.

Para utilizar a função FREQUÊNCIA, devemos criar uma tabela similar à anterior mas com os códigos, 1 e 2. Na célula ao lado do código 1 insira a função FREQUÊNCIA. Na janela de argumentos da função, o primeiro argumento que a função FREQUÊNCIA nos pede é a "Matriz\_dados". Neste caso, selecione o intervalo de dados (toda a coluna da variável "sexo", exceto a primeira célula que contém o nome da variável). Na "Matriz\_bin" será necessário informar o intervalo referente a coluna com os códigos digitados, portanto selecione as células em que foram inseridos os códigos 1 e 2. A resposta somente retornará o equivalente ao código 1. Para atribuir o valor do código 2, seleciona-se o intervalo de células onde queremos exibir os resultados e, na barra de fórmulas, pressione as teclas de atalho "Ctrl + Shift + Enter". Isso irá retornar o valor do código 2, portanto, teremos os valores das duas categorias.

A análise descritiva permite ao pesquisador uma melhor compreensão dos dados por meio de medidas de resumo, identificando tendências, variabilidade e valores extremos. É importante para conhecermos as principais medidas estatísticas.

Para exemplificar, vamos realizar a análise descritiva de medidas de tendência central e de dispersão da variável quantitativa "Idade". Para isso, construa uma tabela para a variável com os dados de média (função MEDIA), mediana (função MED), moda (função MODO.UNICO), valor mínimo (função MÍNIMO), valor máximo (função MÁXIMO), desvio padrão (função DESVPAD.A), variância (função VAR.A), quartis (função QUARTIL.INC) e percentis (função PERCENTIL.INC). Para preencher esta tabela, utilize as respectivas funções. Para as funções QUARTIL.INC e PERCENTIL.INC também é necessário informar os quartis ou percentis, incluindo os valores extremos (mínimo e máximo). No caso dos quartis, como abordado no capítulo anterior, dividimos os dados ordenados em quatro partes iguais; assim, temos: valor mínimo = 0, primeiro quartil (25%) = 1, valor mediano (50%) = 2, terceiro quartil (75%) = 3 e, valor máximo = 4. Para os percentis, dividimos os dados ordenados em k percentil de valores em um intervalo, em que k está no intervalo de 0 a 1. Ou seja, se determinarmos k como sendo 0,33, teremos um percentil de 33%. É muito importante observar os valores mínimo e máximo, comparar média e mediana, para verificar se não há inconsistências nos dados e, assim, identificar erros de digitação no banco de dados. Na interpretação dos dados, para variáveis com distribuição normal, apresenta-se os resultados de média e desvio padrão e, para as variáveis com distribuição não-normal, apresenta-se os dados de mediana e valores mínimo e máximo ou quartis.

Existe uma forma subjetiva para avaliar a distribuição dos dados comparando-se os valores de média e mediana. Valores muito próximos ou iguais de média e mediana indicam uma tendência central, ou seja, distribuição normal. No caso do exercício proposto, os valores de média e mediana são próximos, porém não pode-se afirmar a normalidade sem um teste estatístico adequado, como o de Shapiro-Wilk.

No banco de dados do exercício, podemos observar que há uma célula sem valor para a variável "Idade", ou seja, com valor omisso. Esse valor também pode ser chamado de "*missing value*" ou perda de dado. Quando há valores faltantes é importante mencionar o número de dados analisados na interpretação dos dados nas tabelas e/ou nos gráficos. Isso pode ser sinalizado ao lado do nome da variável na tabela com um símbolo (\*) ou

indicando o "n", ou no rótulo da tabela. O Excel não sinaliza valores faltantes utilizando as funções descritas.

Há ainda outra opção para a análise de dados disponível no menu "Dados", na "Análise de dados". Esta ferramenta nem sempre está disponível no Excel. Para disponibilizar a ferramenta, basta clicar no menu "Arquivo", em "Opções", selecione "Suplementos", na opção "Gerenciar" marque a opção "Suplementos do Excel" e clique em "Ir". Na janela "Suplementos", selecione "Ferramentas de Análise" e clique "Ok".

A opção "Análise de dados" resulta em uma janela onde seleciona-se a opção "Estatística Descritiva" e, em seguida, clique em "Ok". Na janela "Estatística descritiva", no campo "Intervalo de dados", é necessário selecionar os dados referentes à variável de interesse, no caso, "Idade". Atenção! Ao utilizar essa função, devemos selecionar inclusive a célula com o nome da variável e marcar a opção "Rótulo na primeira linha". Pode-se optar por exportar os dados para uma nova planilha ou para uma célula selecionada ("Intervalo de saída"). Selecione as opções de "Resumo Estatístico" e "Nível de confiabilidade para média". O Excel assume a distribuição normal dos dados e só calcula nível de confiabilidade para média, sendo essa uma das limitações do uso do programa para execução de análises estatísticas. Ao clicar em "Ok", teremos uma tabela com a análise descritiva da variável "Idade". No item "Contagem", o Excel vai retornar apenas o número de células com valores computados na análise, facilitando a identificação de dados faltantes.

Utilizando-se essa opção, o Excel apresenta também dados de curtose e assimetria, que referem-se à distribuição da variável. Se os valores de assimetria forem próximos a zero, a distribuição é aproximadamente simétrica ou normal. A curtose refere-se à nitidez do pico da curva central.

## Gráficos

O Excel oferece suporte para vários tipos de gráficos com a finalidade de ajudar a exibir os resultados. A representação gráfica deve obedecer requisitos fundamentais para ser realmente útil:

- Simplicidade: o gráfico deve ser destituído de detalhes, bem como de traços desnecessários;
- Clareza: deve possibilitar uma correta interpretação dos valores representativos da amostra em estudo;
- Veracidade: deve expressar a verdade sobre a amostra em estudo.

No menu "Inserir", na opção "Gráficos", é possível criar diferentes modelos de gráfico. Como exercício, vamos criar um gráfico com os resultados de frequência da variável "Sexo", que analisamos anteriormente. Para isso, selecione as células com as categorias "Masculino" e "Feminino", bem como as células com os valores de frequência absoluta obtidos na tabela 1. Clique na opção de "Gráfico de Pizza", e selecione a primeira opção, gráfico 2D. Esse processo já resultará no gráfico, onde é possível editar detalhes como título do gráfico, alterar as cores, inserir o rótulo dos dados (ao clicar com o botão direito sobre a área do gráfico, clique em "Adicionar Rótulo de Dados". Clicando com o botão direito sobre o valor absoluto do rótulo inserido, nas "Opções de rótulo", é possível inserir o valor relativo). O programa também sugere modelos prontos de gráficos. Naturalmente, esses dados podem ser apresentados através de outros tipos de gráficos, como em barras, colunas, como visto no capítulo anterior.

O Excel ainda tem disponível uma infinidade de outras opções, como correlação, teste t, ANOVA, regressão. Lembrando que o Excel não é um software de estatística e sim de planilha, portanto pode haver discrepância ou variação nos resultados, diminuindo a confiabilidade dos dados obtidos. Comparativamente, o Excel fornece uma redução na redundância de dados, enquanto o SPSS é mais rápido e oferece um desempenho confiável. O Excel não fornece a construção de função, ou seja, precisa de muito trabalho para executar uma tarefa específica. O SPSS possui ferramentas de manipulação de dados integradas para gravação e transformação de variáveis.

## Viés em Pesquisa

Para falar de viés, precisamos falar da validade de um estudo, que pode ser interna ou externa. A validade interna reflete as conclusões de uma investigação em relação à amostra estudada. Já a validade externa é relativa à extrapolação dessas conclusões para a população de onde a amostra foi retirada ou, até mesmo, para outras populações.

A validade externa gira em torno da questão da capacidade de generalização dos resultados da pesquisa se manterem verdadeiros em outras situações ou para outras populações. Ou seja, as conclusões do estudo podem ser generalizadas para a população de interesse? A amostra representa essa população?

A validade interna reflete a qualidade do delineamento, da condução e da análise da pesquisa. Pode ser definida como a ausência de erro sistemático nos achados do estudo. Devemos nos perguntar até que ponto os resultados do estudo estão distorcidos em decorrência de erros metodológicos na concepção do estudo?

### *Tipos de erro*

**Erro Aleatório:** trata-se de uma divergência entre o valor da amostra, que é medida pelo estudo, e o valor verdadeiro, encontrado na população. Este tipo de erro ocorre exclusivamente em decorrência do acaso, diminuindo a precisão (confiabilidade) do estudo e tendendo a diminuir a medida de associação. O erro aleatório pode ser consequência da variação biológica natural, de erro de amostragem e/ou erros de medida. O erro de amostragem decorre geralmente da falta de representatividade da amostra, que não contempla toda a variabilidade da população. A melhor forma de reduzir o erro de amostragem é aumentar o tamanho amostral. De fato, não há como garantir a eliminação do erro aleatório, pois os estudos geralmente são conduzidos com uma pequena amostra da população. Na estatística, uma forma de se evitar esse erro é computar a significância estatística e os intervalos de confiança.

**Erro Sistemático:** trata-se da diferença sistemática entre a medida real de efeito e a medida obtida no estudo. Decorre de problemas no delineamento, planejamento, condução ou análise

do estudo, não podendo ser corrigido com aumento do tamanho de amostra, nem durante a análise dos dados. Pode ser evitado na elaboração do estudo, corrigindo o processo de seleção da amostra e coleta das informações. Do ponto de vista estatístico, viés é quando o valor médio da medida de associação obtido a partir de um número infinito de estudos não corresponde ao valor verdadeiro de associação. Em outras palavras, viés refere-se à diferença entre a média encontrada de uma estimativa computada após múltiplas análises randômicas e o valor verdadeiro do parâmetro a ser estimado. Do ponto de vista epidemiológico, é o desvio dos resultados ou inferências em relação à verdade, ou o processo que leva a esse desvio. Se refere a qualquer tendência na coleta, análise, interpretação, publicação ou revisão de dados que possa conduzir a conclusões sistematicamente diferentes da verdade.

### ***Viés de seleção***

Ocorre quando há uma diferença sistemática no momento da seleção ou alocação dos participantes da pesquisa, resultando em diferenças entre os indivíduos alocados e não alocados para o estudo. Por exemplo, quando os grupos de comparação não provêm da mesma população-base do estudo, impactando na medida de associação entre exposição e desfecho, ou seja, são grupos de sujeitos não comparáveis.

O viés de seleção pode ocorrer em todos os tipos de delineamentos epidemiológicos, mas é uma grande preocupação nos estudos do tipo caso-controle, quando a seleção de casos e de controles é baseada em critérios diferentes e estes estão relacionados à exposição. De uma forma geral, este viés é minimizado em estudos prospectivos, pois a seleção dos indivíduos expostos e não-expostos é feita antes do registro da resposta.

A seletividade da amostra pode render resultados distorcidos porque a natureza das relações entre variáveis nos casos incluídos na amostra não é necessariamente igual às relações que observaríamos em uma amostra aleatória.

Pode ser causada pela forma de randomização (em ensaios clínicos randomizados), de seleção ou admissão, pela

perda de seguimento ou não-resposta dos participantes. Nestes dois últimos casos, a preocupação é a probabilidade de que as perdas possam ter ocorrido de forma seletiva e, portanto, seja relacionada à exposição, à resposta ou ambos. Pode-se controlar esse viés através de planejamento e amostragem aleatória (cada indivíduo tem probabilidade conhecida e diferente de zero de fazer parte da amostra). Pode-se ter diferentes tipos de viés de seleção, que estão resumidos no Quadro 2.

Nos estudos de coorte dificilmente ocorre viés de seleção, pois os grupos são determinados de acordo com o fator de exposição e não há desfecho algum no momento da seleção. No entanto, é comum ocorrerem vieses do tipo de perda de seguimento. Alguns indivíduos expostos a determinado fator de risco em um estudo no qual estes seriam acompanhados após a alta hospitalar podem evoluir a óbito, não comparecendo ao seguimento. Essa perda está relacionada tanto com exposição quanto com desfecho.

### ***Viés de informação (ou classificação)***

O viés de informação em estudos de coorte é mais frequente no momento de aferir o desfecho. Se os investigadores conhecem a exposição, podem ficar sugestionados ao aferir o desfecho. Erros na mensuração da exposição ou do desfecho que ocorrem no momento da coleta de informações podem distorcer a classificação tanto da exposição, quanto do desfecho nos sujeitos da pesquisa, em outras palavras, os métodos utilizados não medem corretamente o desfecho ou exposição. Pode ocorrer através de uso de meios inadequados para obter informações.

Deriva de variação individual, variação entre observadores, instrumentos de medição ou erro de aferição ou calibração. Pode ser controlado através da padronização do método, treinamento dos avaliadores e utilização de instrumento de pesquisa adequado. Se sua ocorrência for igual entre os grupos é chamado de viés não-diferencial e pode ser responsável pela discrepância entre resultados de diferentes estudos epidemiológicos. Existem vários tipos de viés da informação, resumidos no Quadro 3.

Quadro 2 - Tipos de viés de seleção (adaptado de Bonita e cols., 2010).

Viés de Seleção	Definição
Alocação	O investigador escolhe um método não-randomizado para distribuir sujeitos nos grupos de estudo.
Detecção	Resultado da falha em detectar um caso da doença, um possível fator causal ou um efeito de interesse.
Participação	Quando fatores afetam a participação dos sujeitos da pesquisa, como exclusões inadvertidas ou recusa em participar.
Amostragem	A amostra não representa o espectro de características da população alvo. Alguns indivíduos têm maior chance de serem selecionados. Por exemplo, um estudo sobre desnutrição em áreas rurais, onde os investigadores evitam as moradias isoladas e examinam apenas os moradores de regiões próximas à estrada.
Trabalhador Saudável/ Auto-seleção/ Afiliação	Indivíduos com determinadas características (ligada à doença ou exposição) têm maior probabilidade de entrar no estudo. Pessoas empregadas são relativamente mais saudáveis que as desempregadas, aposentadas ou incapacitadas. Provavelmente menos pessoas empregadas se candidatam a participar de um estudo.
Não-respondentes	Pode ser encontrado em estudos transversais, onde os indivíduos não localizados apresentam tipicamente maiores prevalências da maioria das doenças do que os que são entrevistados. Em estudos de caso-controle, por exemplo, pode ocorrer se a taxa de não-resposta for diferente entre casos e controles e se a não resposta estiver associada com a exposição.
Perdas de acompanhamento/ seguimento	Pode ocorrer em estudos de coorte ou de intervenção. Por exemplo, os indivíduos que deixaram de ser acompanhados podem ser diferentes daqueles efetivamente estudados.
Hospitalização/ Admissão/ Diagnóstico (viés de Berkson)	Pacientes com determinadas características podem ser mais frequentemente internados que outros sem tal característica com a mesma doença.
Prevalência/ Incidência	São excluídos os sujeitos curados e falecidos (prevalência) ou os sujeitos com coorte de início (incidência). Também chamado de viés de sobrevivência seletiva ou de Neymann.
Sobrevida Seletiva	Após o diagnóstico, o sujeito muda os hábitos associados à doença. Qualquer característica associada com a duração da doença afetará os resultados em um estudo de casos prevalentes.
Viés de sobrevivência	Refere-se às situações nas quais você procura pelos erros analisando os acertos. O termo surgiu na Segunda Guerra Mundial, quando o matemático Wald resolveu um problema de blindagem na fuselagem dos aviões. Inicialmente, os locais de blindagem eram decididos pela frequência dos locais de marcas de tiros dos aviões que retornavam à base. Wald entendeu que, se os aviões com marcas de tiros naqueles locais retornavam, era porque os outros locais não atingidos que precisariam do reforço.



Quadro 3 - Tipos de viés de Informação (adaptado de Bonita e cols., 2010).

Viés de Informação	Definição
Recordação/Memória	Quando o estudo é conduzido de forma retrospectiva, os casos podem lembrar-se melhor de sua história de exposição do que os indivíduos sem a carga da doença. Por exemplo, pacientes com refluxo gástrico podem se lembrar melhor de certas características da dieta no passado do que indivíduos saudáveis.
Registro	Indivíduos com doenças mais graves tendem a ter registros mais completos sobre exposições, o que pode resultar em medidas errôneas de associação.
Aferição/ Mensuração/ Instrumento	Ocorre quando os métodos de aferição (instrumentos) empregados diferem entre os grupos, favorecendo resultados inadequados a um subgrupo de pacientes. Pode ocorrer por diferenças no método, no preparo dos observadores, medidas basais ou dados de seguimento, falta de validade dos instrumentos de coleta de dados, subjetividade de escalas de medidas ou uso de técnicas inadequadas.
Publicação	Tendência de que resultados publicados sejam sistematicamente diferentes da realidade. A não publicação dos resultados pode ser por decisão do autor, do financiador do estudo ou de editores por não submeterem ou aceitarem trabalhos com resultados negativos ou desfavoráveis.
Deteção/ Diagnóstico	Características ligadas à exposição podem aumentar a chance de que uma doença seja detectada. Por exemplo, mulheres em uso de anticoncepcionais podem fazer exame citopatológico com mais frequência, portanto têm maior risco de deteção de alguma alteração citopatológica.
Entrevistador/ Observador	Quando o entrevistador coleta de modo diferente as informações de doentes e saudáveis ou expostos e não-expostos. Ou ainda, quando há direcionamento da resposta do participante. Pode ser contornado com o cegamento dos investigadores.
Duração	Quando os casos mais moderados e mais indolentes de uma doença são detectados desproporcionalmente em programas populacionais de rastreamento.
Exame tardio	Ocorre quando casos moderados de uma doença que progride lentamente são preferencialmente detectados em um inquérito.
Espectro de doença	Quando somente um espectro limitado da doença é incluído.
Tempo Ganho	Ocorre quando o rastreamento detecta uma doença mais precocemente na sua história natural do que teria acontecido de outra maneira.
Extensão	Ocorre quando o espectro completo de um tumor em particular é composto de cânceres que variam de agressivos a indolentes.

## ***Viés de confirmação***

É a tendência de que, uma vez adotada uma hipótese pelo pesquisador, este busque exemplos que a confirme. É a preferência que temos por confirmar nossas hipóteses a ter de refutá-las. Pode levar à pesquisa tendenciosa, polarização de opiniões e interpretação tendenciosa.

## **Fator de confundimento ou de confusão**

Acontece quando a associação observada ocorre devido a um terceiro fator (variável de confusão, podendo ser um fator de risco ou de proteção) que se relaciona com a variável de exposição e com o desfecho, mas não faz parte da cadeia causal da doença (Figura 1).

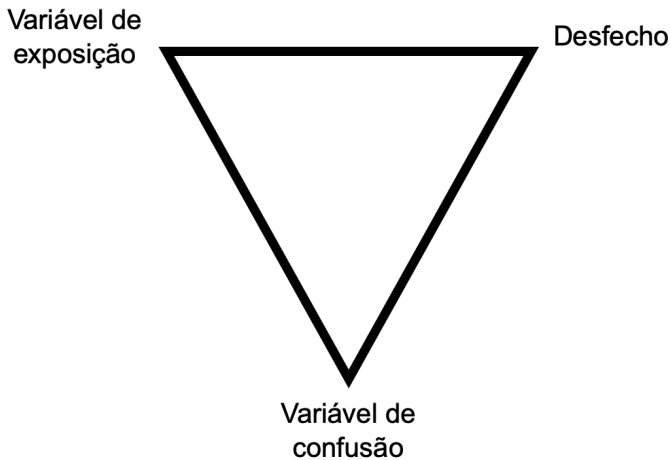


Figura 1. Fator de confusão se relaciona com a variável de exposição e com o desfecho, mas não faz parte da cadeia causal da doença.

Pode também ser definido como a existência de outra exposição que esteja associada com o desfecho. Ocorre devido a uma diferença não aleatória na distribuição dos fatores de risco entre

os grupos, ou seja, a variável de confundimento está distribuída desigualmente entre os grupos comparados. Isso pode alterar a estimativa de risco (subestimar ou superestimar), gerar falso-negativo e/ou falso-positivo e, inclusive, inverter a direção de associação.

O fator de confusão não é considerado um viés, pois não resulta de erro sistemático relacionado ao delineamento de pesquisa. Pode ser controlado tanto no delineamento quanto na análise estatística, através de identificação prévia, presença de grupo controle, randomização, pareamento ou emparelhamento, restrição de categorias, critérios de inclusão e exclusão, estratificação e análise multivariável.

## Referências

Bonita R, Beaglehole R, Kjellström T. Epidemiologia básica. 2. ed. São Paulo: Santos, 2010. 213p.

Guerrero, H. Excel Data Analysis: Modeling and Simulation. 2. ed. Springer Nature Switzerland AG. 358 p.

Guia básico/intermediário. PET Civil UFRGS, 2018, Disponível em: <https://www.ufrgs.br/petcivil/wp-content/uploads/2018/06/Apostila-Excel-2018-1-site.pdf>.

Medronho RA, Bloch KV, Luiz RR, Werneck GL. Epidemiologia. 2. ed. São Paulo: Atheneu, 2008. 676p.

Schoenbach VJ, Rosamond WD. Understanding the Fundamentals of Epidemiology an evolving text. Chapel Hill, North Carolina. 2000. 584 p.

## Exercícios sugeridos

1. No "Banco de dados 1.xls", qual a frequência do nível de pós-graduação entre os alunos da disciplina? Interprete os resultados e construa o gráfico (pizza, barras ou retângulo).
2. No "Banco de dados 2.xls", que traz dados antropométricos e de pressão arterial de uma amostra de sujeitos com obesidade, realize a análise descritiva, construa a tabela e interprete os resultados de idade, Índice de Massa Corporal (IMC), circunferência da cintura e pressões arteriais sistólica e diastólica. Suponha distribuição não normal dos dados para interpretação.
3. No "Banco de dados 2.xls", qual a frequência dos graus de obesidade? E de hipertensos e não hipertensos? Interprete os resultados e construa os gráficos (pizza, barras ou retângulos). Suponha distribuição não normal dos dados para interpretação.