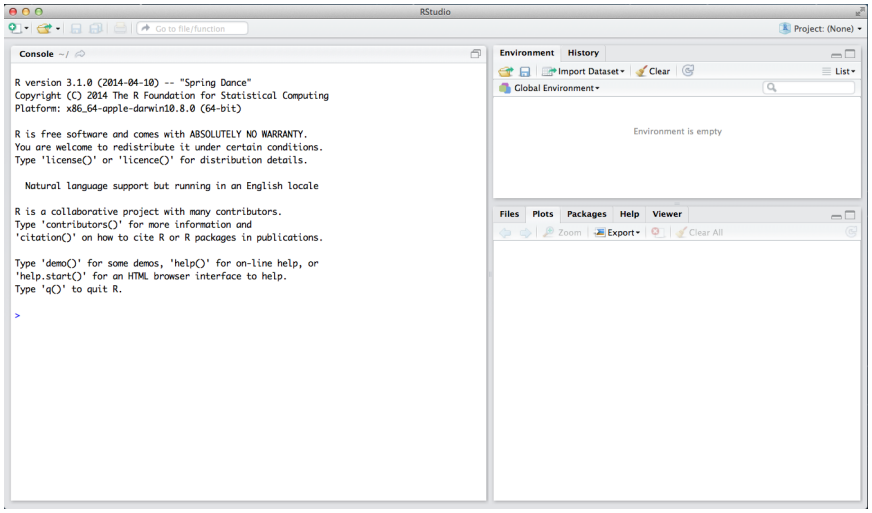
Introdução ao R e ao RStudio

O objetivo deste laboratório é introduzir ao R e ao RStudio, os programas que você usará ao longo do curso tanto para aprender os conceitos estatísticos discutidos no livro quanto para analisar dados reais e chegar a conclusões informadas. Para já distinguir qual é qual: R é o nome da linguagem de programação e RStudio é uma interface gráfica conveniente para utilizar o R.

À medida que os laboratório avançarem, você é encorajado a explorar além do que os laboratórios propõem; a vontade de experimentar o fará um programador muito melhor. Antes de chegarmos a este estágio, contudo, você precisa desenvolver alguma fluência básica em R. Hoje nós começaremos com os blocos fundamentais do R e do RStudio: a interface, importação de dados, e comandos básicos.



O painel na parte superior-direita contém seu *espaço de trabalho* e também um histórico dos comandos que você utilizou anteriormente. Quaisquer gráficos que você gerar aparecerá no painel no canto inferior direito.

O painel à esquerda é onde a ação acontece. Ele é chamado de *console*. Toda vez que você iniciar o RStudio, ele terá o mesmo texto no topo do console dizendo qual versão do R você está rodando. Abaixo desta informação está o *comando de linha*. Como o nome sugere, ele interpreta qualquer entrada como um comando a ser executado. Inicialmente, a interação com o R é feita principalmente pela digitação de comandos e a interpretação dos resultados. Esses comandos e sua sintaxe evoluíram ao longo de décadas (literalmente) e agora proporcionam o que muitos usuários acreditam ser um forma bastante natural de acessar dados e organizar, descrever e invocar computações estatísticas.

Para iniciar, entre o seguinte comando no comando de linha do R (i.e. logo depois do $>$ no comando de linha). Você pode digitar o comando manualmente ou copiar e colar deste documento.

```
source("http://www.openintro.org/stat/data/arbuthnot.R")
```

Este comando instrui o R a acessar o website da OpenIntro e buscar alguns dados: a contagem de batismos de meninos e meninas coletada por Arbuthnot . Você deve ver que a área do espaço de trabalho no canto superior direito da janela do RStudio agora lista um conjunto de dados chamado `arbuthnot` que tem 82 observações de três variáveis. À medida que você interage com o R, você criará uma série de objetos. Às vezes você os carregará como nós fizemos aqui, e às vezes você os criará por conta própria como o produto de uma computação ou alguma análise que você realizou. Preste atenção que, por você estar acessando os

dados a partir da internet, esse comando (e todas as tarefas) funcionará num laboratório de informática, na biblioteca, ou na sua casa; em qualquer lugar que você tenha acesso à internet.

Os Dados: Registro de Batismos do Dr. Arbuthnot

O conjunto de dados Arbuthnot se refere ao Dr. John Arbuthnot, um médico, escritor e matemático do século 18. Ele se interessou pela razão de meninos e meninas recém-nascidos, e para isso ele coletou os registros de batismo de crianças nascidas em Londres todos os anos entre 1629 e 1710. Nós podemos dar uma olhada nos dados digitando seu nome no comando de linha.

```
arbuthnot
```

Você deve ver quatro colunas de números, com cada linha representando um ano diferente: a primeira entrada em cada linha é simplesmente o número da linha (um índice que podemos usar para acessar os dados de anos individuais, se quisermos), a segunda é o ano, e a terceira e a quarta são os números de meninos e meninas batizados naquele ano, respectivamente. Use a barra de rolagem à direita da janela do console para examinar o conjunto de dados completo.

Preste atenção que os números das linhas na primeira coluna não fazem parte dos dados de Arbuthnot. O R os adiciona como parte das impressões em tela para ajudá-lo a fazer comparações visuais. Pense neles como um índice que costuma ficar no lado esquerdo de uma planilha. A comparação com uma planilha geralmente será útil, de fato. O R armazenou os dados de Arbuthnot em um tipo de planilha ou tabela chamada de *data frame* ou banco de dados.

Você pode ver as dimensões deste banco de dados digitando:

```
dim(arbuthnot)
```

Este comando deve dar como resposta `[1] 82 3`, indicando que há 82 linhas e 3 colunas (nós já voltaremos ao que o `[1]` quer dizer), da mesma forma como está especificado ao lado do objeto em seu espaço de trabalho. Você pode ver os nomes das colunas (ou variáveis) digitando:

```
names(arbuthnot)
```

Você deve ver que o banco de dados contém as colunas `year` (ano), `boys` (meninos), e `girls` (meninas). A essa altura, você deve ter notado que muitos dos comandos no R se parecem muito com funções matemáticas; ou seja, invocar comandos do R significa passar a uma função um certo número de argumentos. Os comandos `dim` e `names`, por exemplo, precisaram de um único argumento cada um: o nome do banco de dados.

Uma vantagem do RStudio é que ele vem com um visualizador de dados embutido. Clique no nome `arbuthnot` no canto superior direito da janela que lista os objetos em seu espaço de trabalho. Isso fará com que uma visualização alternativa das contagens de Arbuthnot apareça na janela superior esquerda. Você pode fechar o visualizador de dados clicando no "x" no canto superior esquerdo.

Explorando

Vamos começar a examinar os dados um pouco mais de perto. Nós podemos acessar separadamente os dados de uma única coluna da base de dados usando um comando como

```
arbuthnot$boys
```

Este comando mostrará somente o número de meninos batizados em cada ano.

> **Exercício 1** Qual comando você utilizaria para extrair somente a contagem de meninas batizadas? Experimente!

Preste atenção que a maneira como o R imprimiu esses dados é diferente. Quando nós visualizamos o banco de dados completo, vimos 82 linhas, uma em cada linha do console. Esses dados não estão mais estruturados em uma tabela com outras variáveis, então eles são dispostos um ao lado do outro. Objetos que são impressos na tela desta maneira são chamados de *vetores*; eles representam um conjunto de números. O R adicionou números em [colchetes] no lado esquerdo dos resultados para indicar localizações dentro do vetor. Por exemplo, `5218` segue `[1]`, indicando que `5218` é a primeira entrada no vetor. E se `[43]` inicia uma linha, então isso significa que o primeiro número naquela linha representa a 43ª entrada no vetor.

O R tem algumas funções poderosas para criar gráficos. Podemos criar uma gráfico simples do número de meninas batizadas por ano com o comando

```
plot(x = arbuthnot$year, y = arbuthnot$girls)
```

Por padrão, o R cria uma gráfico de dispersão com cada par x,y indicado por um círculo aberto. O gráfico deve aparecer sob a aba "Plots" no canto inferior direito do RStudio. Repare que o comando acima também se parece com uma função, desta vez com dois argumentos separados por vírgula. O primeiro argumento na função de gráfico especifica a variável para o eixo x e o segundo para o eixo y. Se nós quiséssemos conectar os pontos dos dados com linhas, nós poderíamos adicionar um terceiro argumento, a letra "l" de linha.

```
plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l")
```

Você pode se perguntar como você poderia saber que era possível adicionar aquele terceiro argumento. Felizmente, o R tem documentações extensivas de todas as suas funções. Para ler o que a função faz e aprender os argumentos disponíveis, basta digitar um ponto de interrogação seguido pelo nome da função na qual vocês está interessado. Tente o seguinte.

```
?plot
```

Veja que o arquivo de ajuda substitui o gráfico no painel no canto inferior direito. Você pode alternar entre gráficos e arquivos de ajuda usando as abas no topo daquele painel.

> **Exercício 2** Há alguma tendência aparente no número de meninas batizadas ao longo dos anos? Como você a descreveria?

Agora, vamos supôr que queiramos fazer um gráfico com o número total de batismos. Para calcular isso, nós podemos nos aproveitar do fato de que o R é, na verdade, apenas uma grande calculadora. Nós podemos digitar expressões matemáticas como

```
5218 + 4683
```

para ver o número total de batismos em 1629. Nós podemos repetir isso para cada ano, mas há um modo mais rápido. Se adicionarmos o vetor de batismo para meninos e meninas, o R irá computar todas as somas simultaneamente.

```
arbuthnot$boys + arbuthnot$girls
```

O que você verá são 82 números (naquela exibição compacta, porque não estamos analisando um banco de dados), cada um representando a soma que nós queremos. Dê uma olhada em alguns deles e verifique se eles estão corretos. Portanto, nós podemos criar um gráfico com o total de batismos por ano com o comando

```
plot(arbuthnot$year, arbuthnot$boys + arbuthnot$girls, type = "l")
```

Desta vez, veja que nós deixamos de fora os nomes dos dois primeiros argumentos. Nós podemos fazer isso porque o arquivo de ajuda mostra que o padrão para o comando `plot` é ter a variável x como primeiro argumento e a variável y como segundo argumento.

De maneira similar como calculamos a proporção de meninos, podemos computar a razão entre o número de meninos e o número de meninas batizadas em 1629 com

```
5218 / 4683
```

ou podemos utilizar os vetores completos com a expressão

```
arbuthnot$boys / arbuthnot$girls
```

A proporção de recém-nascidos que são meninos

```
5218 / (5218 + 4683)
```

ou também pode ser calculado para todos os anos simultaneamente:

```
arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls)
```

Preste atenção que usando o R como sua calculadora, você precisa prestar atenção da ordem das operações. Aqui, nós queremos dividir o número de meninos pelo total de recém-nascidos, portanto precisamos usar parênteses. Sem eles, o R efetuará primeiro a divisão, depois a adição, dando como resultado algo que não é uma proporção.

> **Exercício 3** Agora, crie um gráfico das proporções dos meninos com relação ao tempo. O que você percebe? Dica: se você usar as teclas de flecha para cima e para baixo, você pode retomar os comando prévios, chamado de histórico de comandos. Você também pode acessá-lo clicando na aba "history" no painel no canto superior direito. Isto irá lhe economizar várias digitações no futuro!

Por fim, além de operadores matemáticos simples como subtração e divisão, você pode pedir para o R fazer comparações como mair que, $>$, menor que, $<$, e igualdade, $==$. Por exemplo, podemos perguntar se o número de meninos é maior que de meninas em cada ano com a expressão

```
arbuthnot$boys > arbuthnot$girls
```

Este comando retorna 82 valores ou do tipo `TRUE` (verdadeiro) se aquele ano teve mais meninos batizados do que meninas, ou `FALSE` (falso) se naquele ano foi o contrário (a resposta pode surpreendê-lo). Esse resultado mostra um tipo diferente de variável daquelas que vimos até agora. No banco de dados `arbuthnot` nossos dados são numéricos (o ano, o número de meninos e meninas). Aqui, nós pedimos para o R criar dados *lógicos*, dados cujos valores são `TRUE` (verdadeiro) ou `FALSE` (falso). De modo geral, a análise de dados envolverá vários tipos diferentes de dados, e uma razão para usar o R é que ele consegue representar e realizar computações com vários tipos de dados.

Já é o bastante para seu primeiro laboratório, então vamos parar por aqui. Para sair do RStudio você pode clicar no "x" no canto superior direto da janela do aplicativo. Você será questionado se quer salvar seu espaço de trabalho. Se você clicar em "save" (salvar), o RStudio salvará seu histórico e todos os objetos de seu espaço de trabalho para que na próxima vez que você inicializar o RStudio, você verá o objeto `arbuthnot` e você terá acesso aos comando que você digitou nas suas sessões prévias. Por enquanto, clique em "save", e depois reinicialize o RStudio.

Sua Vez

Nas páginas anteriores, você recriou algumas das exposições e análises preliminares dos dados de batismo de Arbuthnot. Sua tarefa consiste repetir essas etapas, mas para os registros atuais de nascimento dos Estados Unidos. Carregue os dados atuais com o seguinte comando.

```
source("http://www.openintro.org/stat/data/present.R")
```

Os dados serão armazenados num banco de dados chamado `present`.

1. Quais anos estão incluídos neste conjunto de dados? Quais são as dimensões da base de dados e quais são os nomes das colunas ou variáveis?
2. Como estas contagens se comparam aos dados de Arbuthnot? Eles estão numa escala similar?
3. A observação de Arbuthnot de que os meninos nascem numa proporção maior que as meninas se mantém nos EUA?
4. Crie um gráfico que mostre a razão de meninos para meninas para cada ano do conjunto de dados. O que você pode verificar?
5. Em qual ano se verifica o maior número de nascimentos nos EUA? Você pode utilizar os arquivos de ajuda ou o cartão de referência do R (*http://cran.r-project.org/doc/contrib/Short-refcard.pdf*) para encontrar comandos úteis.

Esses dados são provenientes de uma pesquisa realizada pelo Centro de Controle de Doenças (Center For Disease Control) (*http://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_20.pdf*). Confira-o se você desejar ler mais sobre a análise da razão entre os sexos nos nascimentos nos Estados Unidos.

Esta foi uma curta introdução ao R e ao RStudio, mas nós forneceremos mais funções e um sentido mais completa da linguagem ao longo do curso. Sinta-se livre para procurar na internet pelo R *http:*

//www.r-project.org e o RStudio *http://rstudio.org* se vocês estiver interessado em aprender mais, ou encontre mais laboratórios para praticar em *http://openintro.org*.

Laboratório 1: Introdução à Análise de Dados

Algumas pessoas definem a Estatística como a ciência que tem por objetivo transformar informação em conhecimento. O primeiro passo no processo é sumarizar e descrever a informação bruta - os dados. Neste laboratório, você obterá novos conhecimento sobre saúde pública gerando sumários gráficos e numéricos de um conjunto de dados coletados pelo Centro para o Controle e Prevenção de Doenças ("Centers for Disease Control and Prevention", CDC). Como esse conjunto de dados é grande, ao longo do caminho você também aprenderá as habilidades indispensáveis de processamento de dados e organização de subconjuntos.

Preparações

O Sistema de Monitoramento de Fatores de Risco Comportamental ("Behavioral Risk Factor Surveillance System", BRFSS) é um *survey* anual realizado por telefone com 350.000 pessoas nos Estados Unidos. Como seu nome implica, o BRFSS foi desenvolvido para identificar fatores de risco na população adulta e relatar tendências emergentes na saúde. Por exemplo, os respondentes são indagados sobre sua dieta e atividades físicas semanais, seu diagnóstico de HIV/AIDS, uso provável de tabaco, e mesmo seu nível de cobertura por planos de saúde. O *website* do BRFSS (*http://www.cdc.gov/brfss*) contém uma descrição completa desta pesquisa, incluindo as questões de pesquisa que motivaram o estudo e muitos resultados interessantes derivados dos dados.

Nós nos focaremos numa amostra aleatória de 20.000 pessoas do BRFSS conduzido em 2000. Ainda que existam mais de 200 variáveis neste conjunto de dados, nós trabalharemos com um subconjunto menor.

Começamos importando os dados das 20.000 observações para dentro do espaço de trabalho do R. Depois de inicializar o RStudio, entre com o seguinte comando.

```
source("http://www.openintro.org/stat/data/cdc.R")
```

O conjunto de dados `cdc` que aparece em seu espaço de trabalho é uma *matriz de dados*, com cada linha representando um *caso* e cada coluna representando uma *variável*. O R denomina este formato de dados como *banco de dados* (*data frame*), que será um termo utilizado ao longo dos laboratórios.

Para visualizar o nome das variáveis, digite o comando

```
names(cdc)
```

Este comando retorna os nomes `genhlth`, `exerany`, `hlthplan`, `smoke100`, `height`, `weight`, `wtdesire`, `age`, e `gender`. Cada uma dessas variáveis corresponde a uma questão que foi feita na pesquisa. Por exemplo, para `genhlth`, os respondentes foram indagados sobre sua saúde geral, respondendo excelente, muito bom, bom, razoável ou ruim. A variável `exerany` indica se o respondente se exercitou no último mês (1) ou não (0). Da mesma forma, `hlthplan` indica se o respondente tem alguma forma de cobertura (1) ou não (0). A variável `smoke100` indica se o respondente fumou pelo menos 100 cigarros ao longo da vida. As outras variáveis registram a altura (`height`) em polegadas, o peso (`weight`) em libras, bem como o peso desejado (`wtdesire`), idade (`age`) em anos, e gênero (`gender`)

> **Exercício 1** Há quantos casos neste conjunto de dados? Quantas variáveis? Para cada variável, identifique seu tipo de dado (p.e., categorial, discreta).

Nós podemos dar uma olhada nas primeiras entradas (linhas) de nossos dados com o comando

```
head(cdc)
```

e, similarmente, podemos verificar as últimas digitando

```
tail(cdc)
```

Você também pode verificar *toda* a base de dados de uma vez só digitando seu nome no console, mas isso pode não ser muito sábio neste contexto. Sabemos que `cdc` tem 20.000 linhas, portanto verificar o conjunto de dados inteiro significa inundar sua tela. É melhor dar pequenas espiadas nos dados utilizando `head`, `tail`, ou as técnicas de construção de subconjunto que você aprenderá logo em seguida.

Sumários e Tabelas

O questionário do BRFSS é um tesouro enorme de informações. Um primeiro passo útil em qualquer análise é destilar toda essa informação em algumas estatísticas sumárias e gráficos. Como um exemplo simples, a função `summary` retorna um sumário numérico: mínimo, primeiro quartil, mediana, média, segundo quartil, e máximo. Para a variável `weight`, esse sumário é:

```
summary(cdc$weight)
```

O R também funciona como uma calculadora poderosa. Se vocês quisesse calcular o intervalo interquartil para o peso dos respondentes, você pode se basear na saída do comando acima e então digitar

```
190 - 140
```

O R também tem funções embutidas para calcular estatísticas descritivas uma por uma. Por exemplo, para calcular a média, mediana, e variância da variável `weight`, digite

```
mean(cdc$weight)

var(cdc$weight)

median(cdc$weight)
```

Ainda que faça sentido descrever uma variável quantitativa como `weight` em termos destas estatísticas, o que fazer com dados categoriais? Nós podemos considerar a frequência da amostra ou a distribuição relativa de frequência. A função `table` faz isso por você contando o número de vezes que cada tipo de resposta é dada. Por exemplo, para ver o número de pessoas que fumaram 100 cigarros ao longo de sua vida, digite

```
table(cdc$smoke100)
```

Ou então verifique a distribuição de frequência relativa digitando

```
table(cdc$smoke100)/20000
```

Perceba como o R automaticamente divide todas as entradas na tabela por 20.000 no comando acima. Isso é similar a algo que observamos no último laboratório; quando multiplicamos ou dividimos um vetor por um número, o R aplica essa ação a todas as entradas dos vetores. Como vimos acima, isso também funciona para tabelas. Em seguida, criamos um gráfico de barras para as entradas na tabela inserindo a tabela dentro do comando para gráficos de barra.

```
barplot(table(cdc$smoke100))
```

Preste atenção no que fizemos agora! Nós computamos a tabela da variável `cdc$smoke100` e então imediatamente aplicamos a função gráfica, `barplot`. Esta é uma ideia importante: os comandos do R podem ser aninhados. Você também pode dividir esse procedimento em dois passos digitando o seguinte:

```
smoke <- table(cdc$smoke100)

barplot(smoke)
```

Agora, criamos um novo objeto, uma tabela, denominada `smoke` (seu conteúdo pode ser verificado digitando `smoke` no console) e então a utilizamos como entrada para o comando `barplot`. O símbolo especial `<-` realiza uma *atribuição*, tomando a saída de uma linha de código e salvando-a em um objeto no seu espaço de trabalho. Esta é outra ideia importante para a qual retornaremos mais tarde.

> **Exercício 2** Crie um sumário numérico para `height` (altura) e `age` (idade), e calcule o intervalo interquartílico para cada um. Calcule a distribuição de frequência relativa para `gender` e `exerany`. Quantos homens compõem a amostra? Qual proporção da amostra diz estar com saúde excelente?

O comando `table` pode ser utilizado para tabular qualquer número de variáveis que você quiser. Por exemplo, para examinar quais participantes fumam, dividido por gênero, nós podemos utilizar o seguinte código.

```
table(cdc$gender,cdc$smoke100)
```

Aqui, vemos etiquetas de coluna formadas por 0 e 1. Lembre-se que o 1 indica que o respondente fumou pelo menos 100 cigarros. As linhas se referem ao gênero. Para criar um gráfico de mosaico para essa tabela, entramos com o seguinte comando.

```
mosaicplot(table(cdc$gender,cdc$smoke100))
```

Nós poderíamos ter conseguido esse resultado em duas etapas: salvando a tabela em uma linha e aplicando `mosaicplot` em seguida (veja o exemplo de tabela/gráfico de barras acima).

> **Exercício 3** O que o gráfico de mosaico revela sobre os hábitos de fumar e gênero?

Interlúdio: Como o R Pensa a Respeito dos Dados

Mencionamos que o R armazena os dados em bases de dados, que você pode pensar como um tipo de planilha. Cada linha é uma observação diferente (um respondente diferente) e cada coluna é uma variável diferente (a primeira é `genhlth`, a segunda é `exerany` e assim por diante). Nós podemos visualizar o tamanho da base de dados ao lado do nome do objeto na área de trabalho ou podemos digitar

```
dim(cdc)
```

o que faz retornar o número de linhas e colunas. Agora, se quisermos acessar um subconjunto da base de dados completa, nós podemos utilizar a notação de linhas-e-colunas. Por exemplo, para visualizar a sexta variável do 567º respondente, utilize o comando

```
cdc[567,6]
```

que significa que nós queremos o elemento de nosso conjunto de dados que está na 567ª linha (ou seja, a 567ª pessoa ou observação) e na 6ª coluna (nesse caso, o peso). Sabemos que `weight` (peso) é a 6ª variável porque ela é a 6ª entrada na lista de nomes de variáveis.

```
names(cdc)
```

Para visualizar os pesos para os primeiros 10 respondentes, podemos digitar

```
cdc[1:10,6]
```

Nesta expressão, nós pedimos somente pelas linhas no intervalo entre 1 e 10. O R usa o ":" para criar um intervalo de valores, de tal forma que 1:10 se expande para 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Você pode visualizar isso digitando

```
1:10
```

Finalmente, se nós queremos todos os dados dos 10 primeiros respondentes, digite

```
cdc[1:10,]
```

Ao deixar de fora um índice ou intervalo (nós não digitamos nada entre a vírgula e o colchete), nós obtemos todas as colunas. Quando iniciamos o uso do R, isso parece um pouco contra-intuitivo. Como um regra geral, omitimos o número da coluna para ver todas as colunas numa base de dados. Da mesma forma, se deixamos de fora um índice ou intervalo para as linhas, nós acessaríamos todas as observações, não apenas a 567ª, ou as linhas 1 a 10. Experimente o código seguinte para ver o peso de todos os 20.000 respondentes passarem voando por sua tela

```
cdc[,6]
```

Recorde que a coluna 6 representa o peso dos respondentes, e portanto o comando acima mostra todos os pesos no conjunto de dados. Um método alternativo para acessar os dados sobre peso é utilizar o seu nome. Anteriormente, digitamos `names(cdc)` para ver todas as variáveis contidas no conjunto de dados

cdc. Nós podemos utilizar qualquer um dos nomes de variáveis para selecionar itens no seu conjunto de dados.

```
cdc$weight
```

O cifrão informa ao R para recuperar na base de dados cdc a coluna denominada weight. Uma vez que se trata de um único vetor, podemos formar subconjuntos utilizando apenas um único índice dentro dos colchetes. Nós verificamos o peso para o 567º respondente digitando

```
cdc$weight[567]
```

Da mesma forma, para apenas os 10 primeiros respondentes

```
cdc$weight[1:10]
```

O comando acima retorna o mesmo resultado que o comando cdc[1:10,6]. Tanto a notação linha-e-coluna quanto a notação utilizando o cifrão são amplamente utilizadas. Qual você escolhe depende da sua preferência pessoal.

Um Pouco Mais Sobre Formação de Subconjuntos

É frequentemente útil extrair todos os sujeitos (casos) de um conjunto de dados que possuem características específicas. Nós conseguimos isso por meio de comando *condicionais*. Primeiramente, considere expressões como

```
cdc$gender == "m"
```

ou

```
cdc$age > 30
```

Esses comandos produzem uma série de valores TRUE (verdadeiro) e FALSE (falso). Há um valor para cada respondente, sendo que TRUE indica que a pessoa era do sexo masculino (pelo primeiro comando) ou mais velha que 30 anos (segundo comando).

Vamos supor que queiramos extrair apenas os dados para homens na amostra, ou apenas para aqueles acima de 30 anos. Nós podemos utilizar a função do R subset para fazer isso por nós. Por exemplo, o comando

```
mdata <- subset(cdc, cdc$gender == "m")
```

criará um novo conjunto de dados denominado mdata que contém apenas os homens do conjunto de dados cdc. Além de poder encontrá-lo em seu espaço de trabalho junto com suas dimensões, você pode dar uma olhada nas primeiras linhas como já fizemos

```
head(mdata)
```

Este novo conjunto de dados contém as mesmas variáveis mas cerca de metade das linhas. Também é possível pedir para o R manter apenas variáveis específicas, um tópico que abordaremos num laboratório no futuro. Por enquanto, o importante é que podemos desmembrar os dados com base nos valores de uma ou mais variáveis.

Você também pode utilizar vários condicionais em conjunto com `&` e `|`. O `&` é lido como "e" de tal forma que

```
m_and_over30 <- subset(cdc, cdc$gender == "m" & cdc$age > 30)
```

resultará nos dados para homens acima de 30 anos de idade. O caractere `|` é interpretado como "ou" de tal forma que

```
m_or_over30 <- subset(cdc, cdc$gender == "m" | cdc$age > 30)
```

selecionará pessoas que são homens ou então acima de 30 anos (por que esse grupo seria interessante é difícil dizer, mas por enquanto entender o comando é o mais importante). A princípio, você pode utilizar quantos "e" e "ou" você quiser quando formar um subconjunto.

> **Exercício 4** Crie um novo objeto denominado `under23_and_smoke` (ou, se preferir, `abaixo23_e_fuma`) que contém todas as observações dos respondentes com menos de 23 anos que fumaram pelo menos 100 cigarros ao longo de sua vida. Escreva o comando que você utilizou para criar o novo objeto como resposta para esse exercício.

Dados Quantitativos

Com nossas ferramentas para criar subconjuntos a postos, podemos retornar à tarefa de hoje: criar sumários básicos do questionário BRFSS. Nós já olhamos os dados categoriais como `smoke` (fumante) e `gender` (gênero). Agora vamos nos concentrar nos dados quantitativos. Duas formas comuns de visualizar dados quantitativos é por meio de gráfico de caixas e histogramas. Nós podemos construir um gráfico de caixas para uma única variável com o seguinte comando.

```
boxplot(cdc$height)
```

Você pode comparar a localização dos componentes da caixa examinando as estatísticas sumárias.

```
summary(cdc$height)
```

Confirme que a mediana e os quartis superior e inferior informados no sumário numérico batem com os apresentados no gráfico. O objetivo de um gráfico de caixa é prover um pequeno esboço de uma variável com o propósito de comparar entre várias categorias. Podemos, por exemplo, comparar as alturas de homens e mulheres com

```
boxplot(cdc$height ~ cdc$gender)
```

A notação aqui é nova. O caractere ~ pode ser lido como "versus" ou "como uma função de". Estamos, portanto, pedindo ao R para nos dar um gráfico de caixas das alturas no qual os grupos são definidos pelo gênero.

Na sequência, consideremos uma nova variável que não aparece diretamente neste conjunto de dados: o Índice de Massa Corporal (IMC). IMC é uma razão entre peso e altura que pode ser calculado da seguinte maneira:

$$IMC = \frac{peso\ (lbs)}{altura\ (pols)^2} * 703^{\dagger}$$

As duas linhas seguintes criam um novo objeto chamado `bmi` (de *Body Mass Index*) e então criamos um gráfico de caixas para esses valores, definindo grupos pela variável `cdc$genhlth`

```
bmi <- (cdc$weight / cdc$height^2) * 703

boxplot(bmi ~ cdc$genhlth)
```

Perceba que a primeira linha acima é apenas aritmética, mas é aplicada para todos os 20.000 número do conjunto de dados `cdc`. Ou seja, para cada um dos 20.000 participantes, pegamos seu peso, dividimos pelo quadrado de sua altura e multiplicamos por 703. O resultado é 20.000 valores de IMC, um para cada respondente. Essa é uma das razões pela qual gostamos do R: ele nos permite realizar cálculos como esse utilizando expressões bem simples.

> **Exercício 5** O que este gráfico de caixas mostra? Escolha outra variável categorial do conjunto de dados e verifique como ela se relaciona ao IMC. Liste a variável que você escolheu, por que você pensou que ela poderia ter relação com o IMC e indique o que o gráfico parece sugerir.

Por fim, vamos fazer alguns histogramas. Nós podemos verificar o histograma da idade de nossos respondentes com o comando

```
hist(cdc$age)
```

Histogramas são geralmente uma boa maneira de visualizar a forma de uma distribuição, mas essa forma pode mudar dependendo como os dados são divididos entre os diferentes segmentos. Você pode controlar o número de segmentos adicionando um argumento ao comando. Nas próximas duas linhas, primeiro fazemos um histograma padrão da variável `bmi` e depois um com 50 segmentos.

```
hist(bmi)

hist(bmi, breaks = 50)
```

Perceba que você pode alternar entre gráficos que você criou clicando nas flechas de avançar e retroceder na região inferior direita do RStudio, logo acima dos gráficos. Quais as diferenças entre esses histogramas?

A esta altura, fizemos uma boa primeira exposição sobre análise das informações no questionário BRFSS. Nós descobrimos um associação interessante entre fumo e gênero, e nós podemos comentar algo a respeito da relação entre a avaliação de saúde em geral dada pelas pessoas e seu próprio IMC. Nós também nos

[†] 703 é um fator de conversão aproximado para mudar as unidades do sistema métrico (metro e kilograma) para o sistema imperial (polegadas e libras). Isso é necessário porque os dados disponíveis estão no sistema imperial. No sistema métrico basta dividir o peso em quilogramas pelo quadrado da altura em metros.

apropriamos de ferramentas computacionais essenciais – estatísticas sumárias, subconjuntos, e gráficos – que nos servirão bem ao longo deste curso.

Sua Vez

1. Crie um gráfico de dispersão da variável peso em relação ao peso desejado. Defina a relação entre essas duas variáveis.
2. Vamos considerar uma nova variável: a diferença entre o peso desejado (`wtdesire`) e o peso atual (`weight`). Crie esta nova variável subtraindo as duas colunas na base de dados e atribuindo-as a um novo objeto chamado `wdiff`.
3. Que tipo de dado está contido na variável `wdiff`? Se uma observação de `wdiff` é 0, o que isso implica com relação ao peso atual e desejado de uma pessoas? E se o valor de `wdiff` for positivo ou negativo?
4. Descreva a distribuição de `wdiff` em termos de seu centro, forma e variação, incluindo qualquer gráfico que você usar. O que isso nos diz sobre como as pessoas se sentem a respeito do seu peso atual?
5. Utilizando sumários numéricos e um gráfico de caixas lado-a-lado, determine se homens tendem a ver seu peso diferentemente das mulheres.
6. Agora chegou a hora de usar a criatividade. Encontre a média e o desvio padrão de `weight` e determine qual a proporção de pesos que estão a um desvio padrão da média.
7. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

Laboratório 2: Probabilidade

Mãos Quentes

Jogadores de basquete que pontuam várias vezes seguidas costumam ser descritos como tendo as "mãos quentes". Fãs e jogadores acreditam há muito tempo no fenômeno da mão quente, que refuta o pressuposto de que cada lance é independente do próximo. Contudo, um artigo de 1985 escrito por Gilovich, Vallone e Tversky coletou evidência que contradiz essa crença e mostrou que lances sucessivos são eventos independentes.[†] Este artigo iniciou uma grande controvérsia que continua até hoje, como você pode verificar se procurar por "hot hand basketball" no Google.

Não temos a expectativa de resolver esta controvérsia hoje. Entretanto, neste laboratório nós aplicaremos um procedimento para responder a questões como essa. Os objetivos deste laboratório são (1) refletir sobre o efeito de eventos independentes e dependentes, (2) aprender como simular sequências de lances no R, e (3) comparar a simulação com os dados efetivos para determinar se o fenômeno das mãos quentes parece ser real.

Salvando seu Código

Clique em File → New → R Script. Um documento em branco será aberto acima do console. À medida que o laboratório avançar, você pode copiar e colar seu código aqui e salvá-lo. Esta é uma boa maneira de manter um registro do seu código e reutilizá-lo mais tarde. Para executar seu código a partir deste documento, você pode ou copiar e colar os comandos no console, ou selecionar o código e clicar no botão Run (Executar), ou então selecionar o código e pressionar command+enter se estiver utilizando um Mac ou control+enter num PC.

Você também poderá salvar este *script* (documento de código). Para fazer isso basta clicar no ícone de disquete. A primeira vez que você pressionar o botão de salvar, o RStudio pedirá por um nome de arquivo; você pode dar qualquer nome que quiser. Depois de clicar em salvar você verá o arquivo aparecer sob a aba Files no painel inferior direito. Você pode reabrir este arquivo a qualquer momento simplesmente clicando sobre ele.

Preparações

Nossa investigação focará na performance de um jogador: Kobe Bryant do Los Angeles Lakers. Sua performance contra o Orlando Magic nas finais de 2009 da NBA lhe deram o título de "Jogador Mais Valioso" e vários espectadores comentaram como ele parecia demonstrar uma mão quente. Vamos carregar alguns dados desses jogos e analisar as primeiras linhas.

```
download.file("http://www.openintro.org/stat/data/kobe.RData", destfile = "kobe.RData")

load("kobe.RData")

head(kobe)
```

Neste banco de dados, cada linha registra um lance feito por Kobe Bryant. Se ele acertou o lance (fez uma cesta), um acerto, `H` (de *Hit*), é registrado na coluna denominada `basket` (cesta); caso contrário um erro, `M` (de *Miss*), é registrado.

[†] "The Hot Hand in Basketball: On the Misperception of Random Sequences", Gilovich, T., Vallone, R., Tversky, A., 1985. Cognitive Psychology, 17, pp. 295-314.

Apenas olhando para a sequência de acertos e erros pode ser difícil de aferir se é possível que Kobe estava arremessando com as mãos quentes. Uma maneira possível de abordar este problema é considerar a crença de que arremessadores com a mão quente tendem a conseguir uma longa sequências de acertos. Para este laboratório, definiremos o comprimento de uma sequência de acertos como o *número de cestas consecutivas até acontecer um erro*.

Por exemplo, no Jogo 1 Kobe teve a seguinte sequência de acertos e erros de suas nove tentativas de arremessos no primeiro quarto:

H M | M | H H M | M | M | M

Para verificar estes dados no R, use o seguinte comando:

```
kobe$basket[1:9]
```

Dentre as nove tentativas de arremesso há seis sequências, que são separadas por um "|" acima. Seus comprimentos são um, zero, dois, zero, zero, zero (em ordem de ocorrência).

> **Exercício 1** O que uma sequência de comprimento 1 significa, ou seja, quantos acertos e erros existem dentro de um sequência de 1? E de uma sequência de comprimento 0?

A função personalizada `calc_streak`, que foi carregada com os dados, pode ser utilizada para calcular os comprimentos de todas as sequências de acertos e então conferir sua distribuição.

```
kobe_streak <- calc_streak(kobe$basket)

barplot(table(kobe_streak))
```

Perceba que, ao invés de fazer um histograma, escolhemos criar uma gráfico de barras a partir de uma tabela dos dados das sequências. Uma gráfico de barras é preferível neste contexto uma vez que nossa variável é discreta – contagens – ao invés de contínua.

> **Exercício 2** Descreva a distribuição do comprimento das sequências de Kobe nas finais de 2009 da NBA. Qual foi seu tamanho de sequência típico? Quão longa foi sua maior sequência de cestas?

Comparado a quê?

Mostramos que Kobe teve algumas sequências de arremesso longas, mas elas são longas o suficiente para apoiar a crença de que ele tinha mãos quentes? Com o que podemos compará-las?

Para responder a essa pergunta, vamos retornar à ideia de *independência*. Dois processos são independentes se o resultado de um processo não afeta o resultado do outro. Se cada arremesso que o jogador faz é um processo independente, ter acertado ou errado o primeiro arremesso não afetará a probabilidade de ele converter ou errar seu segundo arremesso.

Um arremessador com as mãos quentes terá arremessos que *não* são independente um do outro. Mais especificamente, se o arremessador converte seu primeiro arremesso, o modelo das mãos quentes diz que ele terá uma probabilidade *maior* de converter seu segundo arremesso.

Vamos supor por um momento que o modelo das mãos quente é válido para Kobe. Durante sua carreira, o percentual de vezes que Kobe faz uma cesta (ou seja, sua porcentagem de arremessos) é de cerca de 45%, ou, em notação de probabilidade,

$$P(\text{arremesso 1} = \text{H}) = 0.45$$

Se ele converte o primeiro arremesso e tem as mãos quentes (arremesso *não* independentes), então a probabilidade de ele converter seu segundo arremesso deveria aumentar para, digamos, 60%,

$$P(\text{arremesso 2} = \text{H} \,|\, \text{arremesso 1} = \text{H}) = 0.60$$

Como um resultado do aumento da probabilidade, seria esperado que Kobe tivesse sequências mais longas. Compare com a perspectiva cética de que Kobe *não* tem as mãos quentes, ou seja, que cada arremesso é independente do próximo. Se ele acerta seu primeiro arremesso, a probabilidade de ele acertar o segundo continua sendo 0.45.

$$P(\text{arremesso 2} = \text{H} \,|\, \text{arremesso 1} = \text{H}) = 0.45$$

Ou seja, converter o primeiro arremesso não afeta de maneira alguma a probabilidade de ele converter seu segundo arremesso. Se os arremessos de Kobe são independentes, então ele teria a mesma probabilidade de acertar cada arremesso independentemente de seus arremessos anteriores: 45%.

Agora que reformulamos a situação em termos de arremessos independentes, vamos retornar à questão: como podemos saber se as sequências de arremessos de Kobe são longas o suficiente para indicar que ele tem mãos quentes? Podemos comparar o tamanho de suas sequências a alguém que não tem as mãos quentes: um arremessador independente.

Simulações no R

Apesar de não termos nenhum dado de um arremessador que sabemos fazer arremessos independentes, esse tipo de dado é muito fácil de simular no R. Numa simulação, você define as regras básicas de um processo aleatório e então o computador utiliza números aleatórios para gerar um resultado fiel a essas regras. Como um exemplo simples, você pode simular um lance de uma moeda honesta com o seguinte código:

```
outcomes <- c("heads", "tails")

sample(outcomes, size = 1, replace = TRUE)
```

O vetor `outcomes` (resultados) pode ser entendido como um chapéu com duas tiras de papel dentro dele: numa tira está escrito "cara" ("heads") e na outra "coroa" ("tails"). A função `sample` (amostra) sorteia uma tira de dentro do chapéu e revela se ela é cara ou coroa.

Execute o segundo comando listado acima várias vezes. Da mesma maneira quando jogando uma moeda, algumas vezes você obterá cara, algumas vezes você obterá coroa, mas a longo prazo você esperaria obter um número mais ou menos igual de cada.

Se você quisesse simular o lançamento de uma moeda honesta 100 vezes, você poderia ou rodar a função 100 vezes ou, mais simples, ajustar o argumento `size` (tamanho), que regula quantas amostras retirar (o argumento `replace = TRUE` indica que nós recolocamos a tira de papel de volta no chapéu antes de retirar outra amostra). Salve o vetor resultante de cara ou coroa num novo objeto denominado `sim_fair_coin` (ou, se preferir, `sim_moeda_honesta`).

```
sim_fair_coin <- sample(outcomes, size = 100, replace = TRUE)
```

Para visualizar os resultados desta simulação, digite o nome do objeto e então use o comando `table` pra contar o número de caras e coroas.

```
sim_fair_coin

table(sim_fair_coin)
```

Uma vez que há apenas dois elementos no vetor `outcomes`, a probabilidade de um lance de uma moeda dar cara é 0.5. Digamos que estamos tentando simular uma moeda viciada que sabemos que dá cara somente 20% das vezes. Podemos ajustar adicionando um argumento denominado `prob`, que fornece um vetor de dois pesos de probabilidade.

```
sim_unfair_coin <- sample(outcomes, size = 100, replace = TRUE, prob = c(0.2, 0.8))
```

`prob=c(0.2,0.8)` indica que, para os dois elementos no vetor `outcomes`, nós queremos selecionar o primeiro, `heads` (cara), com probabilidade 0.2, e o segundo, `tails` (coroa), com probabilidade 0.8. [†]

> **Exercício 3** Em sua simulação de lançar uma moeda viciada 100 vezes, quantos lances deram cara?

Num certo sentido, nós reduzimos o tamanho da tira de papel que diz "cara", tornando-o menos provável de ser escolhido, e nós aumentamos o tamanho da tira de papel que diz "coroa", tornando-o mais provável de ser retirado. Quando simulamos a moeda honesta, ambas as tiras de papel tinham o mesmo tamanho. Isso acontece por padrão se você não fornecer o argumento `prob`; todos os elementos no vetor `outcomes` tem igual probabilidade de serem escolhidos.

Se você quiser saber mais sobre a função `sample` ou qualquer outra, lembre-se que você pode sempre conferir seu arquivo de ajuda.

```
?sample
```

Simulando o Arremessador Independente

Para simular um jogador de basquete que arremessa de forma independente, utilizamos o mesmo mecanismo que empregamos para simular o lance de uma moeda. Para simular um único arremesso de um arremessador independente, com um percentual de arremesso de 50%, digitamos

```
outcomes <- c("H", "M")

sim_basket <- sample(outcomes, size = 1, replace = TRUE)
```

Para podermos fazer uma comparação válida entre Kobe e nosso arremessador independente simulado, precisamos alinhar tanto seus percentuais de arremesso quanto seus números de arremessos tentados.

> **Exercício 4** Qual mudança precisa ser feita para que a função sample reflita o percentual de arremessos de 45%? Faça esse ajuste, e então rode a simulação para uma amostra de 133

[†]Outra maneira de pensar sobre esse cenário é imaginar o espaço amostral como um saco contendo 10 fichas, sendo 2 marcadas como "cara" e 8 como "coroa". Portanto, a cada seleção, a probabilidade de retirar uma ficha escrito "cara" é 20% e "coroa" é 80%.

arremessos. Atribua o resultado dessa simulação a um novo objeto chamado `sim_basket` (se preferir, `sim_cestas`).

Perceba que nomeamos o novo vetor como `sim_basket`, o mesmo nome que demos ao vetor anterior correspondente a um percentual de arremesso de 50%. Nessa situação, o R sobrescreve o objeto antigo com o novo, portanto sempre se certifique que você não precisa da informação no vetor antigo antes de atribuir um novo objeto ao seu nome.

Com os resultados da simulação salvos como `sim_basket`, temos os dados necessários para comprar Kobe a nosso arremessador independente. Podemos visualizar os dados de Kobe em conjunto com nossos dados simulados.

```
kobe$basket

sim_basket
```

Ambos os conjuntos de dados representam o resultado de 133 tentativas de arremessos, cada uma com o mesmo percentual de arremesso de 45%. Sabemos que nosso dados simulados são de uma arremessador que arremessa de forma independente. Quer dizer, sabemos que o arremessador simulado não tem as mãos quentes.

Sua vez

Comparando Kobe Bryant ao Arremessador Independente

Utilizando a função `calc_streak`, calcule o comprimento das sequências do vetor `sim_basket`.

1. Descreva a distribuição das sequências de arremessos. Qual é o comprimento de sequência típico para o arremessador independente simulado com um percentual de arremesso de 45%? Quão longa é a sequência mais longa de cestas em 133 arremessos?
2. Se você rodasse a simulação do arremessador independente uma segunda vez, como você acha que seria a distribuição de sequências em relação à distribuição da questão acima? Exatamente a mesma? Mais ou menos parecida? Completamente diferente? Explique seu raciocínio.
3. Como a distribuição dos comprimentos de sequência de Kobe Bryant, analisada na página 2, se comparam à distribuição de comprimentos de sequência do arremessador simulado? Utilizando essa comparação, você tem evidência de que o modelo das mãos quentes se ajusta aos padrões de arremessos de Kobe? Explique.
4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

Laboratório 3: Distribuições de Variáveis Aleatórias

Neste laboratório investigaremos a distribuição de probabilidade que é a mais central para a estatística: a distribuição normal. Se estamos confiantes de que nossos dados são aproximadamente normais, uma porta para métodos estatísticos poderosos é aberta. Aqui nós utilizaremos ferramentas gráficas do R para avaliar a normalidade de nossos dados e também aprender como gerar números aleatórios de uma distribuição normal.

Os Dados

Esta semana trabalharemos com medidas de dimensões do corpo. Este conjunto de dados contém medidas de 247 homens e 260 mulheres, a maioria dos quais foram considerados adultos jovens saudáveis.

```
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")

load("bdims.RData")
```

Vamos dar uma rápida olhada nas primeiras linhas dos dados.

```
head(bdims)
```

Você verá que para cada observação temos 25 medidas, muitas das quais são diâmetros ou circunferências. Uma chave para os nomes das variáveis pode ser encontrada no *site* *http://www.openintro.org/stat/data/bdims.php*, mas nos focaremos em apenas três colunas para iniciar: peso em kg (`wgt`), altura em cm (`hgt`), e `sex` (sexo, `1` indica masculino, `0` indica feminino).

Uma vez que homens e mulheres tendem a ter dimensões corporais diferentes, será útil criar dois conjuntos de dados adicionais: um com os dados dos homens e outro com os dados das mulheres.

```
mdims <- subset(bdims, bdims$sex == 1)

fdims <- subset(bdims, bdims$sex == 0)
```

> **Exercício 1** Elabore um histograma da altura dos homens e um histograma das alturas das mulheres. Como você descreveria os diferentes aspectos das duas distribuições?

A Distribuição Normal

Na sua descrição das distribuições, você utilizou palavras como "em forma de sino" ou "normal"? É tentador afirmar isso quando encontramos uma distribuição simétrica e unimodal.

Para verificar quão precisa é essa descrição, podemos desenhar uma curva de distribuição normal sobre o histograma para ver se os dados seguem uma distribuição normal de perto. Essa curva normal deve ter a mesma média e desvio padrão dos dados da amostra. Trabalharemos com as alturas das mulheres. Por isso, vamos armazená-las como um objeto separado e então calcular algumas estatísticas que serão utilizadas mais adiante.

```
fhgtmean <- mean(fdims$hgt)

fhgtsd   <- sd(fdims$hgt)
```

Em seguida, construímos um histograma de densidade que servirá como pano de fundo e utilizamos a função `lines` para sobrepor a curva de probabilidade normal. A diferença entre um histograma de frequência e um histograma de densidade é que, enquanto no histograma de frequência as *alturas* das barras somadas resultam no número total de observações, num histograma de densidade as *áreas* das barras somadas resultam em 1. A área de cada barra pode ser calculada simplesmente como a altura $\times$ a largura da barra. Um histograma de densidade permite-nos sobrepor corretamente uma curva de distribuição normal sobre o histograma uma vez que a curva é uma função de densidade de probabilidade normal. Histogramas de frequência de densidade tem a mesma forma; eles diferem apenas com relação a seu eixo y. Você pode verificar isso comparando o histograma de frequência que você construiu antes e o histograma de densidade criado pelos comandos abaixo.

```
hist(fdims$hgt, probability = TRUE)

x <- 140:190

y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)

lines(x = x, y = y, col = "blue")
```

Depois de criar o histograma de densidade com o primeiro comando, nós criamos as coordenadas dos eixos x e y para a curva normal. Escolhemos o intervalo de `x` entre 140 e 190, de forma a abarcar o intervalo completo da variável `fheight`. Para criar `y`, utilizamos a função `dnorm` para calcular a densidade de cada um dos valores de x numa distribuição que é normal com média `fhgtmean` e desvio padrão `fhgtsd`. O comando final desenha a curva sobre o gráfico existente (o histograma de densidade) conectando cada ponto especificado por `x` e `y`. O argumento `col` simplesmente estabelece a cor da linha a ser desenhada. Se não especificarmos este argumento, a linha seria desenhada na cor preta.[†]

> **Exercício 2** Baseado neste gráfico, parece que os dados seguem aproximadamente uma distribuição normal?

Avaliando a Distribuição Normal

Verificar visualmente a forma do histograma é uma maneira de determinar se os dados parecem se distribuir de maneira quase normal, mas pode ser frustrante decidir quão próximo o histograma está da curva. Uma abordagem alternativa envolve construir uma gráfico de probabilidade normal, também chamado de gráfico normal Q-Q, de "quantil-quantil".

```
qqnorm(fdims$hgt)

qqline(fdims$hgt)
```

Um conjunto de dados que é aproximadamente normal resultará em um gráfico de probabilidade no qual os pontos seguem de perto a linha. Quaisquer desvios da normalidade conduz a desvios desses pontos

[†]O topo da curva é cortado porque os limites dos eixos x e y são ajustados de forma mais adequada ao histograma. Para ajustar o eixo y você pode adicionar um terceiro argumento à função de histograma: `hist(fdims$hgt, probability = TRUE, ylim = c(0, 0.06))`.

com relação à linha. O gráfico para a altura de mulheres mostra pontos que tendem a seguir a linha mas com alguns pontos errantes na direção das caudas. Voltamos ao mesmo problema que encontramos com o histograma acima: quão perto é perto o suficiente?

Uma maneira útil de endereçar essa questão é reformulá-la da seguinte maneira: como gráficos de probabilidade se parecem para dados que *sabemos* serem provenientes de uma distribuição normal? Podemos responder a essa pergunta simulando dados a partir de uma distribuição normal utilizando a função `rnorm`.

```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
```

O primeiro argumento indica quantos números você gostaria de gerar, que aqui especificamos para ser o mesmo número de alturas no conjunto de dados `fdims` utilizando a função `length`. Os últimos dois argumentos determinam a média e o desvio padrão da distribuição normal a partir da qual a amostra simulada será gerada. Podemos visualizar a forma de nosso conjunto de dados simulado, `sim_norm`, assim como seu gráfico de probabilidade normal.

> **Exercício 3** Faça um gráfico de probabilidade normal do vetor `sim_norm`. Os pontos caem todos em cima da linha? Como este gráfico se compara ao gráfico de probabilidade dos dados reais?

Ainda melhor do que comparar o gráfico original a um único gráfico gerado a partir de uma distribuição normal é compará-lo a vários outros gráficos utilizando a seguinte função. Pode ser útil clicar no botão "zoom" na janela do gráfico.

```
qqnormsim(fdims$hgt)
```

> **Exercício 4** O gráfico de probabilidade normal para `fdims$hgt` parece similar aos gráficos criados para os dados simulados? Quer dizer, os gráficos fornecem evidência de que as alturas de mulheres são aproximadamente normais?

> **Exercício 5** Usando a mesma técnica, determine se os pesos de mulheres parecem ser provenientes de uma distribuição normal.

Probabilidades Normais

Muito bem, agora você tem várias ferramentas para julgar se uma variável se distribui normalmente. Mas por que deveríamos nos importar?

Acontece que os estatísticos conhecem várias coisas sobre a distribuição normal. Uma vez que decidimos que a variável aleatória é aproximadamente normal, podemos responder vários tipos de perguntas sobre aquela variável com relação à probabilidade. Por exemplo, a questão: "Qual é a probabilidade de que uma mulher adulta jovem escolhida por acaso é maior do que 6 pés (cerca de 182 cm)"? [†]

Se assumirmos que as alturas de mulheres são distribuídas normalmente (uma aproximação também é aceitável), podemos encontrar essa probabilidade calculando um escore Z e consultando uma tabela Z (também denominada de tabela de probabilidade da normal). No R, isto pode ser feito rapidamente com a função `pnorm`.

[†]O estudo que publicou esse conjunto de dados deixa claro que a amostra não foi aleatória e que portanto qualquer inferência para a população em geral não é recomendada. Nós fazemos isso aqui apenas como um exercício.

```
1 - pnorm(q = 182, mean = fhgtmean, sd = fhgtsd)
```

Perceba que a função `pnorm` dá como resultado a área sob a curva normal abaixo de um certo valor, `q`, com uma dada média e desvio padrão. Uma vez que estamos interessados na probabilidade de que alguém seja maior do que 182 cm, precisamos calcular 1 menos essa probabilidade.

Presumindo uma distribuição normal nos permitiu calcular uma probabilidade teórica. Se queremos calcular a probabilidade empiricamente, simplesmente precisamos determinar quantas observações se encontram acima de 182 e então dividir este número pelo tamanho total da amostra.

```
sum(fdims$hgt > 182) / length(fdims$hgt)
```

Apesar das probabilidades não serem exatamente as mesmas, elas estão perto o suficiente. Quanto mais perto sua distribuição está da normal, mais precisas as probabilidades teóricas serão.

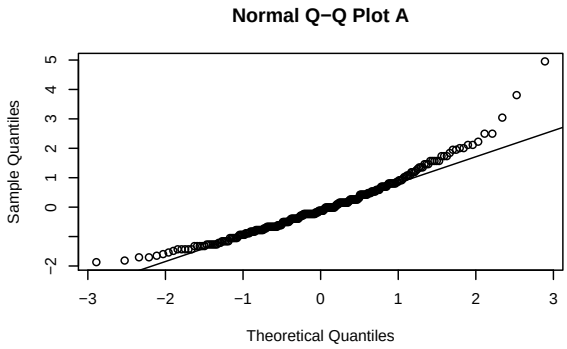
> **Exercício 6** Elabore duas questões de probabilidade que você gostaria de responder; uma com relação à altura de mulheres e outra com relação ao peso de mulheres. Calcule essas probabilidades usando tanto o método teórico da distribuição normal quanto a distribuição empírica (quatro probabilidade no total). Qual variável, altura ou peso, teve uma concordância maior entre os dois métodos?

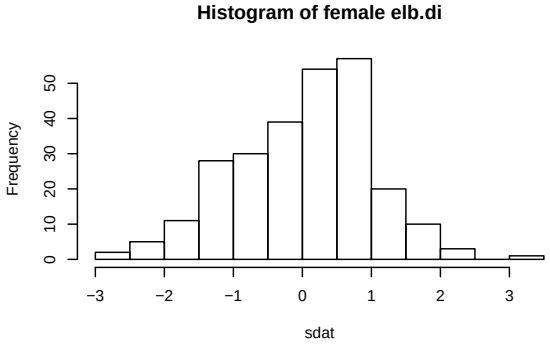
Sua Vez

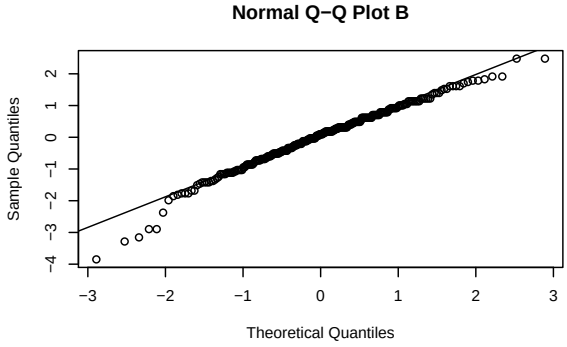
1. Agora vamos analisar outras variáveis no conjunto de dados das dimensões corporais. Utilizando as figuras na próxima página, combine os histogramas com seus gráficos de probabilidade normal. Todas as variáveis foram estandardizadas (primeiro subtraindo a média, e em seguida dividindo pelo desvio padrão), de tal forma que as unidades não serão de qualquer ajuda. Se você estiver incerto com base nessas figuras, gere um gráfico no R para verificar.
   (a) O histograma do diâmetro bi-ilíaco (pélvico) feminino (`bii.di`) pertence ao gráfico de probabilidade normal de letra ____.
   (b) O histograma do diâmetro do cotovelo feminino (`elb.di`) pertence ao gráfico de probabilidade normal de letra ____.
   (c) O histograma de idade geral (`age`) pertence ao gráfico de probabilidade normal de letra ____.
   (d) O histograma de profundidade do peito feminino (`che.de`) pertence ao gráfico de probabilidade normal de letra ____.
2. Perceba que os gráficos de probabilidade normal C e D tem um pequeno padrão passo a passo. Por que você acha que eles são assim?
3. Como você pode ver, gráficos de probabilidade normal podem ser utilizados tanto para avaliar a normalidade quanto visualizar a assimetria. Crie um gráfico de probabilidade normal para o diâmetro do joelho feminino (`kne.di`). Baseado neste gráfico de probabilidade normal, você diria que essa variável é simétrica, assimétrica à direita ou assimétrica à esquerda? Utiliza um histograma para confirmar seu resultado.

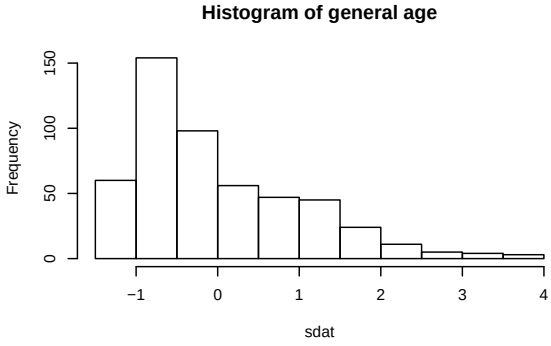
4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

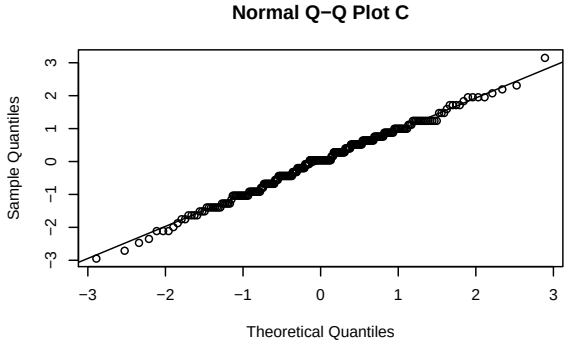

Histogram of female bii.di
Frequency
sdat

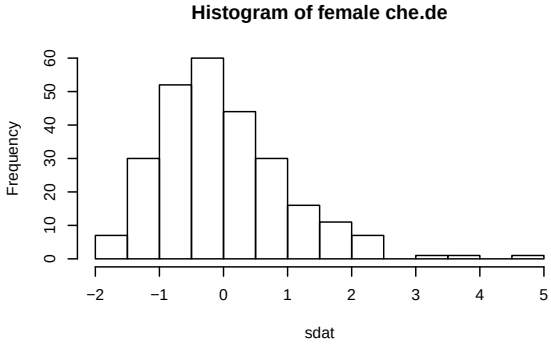

Normal Q−Q Plot A
Sample Quantiles
Theoretical Quantiles

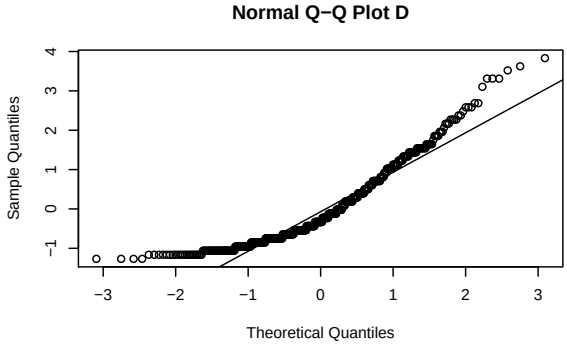

Histogram of female elb.di
Frequency
sdat


Normal Q−Q Plot B
Sample Quantiles
Theoretical Quantiles


Histogram of general age
Frequency
sdat


Normal Q−Q Plot C
Sample Quantiles
Theoretical Quantiles


Histogram of female che.de
Frequency
sdat


Normal Q−Q Plot D
Sample Quantiles
Theoretical Quantiles

Laboratório 4A: Fundamentos para Inferência Estatística - Distribuições Amostrais

Neste laboratório, investigaremos os meios pelos quais as estatísticas de uma amostra aleatória de dados podem servir como estimativas pontuais de parâmetros populacionais. Estamos interessados em formular uma *distribuição amostral* de nossa estimativa para aprender sobre as propriedades da estimativa, como sua distribuição.

Os Dados

Vamos analisar dados do setor imobiliário da cidade de Ames, no estado de Iowa, Estados Unidos. Os detalhes de cada transação imobiliária na cidade de Ames é registrada pelo escritório da Secretaria Municipal da Receita da cidade. Nosso foco particular para este laboratório será todas as vendas de casa em Ames entre 2006 e 2010. Essa coleção representa nossa população de interesse. Neste laboratório queremos aprender sobre essas vendas de casa retirando pequenas amostra da população completa. Vamos importar os dados.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")

load("ames.RData")
```

Vemos que há muitas variáveis em nosso conjunto de dados, o suficiente para realizar uma análise aprofundada. Para este laboratório, restringiremos nossa atenção para somente duas variáveis: a área habitável da casa acima do nível do solo em pés quadrados (`Gr.Liv.Area`) e o preço da venda (`SalePrice`). Para economizar esforços ao longo do laboratório, crie duas variáveis com nomes curtos para representar essas duas variáveis do conjunto de dados.

```
area <- ames$Gr.Liv.Area

price <- ames$SalePrice
```

Vamos dar uma olhada na distribuição da área em nossa população de vendas de casas calculando algumas estatísticas sumárias e criando um histograma.

```
summary(area)

hist(area)
```

Exercício 1 Descreva a distribuição da população.

A Distribuição Amostral Desconhecida

Neste laboratório nós temos acesso à população inteira, mas isso raramente acontece na vida real. Reunir informação sobre uma população inteira costuma ser muito custoso ou impossível. Por essa razão,

frequentemente retiramos uma amostra da população e a utilizamos para compreender propriedades da população.

Se estivermos interessados em estimar a área habitável média na cidade de Ames com base numa amostra, podemos utilizar o seguite comando para sondar a população.

```
samp1 <- sample(area, 50)
```

Esse comando retira uma amostra aleatória simples de tamanho 50 do vetor `area`, que é atribuída à variável `samp1`. É como se fôssemos ao banco de dados da Secretaria Municipal da Fazendo e retirássemos os arquivos de 50 vendas de casas aleatoriamente. Trabalhar com esses 50 arquivos seria consideravelmente mais simples do que lidar com todas as 2930 vendas de casas.

> **Exercício 2** Descreva a distribuição desta amostra. Como ela se compara à distribuição da população?

Se estamos interessados em estimar a área habitável média nas casas da cidade de Ames utilizando esta amostra, nossa melhor suposição é a média da amostra.

```
mean(samp1)
```

Dependendo de quais foram as 50 casas que foram sorteadas, sua estimativa como estar um pouco acima ou abaixo da média populacional verdadeira de 1499,69 pés quadrados. De maneira geral, mesmo assim, a média da amostra costuma ser uma estimativa muito boa da média da área habitável, e nós a obtemos por meio de uma amostra de menos de 3% da população.

> **Exercício 3** Retire uma segunda amostra, também de 50 casos, e a atribua a uma variável de nome `samp2`. Como a média de `samp2` se compara à média de `samp1`? Vamos supor que retiremos mais duas amostras, uma de de 100 casos e outra de 1000 casos. Qual você acha que daria uma estimativa mais precisa da média populacional?

Não é surpreendente que, a cada vez que retiramos uma nova amostra aleatória, obtemos uma média amostral diferente. É útil ter uma ideia de quanta variabilidade podemos esperar quando estimamos a média populacional desta maneira. A distribuição das médias amostrais, denominada de *distribuição amostral*, pode nos ajudar a compreeder essa variabilidade. Neste laboratório, uma vez que temos acesso à população, podemos elaborar a distribuição amostral para a média amostral repetindo os passos acima várias vezes. Agora geraremos 5000 amostras e calcularemos a média amostra de cada uma.

```
sample_means50 <- rep(0, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   }

hist(sample_means50)
```

Se você quiser ajustar a largura dos segmento do seu histograma para exibir um pouco mais de detalhes, você pode fazê-lo mudando o argumento `breaks`.

```
hist(sample_means50, breaks = 25)
```

Nós utilizamos o R para retirar 5000 amostras de 50 casos da população geral, calcular a média de cada amostra, e registrar cada resultado num vetor denominado `sample_means50`. Na próxima página, compreenderemos como esse conjunto de códigos funciona.

> **Exercício 4** A variável `sample_means50` contém quantos elementos? Descreva a distribuição amostral, e certifique-se de prestar atenção especificamente em seu centro. Você acha que a distribuilção mudaria se coletássemos 50.000 médias amostrais?

Interlúdio: O Comando `for` para Repetições

Vamos nos afastar da estatística por um momento para comprender melhor o último bloco de código. Você acabou de rodar seu primeiro *loop*, uma repetição de uma mesma sequência de instruções que é fundamental para a programação de computadores. A ideia por trás do *loop* é a noção de *iteração*: ele permite que você execute um código quantas vezes quiser sem ter que digitar cada iteração. Na caso acima, nós queríamos iterar as duas linhas de código que estão dentro das chaves, que retiram uma amostra aleatória de 50 casos da variável `area` e então salva a média da amostra no vetor `sample_means50`. Sem o *loop*, programar isso seria tedioso:

```
sample_means50 <- rep(0, 5000)

samp <- sample(area, 50)
sample_means50[1] <- mean(samp)

samp <- sample(area, 50)
sample_means50[2] <- mean(samp)

samp <- sample(area, 50)
sample_means50[3] <- mean(samp)

samp <- sample(area, 50)
sample_means50[4] <- mean(samp)
```

e assim por diante...

Usando o comando `for` para implementar um *loop*, essas milhares de linhas de código são comprimidas em um punhado de linhas. Adicionamos uma linha extra ao código abaixo, que imprime a variável `i` em cada iteração do *loop*. Rode este código.

```
sample_means50 <- rep(0, 5000)

for(i in 1:5000){
   samp <- sample(area, 50)
   sample_means50[i] <- mean(samp)
   print(i)
   }
```

Vamos examinar este código linha a linha para compreender o que ele faz. Na primeira linha nós *inicializamos um vetor*. Nesse caso, criamos um vetor com 5000 zeros denominado `sample_means50`. Esse vetor armazenará os valores gerados dentro do *loop*.

A segunda linha executa o *loop*. A sintaxe pode ser lida mais ou menos como, "para cada elemento `i` de 1 a 5000, execute as seguintes linhas de código". Você pode interpretar o `i` como um contador que mantém o registro de qual *loop* você está. Portanto, mais precisamente, o *loop* será executado uma vez quando `i=1`, e então mais uma vez quando `i=2`, e assim por diante até `i=5000`.

A parte principal do *loop* se encontra dentro das chaves, e esse conjunto de linhas de código é executado para cada valor de `i`. Aqui, em cada iteração, selecionamos uma amostra aleatória de 50 elementos a partir da variável `area`, calculamos sua média, e registramos seu valor como o $i^{\text{ésimo}}$ elemento do vetor `sample_means50`.

Para demonstrar que isso está de fato acontecento, pedimos ao R para imprimir o valor de `i` em cada iteração. Esta linha de código é opcional e é usada somente para mostrar o que está acontecendo enquando o *loop* do comando `for` está em execução.

O *loop* nos permite não somente rodar o código 5000 vezes, mas também armazenar os resultados ordenadamente, elemento por elemento, num vetor vazio que inicializamos nas primeiras linhas.

> **Exercício 5** Para certificar que você compreendeu o que você fez neste *loop*, experimente rodar uma versão menor. Inicialize um vetor com 100 zeros com o nome `sample_means_small`. Execute um *loop* que retira uma amostra de 50 elementos da variável `area` e armazena o média amostral no vetor `sample_means_small`, mas que repete a iteração de 1 a 100. Imprima o resultado em sua tela (basta digitar `sample_means_small` no console e pressionar enter). Há quantos elementos no objeto `sample_means_small`? O que cada elemento representa?

Tamanho da Amostra e Distribuição Amostral

À parte dos aspectos mecânicos de programação, vamos retomar a razão pela qual utilizamos o *loop* do comando `for`: para calcular uma distribuição amostral, especificamente, esta aqui:

```
hist(sample_means50)
```

A distribuição amostral que calculamos nos informa bastante sobre as estimativas da área habitável nas casas na cidade de Ames. Uma vez que a média amostral é um estimador não-enviesado, a distribuição amostral está centrada na verdadeira média da área habitável da população, e a dispersão da distribuição indica quanta variabilidade é possível ao se amostra somente 50 vendas de casas.

Para ter uma ideia melhor do efeito do tamanho da amostra na distribuição amostral, vamos construir mais duas distribuições amostrais: uma baseada numa amostra de 10 elementos e outra baseada numa amostra de 100.

```
sample_means10 <- rep(0, 5000)
sample_means100 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(area, 10)
  sample_means10[i] <- mean(samp)
  samp <- sample(area, 100)
  sample_means100[i] <- mean(samp)
}
```

Aqui podemos utilizar um único *loop* para construir duas distribuições adicionando mais algumas linhas dentro das chaves. Não se preocupe com o fato de que `samp` é utilizado como o nome de dois objetos diferentes. No segundo comando do *loop*, a média de `samp` é salva em seu devido lugar no vetor `sample_means10`. Com a média já salva, podemos sobrescrever o objeto `samp` com uma nova amostra,

desta vez de com 100 elementos. De maneira geral, quando você cria um objeto utilizando um nome que já está em uso, o objeto antigo será substituído pelo novo.

Para verificar o efeito que diferentes tamanhos de amostra tem na distribuição amostral, crie gráficos das três distribuições, um em cima do outro.

```
par(mfrow = c(3, 1))

xlimits = range(sample_means10)

hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

O primeiro comando especifica que você quer dividir a área do gráfico em três linhas e uma coluna para cada um dos gráficos[†]. O argumento `breaks` ("quebras") especifica o número de segmentos utilizados para construir o histograma. O argumento `xlim` especifica o intervalo no eixo x no hisograma, e ao definilo como igual a `xlimits` para cada histograma, certificamo-nos de que todos os três histogramas serão criados com os mesmos limites no eixo x.

> **Exercício 6** Quando o tamanho da amostra é maior, o que acontece com o centro da distribuição? E com a dispersão?

Sua Vez

Até agora, nós nos ocupamos em estimar a média da área habitável nas casas do município de Ames. Agora você tentará estimar a média dos preços das casas.

1. Retire uma amostra aleatória de 50 elementos da variável `price` (`preço`). Com essa amostra, qual é sua melhor estimativa pontual para a média populacional?
2. Já que você tem acesso à população, simule a distribuição amostral de $\bar{x}_{price}$ retirando 5000 amostras de 50 elementos da população e calculando 5000 médias amostrais. Armazene essas médias em um vetor com o nome `sample_means50`. Crie um gráfico com os resultados, e então descreva a forma dessa distribuição amostral. Baseado nessa distribuição amostral, qual seria seu palpite para a média dos preços das casas na população? Por fim, calcule e informe a média populacional.
3. Mude o tamanho da sua amostra de 50 para 150, e então calcule a distribuição amostral utilizando o mesmo método descrito acima, e guarde as médias em um novo vetor com o nome `sample_means150`. Descreva a forma dessa distribuição amostral e compare-a com a distribuição amostral para a amostra de 50 elementos. Com base nessa distribuição amostral, qual seria seu palpite sobre a média dos preços de vendas de casas no município de Ames?
4. Das distribuições amostrais calculadas nos exercícios 2 e 3, qual tem menor dispersão? Se estamos interessados em estimativas que estão mais próximas do valor verdadeiro, preferiríamos uma distribuição com uma dispersão pequena ou grande?

[†]Talvez você precise esticar um pouco sua janela com os gráficos para acomodar os gráficos extras. Para retornar para a configuração padrão de criar um gráfico por vez, rode o seguinte comando:

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

5. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

Laboratório 4B: Fundamentos para Inferência Estatística - Intervalos de Confiança

Amostragem de Ames, Iowa

Se você tem acesso aos dados de um população inteira, por exemplo o tamanho de cada casa na cidade de Ames, Iowa, Estado Unidos, é fácil e direto responder a questões como "Qual é o tamanho de uma casa típica na cidade de Ames?" e "Quanta variação existe no tamanho das casas?". Se você tem acesso somente a uma amostra da população, como costuma ser o caso, responder a essas perguntas fica mais complicado. Qual é sua melhor estimativa para o tamanho típico de uma casa se você só sabe o tamanho de algumas dezenas de casas? Esse tipo de situação requer que você use sua amostra para fazer inferências a respeito da população em geral.

Os Dados

Na laboratório anterior nós exploramos os dados populacionais das casa da cidade de Ames, Iowa. Vamos começar carregando esse conjunto de dados.

```
download.file("http://www.openintro.org/stat/data/ames.RData", destfile = "ames.RData")

load("ames.RData")
```

Neste laboratório começaremos com uma amostra aleatória simples de 60 elementos da população. Perceba que o conjunto de dados contém informações sobre várias variáveis relativas às casas, mas para a primeira parte do laboratório focaremos no tamanho da casa, representada pela variável `Gr.Liv.Area`.

```
population <- ames$Gr.Liv.Area

samp <- sample(population, 60)
```

Exercício 1 Descreva a distribuição da sua amostra. Qual é o tamanho "típico" dentro da sua amostra? Procure esclarecer também como você interpretou o significado de "típico".

Exercício 2 Você acha que a distribuição de outro aluno seria idêntica a sua? Você acha que ela seria similar? Por quê, ou por quê não?

Intervalos de Confiança

Uma das maneiras mais comuns para se descrever o valor típico ou central de uma distribuição é por meio da média. Neste caso podemos calcular a média da amostra utilizando

```
sample_mean <- mean(samp)
```

Retome agora a pergunta que motivou este laboratório: baseado nesta amostra, o que podemos inferir sobre a população? Baseado apenas nesta única amostra, a melhor estimativa da área habitacional das casas vendidas em Ames seria a média amostral, geralmente representada como $\bar{x}$ (aqui denominaremos de `sample_mean` ("média amostral")). A média amostral serve como uma boa *estimativa pontual*, mas seria interessante também deixar claro quanta incerteza temos desta estimativa. Isso pode ser feito pelo uso de um *intervalo de confiança*.

Podemos calcular um intervalo de confiança de 95% para a média amostral adicionando e subtraindo 1.96 erros padrão da estimativa pontual.[†]

```
se <- sd(samp)/sqrt(60)

lower <- sample_mean - 1.96 * se

upper <- sample_mean + 1.96 * se

c(lower, upper)
```

Acabamos de fazer uma inferência importante: mesmo que não saibamos como a população inteira se distribui, temos 95% de confiança de que a média verdadeira do tamanho das casas em Ames se encontra entre os valores `lower` (limite inferior do intervalo de confiança) e `upper` (limite superior do intervalo de confiança). Contudo, existem algumas condições que precisam ser atendidas para esse intervalo ser válido.

> **Exercício 3** Para o intervalos de confiança ser válido, a média amostral precisa ter distribuição normal e ter um erro padrão igual a $s/\sqrt{n}$. Quais condições precisam ser atendidas para isso ser verdadeiro?

Níveis de Confiança

> **Exercício 4** O que significa "95% de confiança"? Se você não tem certeza, retome a Seção 4.2.2.

Neste caso nós temos a comodidade de saber a verdadeira média populacional, uma vez que temos os dados da população inteira. Este valor pode ser calculado utilizando o seguinte comando:

```
mean(population)
```

> **Exercício 5** O seu intervalo de confiança contém a verdadeira média do tamanho das casas em Ames? Se você está trabalhando neste laboratório em uma sala de aula, o intervalo de seus colegas também contém esse valor?

> **Exercício 6** Cada aluno de sua turma deve ter obtido um intervalo de confiança um pouco diferente. Que proporção desses intervalos você espera que contenha a verdadeira média populacional? Por quê? Se você está trabalhando neste laboratório em um sala de aula, reúna informações sobre os intervalos criados pelos outros alunos da turma e calcule a proporção de intervalos que contém a verdadeira média populacional.

Utilizando o R, vamos criar várias amostra para aprender um pouco mais a respeito de como as médias

[†]Confira a seção 4.2.3 se você não está familiarizado com essa fórmula.

amostrais e os intervalos de confiança variam de uma amostra para outra. *Loops* são úteis para isso.[§]

Eis o esboço do processo:

(1) Obter uma amostra aleatória.

(2) Calcular a média e o desvio padrão da amostra.

(3) Utilizar estas estatísticas para calcular um intervalos de confiança.

(4) Repetir as etapas (1)-(3) 50 vezes.

Mas antes de implementar esse processo, precisamos primeiro criar vetores vazios nos quais possamos salvar as médias e desvios padrão que serão calculados para cada amostra. Ao mesmo tempo, vamos também armazenar o tamanho da amostra como `n`.

```
samp_mean <- rep(NA, 50)

samp_sd <- rep(NA, 50)

n <- 60
```

Agora estamos prontos para o *loop*, com o qual calculamos as médias e desvios padrão de 50 amostras aleatórias.

```
for(i in 1:50){
  samp <- sample(population, n) # obtém uma amostra de n = 60 elementos da população
  samp_mean[i] <- mean(samp)    # salva a média amostral no i-ésimo elemento de samp_mean
  samp_sd[i] <- sd(samp)        # salva o dp da amostra como o i-ésimo elemento de samp_sd
}
```

Por fim, construímos os intervalos de confiança.

```
lower_vector <- samp_mean - 1.96 * samp_sd / sqrt(n)

upper_vector <- samp_mean + 1.96 * samp_sd / sqrt(n)
```

Os limites inferiores destes 50 intervalos de confiança são armazenados na vetor `lower_vector`, e o limites superiores são armazenados no vetor `upper_vector`. Vamos visualizar o primeiro intervalo.

```
c(lower_vector[1],upper_vector[1])
```

Sua Vez

1. Utilizando a seguinte função (que foi carregada junto com o conjunto de dados), crie gráficos de todos os intervalos. Que proporção dos intervalos de confiança contém a verdadeira média populacional?

[§]Se você não está familiarizado com *loops*, revise o Laboratório 4A.

Essa proporção é exatamente igual ao nível de confiança? Se não, explique por quê.[†]

```
plot_ci(lower_vector, upper_vector, mean(population))
```

2. Escolha um intervalo de confiança de sua preferência, desde que não seja de 95%. Qual é o valor crítico apropriado?
3. Calcule 50 intervalos de confiança utilizando o nível de confiança que você escolheu na questão anterior. Você não precisa obter novas amostras: simplesmente calcule os novos intervalos baseado nas médias amostrais e desvios padrão que você já coletou. Utilizando a função `plot_ci`, crie gráficos de todos os intervalos e calcule a proporção de intervalos que contém a verdadeira média populacional. Compare essa proporção com o nível de confiança escolhido para os intervalos.
4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

[†]Essa figura pode parecer familiar (Verifique a Seção 4.2.2.)

Laboratório 5: Inferência para Dados Numéricos

Nascimentos na Carolina do Norte

Em 2004, o estado da Carolina do Norte, Estado Unidos, disponibilizou um grande conjunto de dados contendo informações sobre os nascimentos registrados no estado. Esse conjunto de dados é útil para pesquisadores que estudam a relação entre hábitos e práticas de gestantes e o nascimento de seus filhos. Nós trabalharemos com uma amostra aleatória das observações deste conjunto de dados.

Análise Exploratória

Carregue o conjunto de dados `nc` em seu espaço de trabalho.

```
download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")

load("nc.RData")
```

Temos dados de 13 variáveis diferentes, algumas categoriais e outras numéricas. Cada variável representa as seguintes informações:

`fage`	idade do pai em anos.
`mage`	idade da mãe em anos.
`mature`	maioridade da mãe.
`weeks`	duração da gestação em semanas.
`premie`	se o nascimento é classificado como prematuro ou a termo.
`visits`	número de visitas hospitalares durante a gravidez.
`marital`	se a mãe estava `casada` (`married`) ou `solteira` (`not married`) no momento do nascimento.
`gained`	peso ganho pela mãe durante a gravidez, em libras.
`weight`	peso do bebê no nascimento, em libras.
`lowbirthweight`	se o bebê foi classificado como tendo baixo peso ao nascer (`low`) ou não (`not low`).
`gender`	sexo do bebê, `feminino` (`female`) ou `masculino` (`male`).
`habit`	se a mãe é `não-fumante` (`nonsmoker`) ou `fumante` (`smoker`).
`whitemom`	se a mãe é `branca` (`white`) ou `não-branca` (`not white`).

Exercício 1 Quais são os casos neste conjunto de dados? Há quantos casos em nossa amostra?

Como um primeiro passo na análise, devemos levar em consideração alguns sumários dos dados. Isso pode ser feito utilizando o comando `summary` ("sumário"):

```
summary(nc)
```

Enquanto você confere os sumários das variáveis, considere quais variáveis são categoriais e quais são numéricas. Para as variáveis numéricas, há algum caso atípico, um *outlier*? Se você não tem certeza ou quer dar uma olhada mais aprofundada nos dados, crie um gráfico.

Considere a possibilidade de uma relação entre o hábito de fumar da mãe e o peso de seu bebê. Criar um gráfico com os dados é uma etapa útil porque nos ajuda a visualizar tendências rapidamente, identificar associações fortes, e elaborar questões de pesquisa.

Exercício 2 Crie um gráfico de caixas lado-a-lado das variáveis `habit` (`hábito`) e `weight` (`peso`). O que o gráfico revela sobre a relação entre essas duas variáveis?

O gráfico de caixas permite comparar as medianas das distribuições, mas podemos também comparar as médias das distribuições utilizando a seguinte função para dividir a variável `weight` nos grupos definidos pela variável `habit`, e então calcular a média de cada um utilizando a função `mean`.

```
by(nc$weight, nc$habit, mean)
```

Há uma diferença evidente, mas essa diferença é estatisticamente significante? Para responder a essa questão, vamos realizar um teste de hipótese.

Inferência

Exercício 3 Verifique se as condições necessárias para realizar a inferência são atendidas. Perceba que você precisará obter o tamanho das amostras para verificar as condições. Você pode calcular o tamanho dos grupos utilizando o mesmo comando `by` utilizado acima, mas substituindo a função `mean` pela função `length`.

Exercício 4 Escreva as hipóteses para testar se a média dos pesos dos bebês que nasceram de mães fumantes é diferente daqueles que nasceram de mães não fumantes.

Em seguida, utilizaremos uma nova função, `inference`, que será utilizada para realizar os testes de hipótese e para construir os intervalos de confiança.

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0,
          alternative = "twosided", method = "theoretical")
```

Vamos com calma para analisar cada argumento desta função personalizada.

- O primeiro argumento é `y`, que é a variável resposta na qual estamos interessados: `nc$weight` (peso).
- O segundo argumento é a variável explicativa, `x`, que é a variável que divide os dados em dois grupos, fumantes e não fumantes: `nc$habit`.
- O terceiro argumento, `est`, é o parâmetro no qual estamos interessados: `"mean"` (média) (há outras opções: `"median"` (mediana), ou `"proportion"` (proporção)).
- Em seguida decidimos sobre o tipo de inferência que queremos (`type`): um teste de hipótese (`"ht"`) ou um intervalo de confiança (`"ci"`).
- Quando realizamos um teste de hipótese, também precisamos informar o valor nulo (`null`), que neste caso é `0`, já que a hipótese nula supõe que as duas médias populacionais são iguais uma a outra.
- A hipótese alternativa (`alternative`) pode ser `"less"` (menor), `"greater"` (maior), ou `"twosided"` (bi-caudal).
- Por fim, o método (`method`) de inferência pode ser `"theoretical"` (teórico) ou `"simulation"` (baseado em simulações).

Exercício 5 Mude o argumento `type` (tipo) para `"ci"` para construir e registrar um intervalo de confiança para a diferença entre os pesos dos bebê que nasceram de mães fumantes e não fumantes.

Por padrão, a função utilizada informa um intervalo para a diferença ($\mu_{não-fumante} - \mu_{fumante}$) (a diferença entre médias dos dois grupos). Podemos mudar facilmente essa ordem utilizando o argumento `order` (ordem):

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          order = c("smoker","nonsmoker"))
```

Sua Vez

1. Calcule o intervalo de confiança de 95% para a duração média das gravidezes (`weeks`) e o interprete no contexto do conjunto de dados. Perceba que, uma vez que você está realizando uma inferência sobre um único parâmetro populacional, não há nenhuma variáveis explanatória, e portanto você pode omitir a variável `x` da função.
2. Calcule um novo intervalo de confiança para o mesmo parâmetro com nível de confiança de 90%. Você pode mudar o nível de confiança adicionando um novo argumento à função: `conflevel =0.90`.
3. Realize um teste de hipótese para avaliar se o a média do peso ganho pelas mães mais jovens é diferente da média de peso ganho pelas mães mais velhas.
4. Agora, um tarefa não-inferencial: determine o ponto de corte da idade das mães jovens e maduras. Utilize um método da sua escolha, e explique como seu método funciona.
5. Escolha um par de variáveis, sendo uma numérica e outra categorial, e desenvolva um pergunta de pesquisa para avaliar a relação entre essas variáveis. Formule a questão de maneira que ela possa ser respondida utilizando um teste de hipótese e/ou um intervalo de confiança. Responda a sua questão utilizando a função `inference`, informe os resultados estatísticos, e também elabora uma explicação em linguagem simples.
6. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

Laboratório 6: Inferência para Dados Categoriais

Em agosto de 2012, agências de notícias como Washington Post e o Huffington Post publicaram reportagens sobre o aumento do ateísmo na América do Norte. A fonte da reportagem foi uma pesquisa que perguntou às pessoas, "Independente de você frequentar algum culto religioso ou não, você diria que você é uma pessoa religiosa, não é uma pessoa religiosa ou é um ateu convicto?" Esse tipo de pergunta, que pede para as pessoas se classificarem de uma forma ou outra, é comum em pesquisas de opinião e gera dados categoriais. Neste laboratório vamos explorar a pesquisa sobre ateísmo e investigar o que está em jogo quando fazemos inferências sobre proporções populacionais utilizando dados categoriais.

A Pesquisa de Opinião

Para acessar o comunicado à imprensa da pesquisa de opinião, realizada pela WIN-Gallup International, clique no link abaixo:

http://www.wingia.com/web/files/richeditor/filemanager/Global_INDEX_of_Religiosity_and_Atheism_PR__6.pdf

Revise com cuidado as informações do relatório e então tente resolver as seguintes questões:

Exercício 1 No primeiro parágrafo, vários resultados importantes são relatados. Essas porcentagens parecem ser *estatísticas amostrais* (derivadas dos dados da amostra) ou *parâmetros populacionais*?

Exercício 2 O título do relatório é "Índice Global de Religiosidade e Ateísmo" ("Global Index of Religiosity and Atheism"). Para generalizar os resultados do relatório para a população humana global, o que devemos assumir a respeito do método amostral? Parece ser uma suposição razoável?

Os Dados

Preste atenção na Tabela 6 (páginas 15 e 16), que informa o tamanho da amostra e o percentual de respostas de todos os 57 países que fizeram parte da pesquisa. Mesmo sendo um formato útil para resumir os dados, basearemos nossas análises no conjunto de dados original das respostas individuais à pesquisa. Carregue esse conjunto de dados no R utilizando o seguinte comando.

```
download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")
load("atheism.RData")
```

Exercício 3 A que corresponde cada linha da Tabela 6? A que corresponde cada linha do banco de dados `atheism` (ateísmo)?

Para investigar o elo entre essas duas maneiras de organizar esses dados, dê uma olhada na proporção estimada de ateus nos Estados Unidos. Perto do fim da Tabela 6, verificamos que é 5%. Devemos ser capazes de chegar ao mesmo número usando o banco de dados `atheism`.

Exercício 4 Utilizando o comando abaixo, crie um novo banco de dados denominado `us12` que contém apenas as linhas do banco de dados `atheism` associadas aos respondentes da pesquisa

realizada em 2012 nos Estados Unidos. Em seguida, calcule a proporção de respostas dos que se afirmam ateus. Ela é semelhante à porcentagem da Tabela 6? Se não, por quê?

```
us12 <- subset(atheism, atheism$nationality == "United States" & atheism$year == "2012")
```

Inferência de Proporções

Como foi sugerido pelo Exercício 1, a Tabela 6 apresenta *estatísticas*, ou seja, cálculos feitos a partir da amostra de 51.927 pessoas. O que nós gostaríamos, porém, é obter informações sobre os *parâmetros* populacionais. Você pode responder à pergunta "Qual a proporção de pessoas na amostra que informaram serem ateus?" com uma estatística; por outro lado, uma questão como "Qual a proporção de pessoas na Terra que informariam serem ateus?" é respondida com uma estimativa do parâmetro.

As ferramentas inferenciais para estimar proporções populacional são análogas àquelas utilizadas para as médias no último laboratório: o intervalo de confiança e o teste de hipótese.

> **Exercício 5** Descreva as condições para inferência necessárias para construir um intervalo de confiança de 95% para a proporção de ateus nos Estados Unido em 2012. Você está confiante de que todas as condições são atendidas?

Se as condições para inferência são razoáveis, podemos calcular o erro padrão e construir o intervalo de confiança manualmente, ou deixar que a função `inference` faça isso por nós.

```
inference(y = us12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

Perceba que, uma vez que o objetivo é construir uma estimativa intervalar para uma proporção, é necessário especificar o que constitui um "sucesso", que nesse caso é a resposta `atheist` (ateu).

Apesar de intervalos de confiança formais e testes de hipótese não aparecerem no relatório, sugestões de inferência aparecem no final da página 7: "Em geral, a margem de erro para pesquisas de opinião deste tipo é de $\pm 3 - 5\%$ com 95% de confiança."

> **Exercício 6** Com base nos resultados do R, qual é a margem de erro para a estimativa da proporção de ateus nos EUA em 2012?

> **Exercício 7** Utilizando a função `inference`, calcule os intervalos de confiança para a proporção de ateus em 2012 para dois outros países de sua escolha, e informe as margens de erro associadas a eles. Certifique-se de observar se as condições para inferência são atendidas. Pode ser útil primeiro criar novos conjuntos de dados para cada um dos dois países, e então usar essas conjuntos de dados junto com a função `inference` para construir os intervalos de confiança.

Como a Proporção Afeta a Margem de Erro?

Imagine que você fez um levantamento com 1000 pessoas a respeito de duas questões: você é mulher? E você é canhoto? Uma vez que ambas as proporções amostrais foram calculadas a partir de um mesmo tamanho de amostra, elas devem ter a mesma margem de erro, certo? Errado! Apesar da margem de erro mudar em relação ao tamanho da amostra, ela também é afetada pela proporção.

Lembre-se da fórmula para calcular o erro padrão: $EP = \sqrt{p(1-p)/n}$. O resultado é utilizado na fórmula para calcular a margem de erro para um intervalo de confiança de 95%: $ME = 1.96 \times EP =$

$1.96 \times \sqrt{p(1-p)/n}$. Já que a proporção populacional p se encontra na fórmula para calcular o ME, faz sentido que a margem de erro depende, de alguma forma, da proporção populacional. Podemos visualizar essa relação criando um gráfico relacionando ME com p.

O primeiro passo é criar um vetor `p`, que é uma sequência de 0 a 1 com cada número separado por 0,01. Podemos então criar um vetor para a margem de erro (`me`), associando com cada um dos palores de `p` utilizando a fórmula aproximada já conhecida ($ME = 2 \times SE$). Por fim, fazemos um gráfico com os dois vetores para revelar a relação entre eles.

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2*sqrt(p*(1 - p)/n)
plot(me ~ p)
```

Exercício 8 Descreva a relação entre `p` e `me`.

Condição de Sucesso ou Fracasso

O livro enfatiza que você deve sempre verificar as condições antes de fazer qualquer inferência. Para inferência de proporções, a proporção amostral pode ser considerada como se distribuindo de maneira aproximadamente normal se for baseada numa amostra aleatória de observações independentes e se $np \geq 10$ e $n(1-p) \geq 10$. Essa regra geral é fácil o suficiente de ser seguida, mas deixa aberta a questão: o que há de tão especial com o número 10? A resposta mais curta é: nada. Você pode argumentar que estaríamos bem com 9 ou que deveríamos utilizar 11. O "melhor" valor para essa regra geral é, pelo menos em alguma medida, arbitrário.

Podemos investigar as relações entre n e p e a forma da distribuição amostral utilizando simulações. Para começar, simulamos o processo de retirar 5000 amostra de 1040 elementos de uma população com a verdadeira proporção de ateus igual a 0.1. Para cada uma das 5000 amostras, calculamos $\hat{p}$ e então criamos um histograma para visualizar sua distribuição.

```
p <- 0.1
n <- 1040
p_hats <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats[i] <- sum(samp == "atheist")/n
}

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
```

Esses comandos constroem a distribuição amostral de `p_hats` por meio do *loop* do comando `for` que já nos é familiar. Você pode pode ler o procedimentos amostral da primeira linha de código dentro do *loop* como "retire uma amostra com reposição de n elementos a partir das opções de ateu e não-ateu com probabilidades p e $1-p$, respectivamente." A segunda linha do *loop* diz "calcule a proporção de ateus nesta amostra e registre e valor." O *loop* nos permite repetir esse processo 5.000 vezes para construir uma boa representação da distribuição amostral.

Exercício 9 Descreva a distribuição amostral da proporção com $n = 1040$ e $p = 0.1$. Certifique-se de identificar seu centro, dispersão e forma.
Dica: Lembre-se que o R tem funções como `mean` para calcular estatísticas descritivas.

Exercício 10 Repita a simulação acima mais três vezes mas com diferentes tamanhos de amostra e proporções: com $n = 400$ e $p = 0.1$, $n = 1040$ e $p = 0.02$, e $n = 400$ e $p = 0.02$. Crie histogramas para as quatro distribuições e exiba-os em conjunto utilizando o comando `par(mfrow = c(2,2))`. Talvez você precise expandir a janela do gráfico para acomodar o gráfico maior. Descreva as três distribuições amostrais novas. Com base nesses gráficos limitados, como que n parece afetar a distribuição de $\hat{p}$? Como que p afeta a distribuição amostral?

Depois de terminar, você pode resetar a disposição da janela de gráfico utilizando o comando `par(mfrow = c(1,1))` ou clicando no botão "Clear All" ("Limpar Tudo") logo acima da janela de gráficos (se estiver usando o RStudio). Preste atenção pois a última opção irá apagar todos os gráficos anteriores.

Exercício 11 Se você retomar a Tabela 6, verá que a Austrália tem uma proporção amostral de 0,1 numa amostra de 1040, e que o Equador tem uma proporção amostral de 0,02 com 400 sujeitos. Vamos supor, para esse exercício, que essas estimativas pontuais são verdadeiras. Dada a forma de suas respectivas distribuições amostrais, você acha razoável efetuar inferência e informar a margem de erros, como o relatório faz?

Sua Vez

A questão sobre o ateísmo foi também feita pelo WIN-Gallup International numa pesquisa de opinião parecida realizada em 2005.[†] A Tabela 4 na página 13 do relatório resume os resultados da pesquisa de 2005 a 2012 em 29 países.

1. Responda às duas perguntas seguintes utilizando a função `inference`. Como sempre, descreva as hipóteses para qualquer teste que você realizar e esboce sobre as condições para inferência.

   (a) Há evidência convincente de que a Espanha teve uma mudança em seu índice de ateísmo entre 2005 e 2012?
   Dica: Crie um novo conjunto de dados para os respondentes da Espanha. Depois, utilize suas respostas como a primeira entrada na função `inference`, e utilize a variável `year` (ano) para definir os grupos.

   (b) Há evidência convincente de que os Estados Unidos tiveram uma mudança em seu índice de ateísmo entre 2005 e 2012?

2. Se de fato não houve nenhuma mudança no índice de ateísmo nos países listados na Tabela 4, em quantos países você esperar detectar uma mudança (com um nível de significância de 0,05) simplesmente por acaso?
   Dica: Procure no índice do livro sobre erros do Tipo 1.

3. Suponha que você foi contratado pelo governo local para estimar a proporção de residentes que participam de cultos religiosos semanalmente. De acordo com diretrizes, a estimativa deve ter uma margem de erro inferior a 1% com nível de confiança de 95%. Você não tem nenhuma noção de que valor supor para p. Quanto pessoas você teria que amostrar para garantir que você está dentro das diretrizes?
   Dica: Retome seu gráfico da relação entre p e a margem de erro. Não use o conjunto de dados para responder a essa questão.

4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

[†] Assumimos aqui que o tamanho das amostras permaneceram iguais.

Laboratório 7: Introdução à Regressão Linear

Preparação

O filme *O Homem que Mudou o Jogo* (*Moneyball*) aborda a "busca pelo segredo do sucesso no beisebol". O filme conta a história de um time de baixo orçamento, o *Oakland Athletics*, que acreditava que estatísticas pouco utilizadas, tal como a habilidade de um jogador chegar a uma base, prediziam melhor a habilidade de marcar pontos do que estatísticas mais comuns, como *home runs*, RBIs (*runs batted in*, pontos feitos quando um jogador estava rebatendo), e média de rebatidas. Contratar jogadores que se destacavam nessas estatísticas pouco utilizadas se mostrou muito mais econômico para o time.

Neste laboratório exploraremos os dados de todos os 30 times da Liga Principal de Beisebol dos Estados Unidos e examinaremos a relação linear entre pontos marcados numa temporada e várias outras estatísticas dos jogadores. Nosso objetivo será resumir essas relações de maneira visual e numérica para identificar qual variável, se houver alguma, melhor nos ajuda a predizer os pontos marcados por um time numa temporada.

Os Dados

Vamos carregar os dados da temporada de 2011.

```
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")

load("mlb11.RData")
```

Além dos pontos marcados, este conjunto de dados contém sete variáveis tradicionalmente utilizadas: vez ao taco (*at-bats*), rebatidas (*hits*), *home runs*, média de rebatidas (*batting average*), eliminações (*strikeouts*), roubos de bases (*stolen bases*), e vitórias[†]. Também foram incluídas três novas variáveis: percentual de alcance de base (*on-base percentage*), percentual de potência (*slugging percentage*), e alcance de base mais potência (*on-base plus slugging*). Para a primeira parte da análise consideraremos as sete variáveis tradicionais. Ao final do laboratório, você trabalhará com as novas variáveis por conta própria.

> **Exercício 1** Que tipo de gráfico você utilizaria para mostrar a relação entre `runs` (pontos) e alguma outra variável numérica? Crie um gráfico dessa relação utilizando a variável `at_bats` como preditora. A relação parece ser linear? Se você soubesse o valor de `at_bats` (vez ao taco) de um time, você se sentiria confiante para utilizar um modelo linear para predizer o número de pontos (*runs*)?

Se a relação parece ser linear, podemos quantificar a força da relação utilizando o coeficiente de correlação.

```
cor(mlb11$runs, mlb11$at_bats)
```

 Este laboratório foi adaptado para a OpenIntro por Andrew Bray e Mine Çetinkaya-Rundel de um laboratório escrito por Mark Hansen do departamento de Estatística da UCLA. Tradução para o português por Erikson Kaszubowski.

[†]Apesar de não ser necessário para acompanhar o laboratório, se você quiser se familiarizar com as regras do beisebol e com as estatísticas mais utilizadas, visite *http://en.wikipedia.org/wiki/Baseball_rules* e *http://en.wikipedia.org/wiki/Baseball_statistics*. Como os termos usuais nem sempre tem um tradução exata para o português, manterei entre parênteses o termo original em inglês (N. do T.).

Soma dos Quadrados dos Resíduos

Recorde como descrevemos a distribuição de uma única variável. Lembre-se que discutimos características como tendência central, dispersão e forma. Também é útil poder descrever a relação entre duas variáveis numéricas, como fizemos acima com as variáveis `runs` (pontos) e `at_bats` (vez ao taco).

> **Exercício 2** Examinando os gráficos do exercício anterior, descreva a relação entre essas duas variáveis. Certifique-se de discutir a forma, a direção e a força da relação, bem como quaisquer características incomuns.

Assim como utilizamos a média e o desvio padrão para resumir características importantes de uma única variável, podemos resumir a relação entre essas duas variáveis por meio de uma linha que melhor descreve sua associação. Utilize a seguinte função interativa para selecionar a linha que você acha que cruza a nuvem de pontos da melhor maneira.

```
plot_ss(x = mlb11$at_bats, y = mlb11$runs)
```

Depois de executar esse comando, você será solicitado a clicar em dois ponto no gráfico para definir uma linha. Depois que você fizer isso, a linha que você especificou será mostrada na cor preta e os resíduos na cor azul. Perceba que há 30 resíduos, um para cada uma das 30 observações. Lembre-se que os resíduos são a diferença entre os valores observados e o valor predito pela linha:

$$e_i = y_i - \hat{y}_i$$

A maneira mais comum de se fazer uma regressão linear é selecionar a linha que minimiza a soma dos quadrados dos resíduos. Para visualizar o quadrado dos resíduos, você pode rodar novamente o comando de geração do gráfico e adicionar o argumento `showSquares = TRUE`.

```
plot_ss(x = mlb11$at_bats, y = mlb11$runs, showSquares = TRUE)
```

Perceba que o resultado da função `plot_ss` fornece a inclinação (coeficiente angular) e o intercepto da sua linha, bem como a soma dos quadrados.

> **Exercício 3** Utilizando a função `plot_ss`, escolha uma linha que consiga minimizar a soma dos quadrados. Rode a função várias vezes. Qual foi a menor soma dos quadrados que você obteve? Compare-a com os resultados obtidos por outros alunos.

O Modelo Linear

É bastante cansativo tentar obter a linha dos mínimos quadrados, ou seja, a linha que minimiza a soma dos quadrados dos resíduos, por meio de tentativa e erro. Ao invés disso, podemos utilizar a função `lm` no R para ajustar o modelo linear (também conhecido como linha de regressão).

```
m1 <- lm(runs ~ at_bats, data = mlb11)
```

O primeiro argumento da função `lm` é a fórmula descrita como `y~x`. Aqui ela pode ser entendida como "obtenha o modelo linear de `runs` (pontos) em função de `at_bats` (vez ao taco)." O segundo argumento especifica que o R deve buscar no banco de dados `mlb11` as variáveis `runs` e `at_bats`.

O resultado da função `lm` é um objeto que contém todas as informações que precisamos sobre o modelo linear que acabamos de ajustar. Podemos acessar essa informação utilizando a função `summary`.

```
summary(m1)
```

Vamos analisar o resultado passo a passo. Primeiramente, a fórmula utilizada para descrever o modelo é apresentada no começo. Depois da fórmula você verá o resumo de cinco números dos resíduos. A tabela "Coefficients" (coeficientes) apresentada em seguida é central; sua primeira coluna apresenta o intercepto de y do modelo linear e o coeficiente da variável `at_bats`. Com essa tabela, podemos descrever a linha de regressão de mínimos quadrados para o modelo linear:

$$\hat{y} = -2789.2429 + 0.6305 * atbats$$

Uma última informação que abordaremos do resultado da função `summary` é o R-quadrado Múltiplo, ou de maneira abreviada, R^2. O valor do R^2 representa a proporção de variabilidade na variável desfecho que é explicada pela variável explicatória. Neste modelo, 37,7% da variabilidade dos pontos (*runs*) é explicada pela vez ao taco (*at-bats*).

> **Exercício 4** Ajuste um novo modelo que utilize a variável `homeruns` para predizer `runs` (pontos). Utilizando as estimativas dos resultados do R, escreva a equação da linha de regressão. O que a inclinação (coeficiente angular) nos diz sobre a relação entre o sucesso de um time e seus *home runs*?

Predição e Erro de Predição

Vamos criar uma gráfico de dispersão com a linha dos mínimos quadrados disposta junto com os pontos.

```
plot(mlb11$runs ~ mlb11$at_bats)

abline(m1)
```

A função `abline` traça uma linha baseada em sua inclinação e intercepto. Aqui, utilizamos um atalho fornecendo o modelo `m1`, que contém as estimativas do dois parâmetros. Essa linha pode ser utilizada para predizer y a partir de qualquer valor de x. Quando são feitas predições para valores de x que estão além do intervalo dos dados observados, denominamos essas predições de *extrapolações* e geralmente não é algo recomendável. Contudo, predições feitas dentro do intervalo dos dados são mais confiáveis. Elas também são utilizadas para calcular os resíduos.

> **Exercício 5** Se o gerente de um time visse a linha de regressão dos mínimos quadrados e não os dados reais, quantos ponto (*runs*) ele prediria para um time com 5.578 vezes ao taco (*at-bats*)? Esse valor superestima ou subestima o valor real, e por quanto? Em outras palavras, qual é o resíduo para essa predição?

Diagnósticos do Modelo

Para avaliar se um modelo linear é confiável, precisamos verificar (1) a linearidade, (2) resíduos normalmente distribuídos, e (3) variância constante.

(1) Linearidade: Você já verificou se a relação entre pontos (*runs*) e vezes ao taco (*at-bats*) é linear utilizando o gráfico de dispersão. Deveríamos também verificar essa condição utilizando um gráfico de resíduos

em função da variável vez ao taco (*at-bats*). Lembre-se que todo código após um # é um comentário para auxiliar a compreender o código e é ignorado pelo R.

```
plot(m1$residuals ~ mlb11$at_bats)

abline(h = 0, lty = 3)  # adiciona uma linha pontilhada horizontal em y = 0
```

Exercício 6 Há algum padrão aparente do gráfico de resíduos? O que isso indica sobre a linearidade da relação entre pontos (*runs*) e vezes ao taco (*at-bats*)?

(2) Resíduos normalmente distribuídos: Para verificar essa condição, podemos conferir o histograma dos resíduos:

```
hist(m1$residuals)
```

ou um gráfico de probabilidade normal dos resíduos.

```
qqnorm(m1$residuals)

qqline(m1$residuals)  # adiciona uma linha diagonal ao gráfico de probabilidade normal
```

Exercício 7 Com base no histograma e no gráfico de probabilidade normal, a condição de distribuição normal dos resíduos parece ser atendida?

(3) Variância constante:

Exercício 8 Com base no gráfico criado em (1), a condição de variância constante parece ser atendida?

Sua Vez

1. Escolha outra variável tradicional contida no banco de dados `mlb11` que você acha que poderia ser um bom preditor da variável `runs` (pontos). Crie um gráfico de dispersão das duas variáveis e ajuste um modelo linear. Visualmente, parece haver uma relação linear?
2. Compare essa relação com a relação entre `runs` (pontos) e `at_bats` (vez ao taco). Utilize os valores R^2 do sumário dos dois modelos para compará-los. A variável que vocês escolheu parece predizer `runs` (pontos) melhor do que `at_bats` (vez ao taco)? Como você justificaria sua resposta?
3. Agora que você pode resumir a relação linear entre duas variáveis, investigue a relação entre `runs` (pontos) e cada uma das outras cinco variáveis tradicionalmente utilizadas no beisebol. Qual variável prediz melhor o valor de `runs`? Justifique sua conclusão utilizando métodos gráficos e numéricos que já discutimos (para ser conciso, inclua apenas os resultados da melhor variável, não de todas as cinco).

4. Agora examine as três variáveis mais recentes. Essas são as estatísticas utilizadas pelo autor do filme *O Homem que Mudou o Jogo* para predizer o sucesso de um time. De modo geral, elas são mais ou menos eficazes para predizer os pontos do que as variáveis mais tradicionais? Explique utilizando evidências gráficas e numéricas. De todas as dez variáveis que nós analisamos, qual parece ser o melhor preditor da variável `runs` (pontos)? Utilizando as informações limitadas (ou não tão limitadas) que você conhece sobre estas estatísticas do beisebol, seu resultado faz sentido?

5. Verifique os diagnósticos do modelo para o modelo de regressão com a variável que você escolheu como o melhor preditor dos pontos (*runs*).

6. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

Laboratório 8: Regressão Linear Múltipla

Dando Nota ao Professor

Vários cursos universitários dão aos alunos a oportunidade de avaliar o curso e o professor de maneira anônima ao final do semestre. Contudo, o uso das avaliações dos alunos como um indicador da qualidade do curso e a eficácia do ensino é frequentemente criticado porque essas medidas podem refletir a influência de características não relacionadas à docência, tal como a aparência física do professor. O artigo intitulado "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" (Hamermesh & Parker, 2005)* descreve como professores que são vistos como tendo melhor aparência recebem avaliações mais altas.†

Neste laboratório analisaremos os dados deste estudo para aprender o que influencia uma avaliação positiva de um professor.

Os Dados

Os dados foram coletados a partir das avaliações discentes de final de semestre de uma grande amostra de professores da Universidade do Texas em Austin. Além disso, seis estudantes avaliaram a aparência física dos professores.‡ O resultado é um banco de dados no qual cada linha contém diferentes disciplinas e cada coluna representa as variáveis sobre as disciplinas e os professores.

```
download.file("http://www.openintro.org/stat/data/evals.RData", destfile = "evals.RData")

load("evals.RData")
```

Explorando os Dados

Exercício 1 Esse estudo é observacional ou experimental? O pergunta de pesquisa original proposta no artigo é se a beleza influencia diretamente as avaliações das disciplinas. Levando em consideração o desenho da pesquisa, é possível responder a essa pergunta tal como ela está formulada? Se não, reformule a pergunta.

Exercício 2 Descreva a distribuição da variável `score`. A distribuição é assimétrica? O que sua forma permite dizer sobre a maneira como os alunos avaliam as disciplinas? A forma corresponde ao que você esperava ver? Por quê, ou por que não?

Exercício 3 Com exceção da variável `score`, escolha duas outras variáveis e descreva sua relação utilizando as técnicas apropriadas (gráfico de dispersão, gráfico de caixas lado-a-lado, ou gráfico de mosaico).

 Este laboratório foi adaptado para a OpenIntro por Andrew Bray e Mine Çetinkaya-Rundel de um laboratório escrito por Mark Hansen do departamento de Estatística da UCLA. Tradução para o português por Erikson Kaszubowski.

*"Beleza na sala de aula: a pulcritude do professor e produtividade pedagógica putativa"

†Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity, *Economics of Education Review*, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. (http://www.sciencedirect.com/science/article/pii/S0272775704001165).

‡Esta é uma versão levemente modificada do conjunto de dados original que foi publicado como parte dos dados de reprodução para o livro *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill, 2007).

score	pontuação média da avaliação do docente: (1) muito insatisfatório - (5) excelente.
rank	nível do professor: horista (teaching), assistente (tenure track), titular (tenured).*
ethnicity	etnia do professor: não-minoria, minoria.
gender	sexo do professor: feminino, masculino.
language	língua da universidade frequentada pelo professor: inglês ou não-inglês.
age	idade do professor.
cls_perc_eval	percentual de alunos na turma que completaram a avaliação.
cls_did_eval	número de alunos na turma que completaram a avaliação.
cls_students	número total de alunos na turma.
cls_level	nível da disciplina: introdutória, avançada.
cls_profs	número de professores ministrando módulos na disciplina dentro da amostra: único, múltiplos.
cls_credits	número de créditos da disciplina: um crédito, múltiplos créditos.
bty_f1lower	avaliação da beleza do professor por aluna de nível inicial: (1) mais baixo - (10) mais alto.
bty_f1upper	avaliação da beleza do professor por aluna de nível avançado: (1) mais baixo - (10) mais alto.
bty_f2upper	avaliação da beleza do professor por segunda aluna de nível avançado: (1) mais baixo - (10) mais alto.
bty_m1lower	avaliação da beleza do professor por aluno de nível inicial: (1) mais baixo - (10) mais alto.
bty_m1upper	avaliação da beleza do professor por aluno de nível avançado: (1) mais baixo - (10) mais alto.
bty_m2upper	avaliação da beleza do professor por segundo aluno de nível avançado: (1) mais baixo - (10) mais alto.
bty_avg	média da avaliação da beleza do professor.
pic_outfit	roupa do professor na foto avaliada: informal, formal.
pic_color	cor da foto avaliada: colorida, preto e branco.

Regressão Linear Simples

O fenômeno proposto pelo estudo é que professores com melhor aparência são avaliados de maneira mais favorável. Vamos criar um gráfico de dispersão para verificar se isso é verdade:

```
plot(evals$score ~ evals$bty_avg)
```

Antes de tirar conclusões sobre a tendência, compare o número de observações no banco de dados com o número de pontos no gráfico de dispersão. Há algo de errado?

Exercício 4 Refaça o gráfico de dispersão, mas agora utilize a função `jitter()` no eixo y ou x. (Utilize o comando `?jitter` para aprender mais a respeito.) O que estava errado no gráfico de dispersão inicial?

Exercício 5 Vamos verificar se a tendência aparente no gráfico é algo além de variação natural. Ajuste um modelo linear denominado `m_bty` para predizer a avaliação média de um professor a partir da média da avaliação da beleza e adicione a linha ao gráfico utilizando o comando `abline(m_bty)`. Escreva a equação do modelo linear e interprete a inclinação da reta. A média da avaliação da beleza é um preditor estatisticamente significante? Essa variável parecer ser um preditor com significância prática?

Exercício 6 Utilize gráficos de resíduos para avaliar se as condições para uma regressão utilizando mínimos quadrados são plausíveis. Utilize gráficos e comente cada uma deles (retome o Laboratório 7 para relembrar como criá-los).

Regressão Linear Múltipla

O conjunto de dados contém diversas variáveis sobre a avaliação de beleza do professor: avaliações individuais de cada um dos seis estudantes que foram convidados a avaliar a aparência física dos professores e a média dessas seis avaliações. Vamos dar uma olhada na relação entre uma dessas avaliações e a média da avaliação da beleza.

```
plot(evals$bty_avg ~ evals$bty_f1lower)
cor(evals$bty_avg, evals$bty_f1lower)
```

Como esperado, a relação é bem forte – afinal, a média das avaliações é calculada utilizando a avaliações individuais. Podemos dar uma olhada nas relações entre todas as variáveis relativas à beleza (colunas 13 a 19) utilizando o seguinte comando:

```
plot(evals[,13:19])
```

Essas variáveis são colineares (correlacionadas), e adicionar mais do que uma delas ao modelo não agregaria muito valor. Neste caso, com esses preditores com altos índices de correlação, é melhor utilizar a média das avaliações da beleza como o único representante dessas variáveis.

Para verificar se a beleza ainda é um preditor significante da avaliação docente depois que consideramos o sexo do professor, podemos adicionar um termo para o sexo no modelo.

```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

> **Exercício 7** Valores p e estimativas dos parâmetros só são confiáveis se as condições para a regressão são plausíveis. Verifique se as condições para esse modelo são plausíveis utilizando gráficos de diagnóstico.

> **Exercício 8** A variável `bty_avg` continua sendo um preditor significante de `score`? A adição da variável `gender` ao modelo alterou a estimativa do parâmetro de `bty_avg`?

Perceba que a estimativa para `gender` é agora denominada de `gendermale`. Você verá essa mudança de nome sempre que adicionar uma variável categorial ao modelo. O motivo é que o R recodifica `gender`, alterando seus valores iniciais `female` (feminino) e `male` (masculino) para uma variável indicativa denominada `gendermale` que tem o valor 0 para mulheres e o valor 1 para homens (tais variáveis são frequentemente chamada de variável "dummy" (falsa ou postiça)).

O resultado, para mulheres, é que o parâmetro estimado é multiplicado por zero, deixando a forma do intercepto e da inclinação similar à regressão simples.

$$\begin{aligned}\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg\end{aligned}$$

Podemos traçar essa linha e a linha correspondente aos homens com a seguinte função personalizada:

```
multiLines(m_bty_gen)
```

> **Exercício 9** Qual é a equação da linha correspondente aos homens? (*Dica:* Para os homens, a estimativa do parâmetro é multiplicada por 1.) Para dois professores que receberam a mesma avaliação de beleza, qual gênero tende a ter as avaliações mais altas?

A decisão de chamar a variável indicativa de `gendermale` ao invés de `genderfemale` não tem nenhum significado profundo. O R simplesmente codifica a categoria que vem em primeiro lugar na ordem alfabética

como um 0.[§]

> **Exercício 10** Crie um novo modelo denominado `m_bty_rank` removendo a variável `gender` e adicionando a variável `rank`. Como o R maneja variáveis categoriais que tem mais de dois níveis? Perceba que a variável `rank` tem três níveis: horista (teaching), assistente (tenure track) e titular (tenured).

A interpretação dos coeficientes na regressão múltipla é um pouco diferente da regressão simples. A estimativa do coeficiente da variável `bty_avg` reflete quanto mais um grupo de professores deve receber na avaliação da disciplina se sua avaliação de beleza é um ponto maior *mantendo todas as outras variáveis constantes*. Neste caso, isso significa considerar somente professores do mesmo nível com avaliações de `bty_avg` que estão separadas por um ponto.

A Busca pelo Melhor Modelo

Vamos começar com um modelo completo que prediz a avaliação docente com base no nível, etnia, sexo, língua da universidade onde obteve seu diploma, idade, proporção de alunos que completaram as avaliações, tamanho da turma, nível da disciplina, número de professores, número de créditos, média da avaliação da beleza, roupa e cor da foto avaliada.

> **Exercício 11** Qual variável você acha que teria o maior valor p neste modelo? Por quê? *Dica:* Pense em qual variável você esperaria não estar associada à avaliação docente.

Vamos rodar o modelo...

```
m_full <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_profs + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full)
```

> **Exercício 12** Verifique suas suspeitas do exercício anterior. Inclua os resultados do modelo em sua resposta.

> **Exercício 13** Interprete o coeficiente associado à variável etnia.

> **Exercício 14** Retire a variável com o maior valor p e reajuste o modelo. Os coeficientes e suas significâncias para as outras variáveis explicativas se alteraram? (Uma das coisas que torna a regressão múltipla interessante é que a estimativa dos coeficientes dependem das outras variáveis que são incluídas no modelo.) Se não, o que isso implica para questão de se a variável retirada era ou não colinear com outras variáveis explicativas?

> **Exercício 15** Utilizando seleção inversa e o valor p como critério de seleção, determine qual é o melhor modelo. Você não precisa mostrar todos os passos na sua resposta, apenas o resultado do modelo final. Também escreva a equação do modelo linear para predizer a avaliação docente com base no modelo final que você estabeleceu.

> **Exercício 16** Verifique se as condições para esse modelo são plausíveis utilizando gráficos de diagnóstico.

[§]Você pode mudar o nível de referência de uma variável categorial, que é o nível codificado como um 0, utilizando a função `relevel`. Utilize o comando `?relevel` para aprender mais a respeito.

Exercício 17 O artigo original descreve como os dados foram obtidos a partir de amostras de professores da Universidade do Texas em Austin e incluindo todas as disciplinas que eles ministraram. Considerando que cada linha representa uma disciplina, essa nova informação poderia ter algum impacto em alguma das condições para a regressão linear?

Exercício 18 Com base no seu modelo final, descreva as características de um professor e de uma disciplina da Universidade do Texas em Austin que estariam associadas com uma avaliação alta.

Exercício 19 Você se sentiria confiante em generalizar suas conclusões para todos os professores, de modo geral (e em qualquer universidade)? Por quê ou por que não?