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.

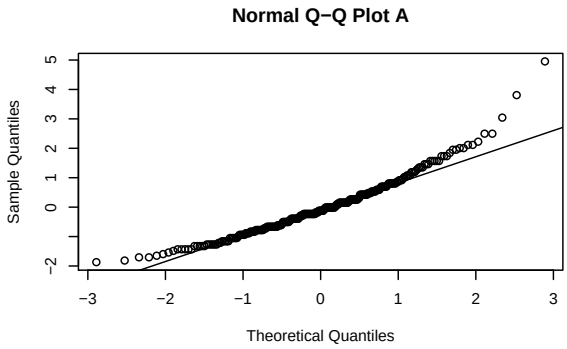
> **Exercício 6** Elabore duas questões de probabilidade que você gostaria de responder; uma com relação à altura de mulheres e outra com relação ao peso de mulheres. Calcule essas probabilidades usando tanto o método teórico da distribuição normal quanto a distribuição empírica (quatro probabilidade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

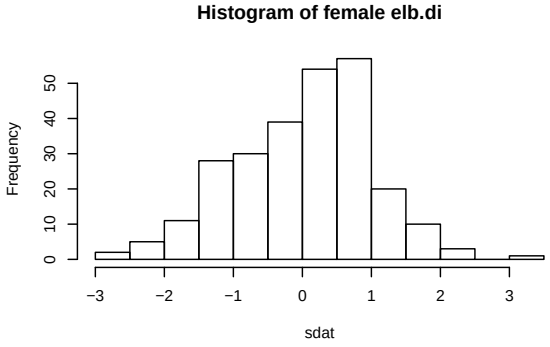
Sua Vez

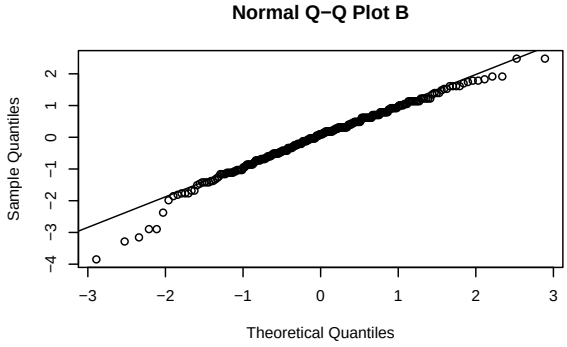
1. Agora vamos analisar outras variáveis no conjunto de dados das dimensões corporais. Utilizando as figuras na próxima página, combine os histogramas com seus gráficos de probabilidade normal. Todas as variáveis foram estandardizadas (primeiro subtraindo a média, e em seguida dividindo pelo desvio padrão), 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.

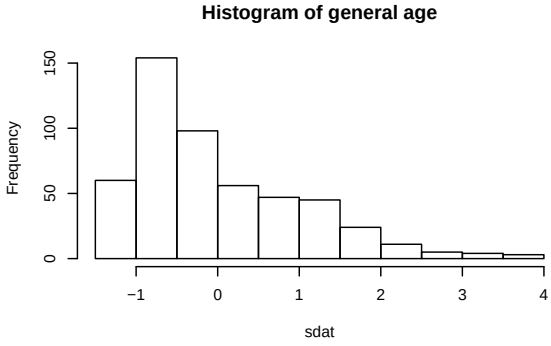
4. Quais conceitos do livro são abordados neste laboratório? Quais conceitos, se houver algum, que não são abordados no livro? Você viu esses conceito em algum outro lugar, p.e., aulas, seções de discussão, laboratórios anteriores, ou tarefas de casa? Seja específico em sua resposta.

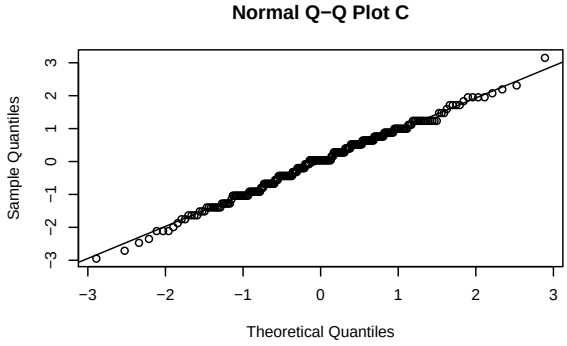

Histogram of female bii.di
Frequency
sdat

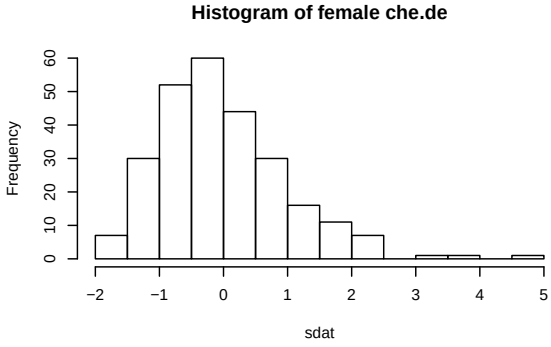

Normal Q−Q Plot A
Sample Quantiles
Theoretical Quantiles

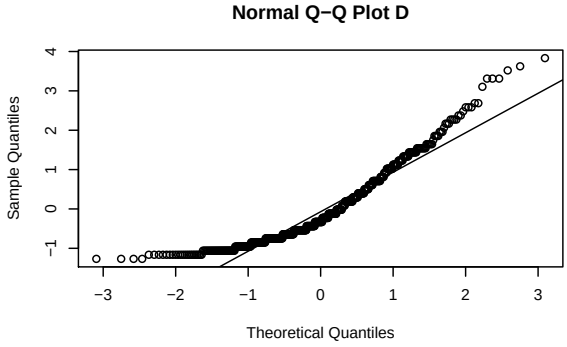

Histogram of female elb.di
Frequency
sdat


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