# Capítulo 1

# Os pássaros de Darwin e o método hipotético-dedutivo.

## *Testes estatísticos e distribuições probabilísticas*

# Introdução

Charles Darwin observou que os pássaros fringilídeos nas ilhas de Galápagos apresentavam variedades de formato e tamanho dos bicos. Sua intuição sobre a origem das variedades a partir de um ancestral comum foi um dos argumentos mais contundentes do “*On the Origin of Species*” (1859). Essa história é o ponto de partida para este capítulo.

Estudamos a relação natural entre ciências empíricas e duas distribuições probabilísticas: a distribuição normal e a distribuição t, relacionadas entre si. A adoção da distribuição normal em trabalhos científicos é popular, porém os motivos são pouco entendidos. O Teorema do Limite Central é fundamental nesse contexto.

Usamos as distribuições citadas para estudar as medidas dos bicos dos tentilhões em pequenas amostras de cada ilha e *fazer inferências* sobre as populações. O racional de testes de hipótese é introduzido.

## Pássaros em Galápagos

Em sua viagem pelo mundo a bordo do Beagle, Charles Darwin descreveu um grupo de pássaros que habita as Ilhas Galápagos, arquipélago localizado a aproximadamente 900 km da costa do Equador (América do Sul). A variedade em tamanhos dos bicos chamou atenção: “*It is very remarkable that a nearly perfect gradation of structure in this one group can be traced in the form of the beak, from one exceeding in* ***dimensions*** *that of the* ***largest*** *gros-beak, to another* ***differing******but little*** *from that of a warbler".[[1]](#footnote-0) The Voyage of the Beagle (1839).*

É interessante notar que a linguagem usada para denotar diferenças é eminentemente quantitativa (dimensions, largest, differing). Darwin não conduziu estudos quantitativos por razões práticas. Neste capítulo, simularemos o mesmo cenário empregando métodos estatísticos para comparar os pássaros.

Antes da publicação de A origem das Espécies, o caso dos fringilíddeos (nome destas aves) já continha um embrião do processo de seleção natural. Na segunda edição, em 1845, ele especula sobre um grupo ancestral comum moldado por fins específicos:

*“Seeing this gradation and diversity of structure in one small, intimately related group of birds, one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends.”[[2]](#footnote-1)*

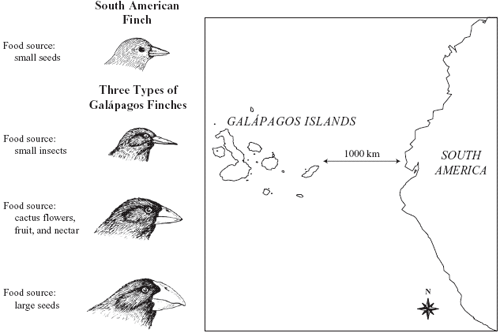


Figura 1. Fringilídeos em Galápagos.

Darwin observou que a variedade dos bicos era adaptada à dieta de cada grupo: frutas, nozes, insetos. Os de bico pontudo conseguem comer frutas e arilo da sementos do cacto, enquanto os de bico curto estraçalham a base do cacto e comem sua polpa.

A inspeção visual de um naturalista treinado foi capaz detectar essas nuances. Sob sua percepção, havia um total de 3 espécies em 4 ilhas: 1 na Ilha Charles, 1 na Ilha Albemarle e 1 nas ilhas James e Chatham. Inicialmente, notou que os pássaros eram semelhantes àqueles vistos no Chile. Darwin coletou 26 pássaros e os levou de volta para que um ornitólogo os estudasse com mais detalhe. John Gould sugeriu que os 26 pássaros representavam 12 espécies completamente novas, número que posteriormente passou para 25. Hoje, os taxonomistas sugerem um número de 15 espécies para os fringilíddeos de Darwin.

Pensaremos como biólogos interessados em estudar quantitativamente o tamanho dos bicos. Usaremos estatística e probabilidades para testar hipóteses e fazer conclusões mais acuradas sobre as medidas, explorando diferenças entre os grupos de pássaros de Galápagos.

## A distribuição normal e um curioso teorema

Em trabalhos empíricos, é comum a suposição de que medidas de uma variável aleatória vêm de uma população com distribuição normal. A seguir, vamos estudar o comportamento dessa função probabilística.

Abraham de Moivre (26 May 1667 – 27 November 1754) , sem financiamento exclusivo para estudos e pesquisa, prestava serviços. Entre eles, cálculos de probabilidades em jogos de azar para clientes. Em 1733, de Moivre percebeu que as probabilidades de uma distribuição binomial, como o lançamento de moedas (p=0.5), aproximam-se de uma curva suave (contínua) à medida em que o *n* aumenta.

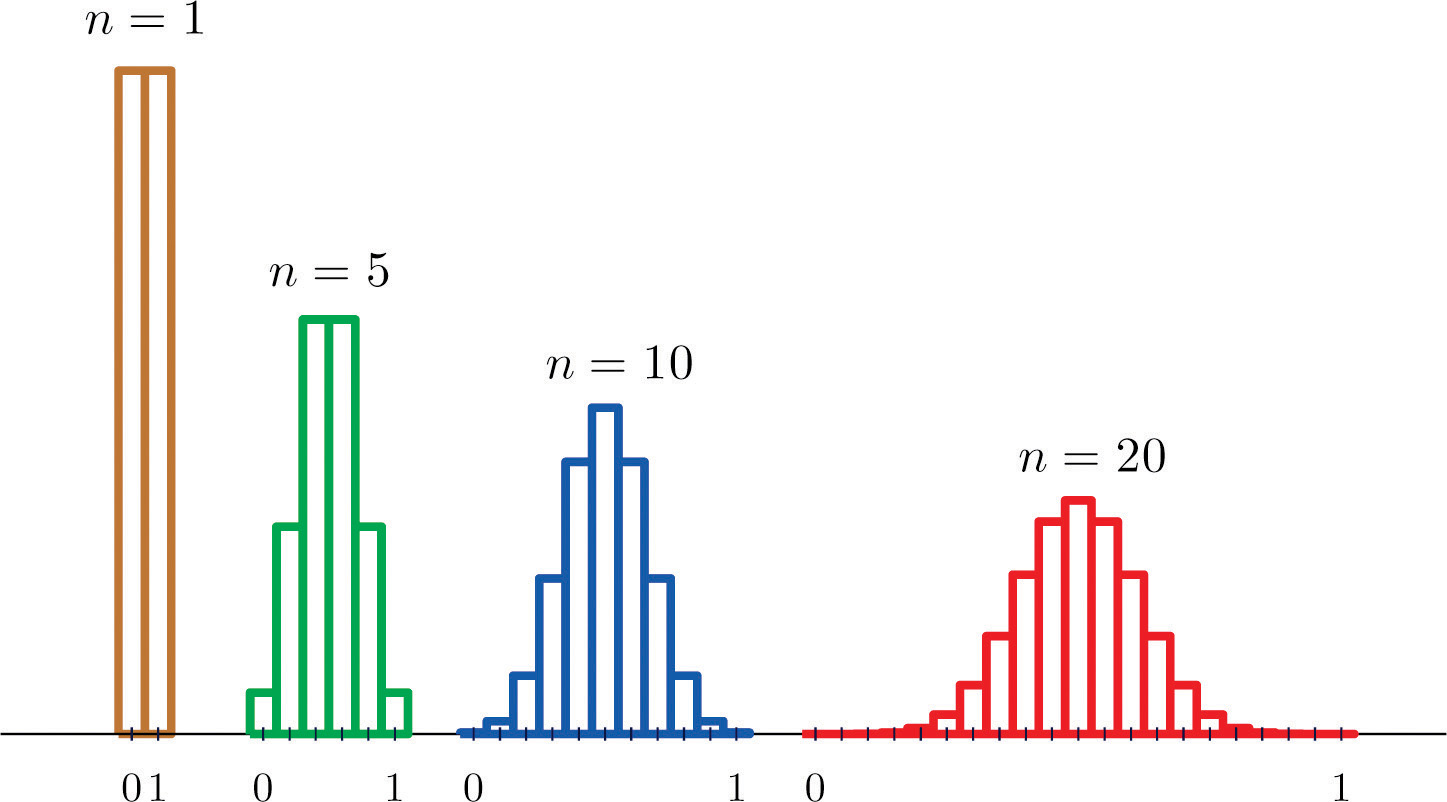


Figura 2. Distribuição binomial com p = 0.5 (e.g: lançamento de uma moeda). Para n >1, valores extremos indicam resultados com apenas caras (cauda à esquerda, 0000…) ou coroas (cauda à direita, 1111…).

Para n = 1, temos uma distribuição uniforme P(1) = P(0) = 0.5. Para um número maior de lançamentos, os resultados mais frequentes são números parecidos de caras (0s) e coroas (1s).

Para n = 10, é muito mais provável obter um número de *caras* próximo a 5 (centro das curvas) que um resultado com 9 ou 10 lançamentos iguais. É possível demonstrar que aumentar o valor de n faz que com que a distribuição se aproxime da seguinte curva contínua:

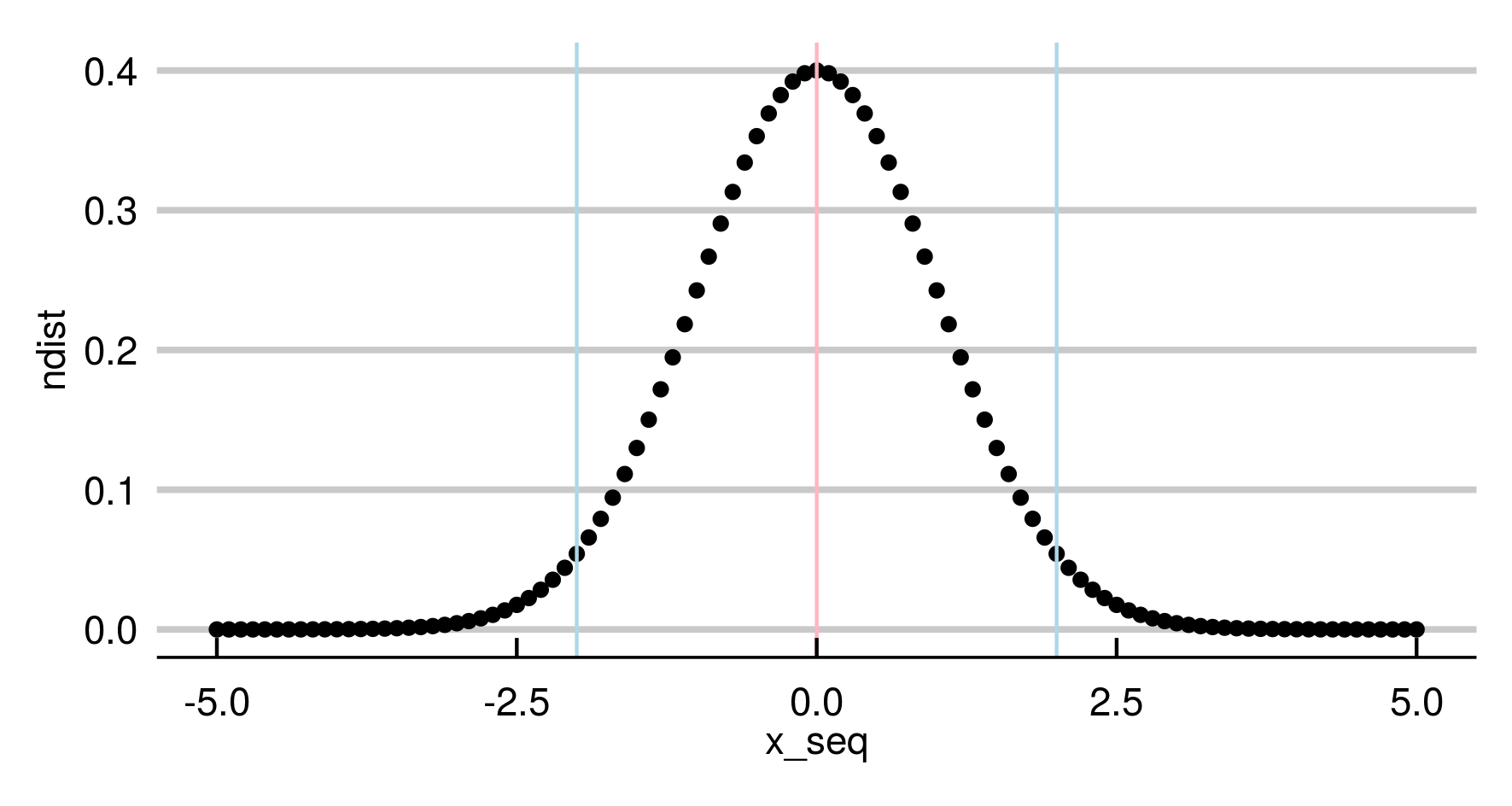
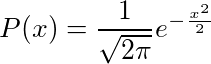
**

Figura 3: Distribuição normal (gaussiana), cujo formato lembra o de um sino

Algumas características nesta curva são desejáveis e/ou plausíveis para o estudo de fenômenos naturais: simetria; valores mais prováveis concentrados ao redor de um pico, valores extremos improváveis; distribuição governada por poucos parâmetros (média e variância).

De Moivre intuiu que a distribuição de binomiais com muitos lançamentos aproximava o de uma função suave. Mas qual curva? Parecia o de um número elevado aos valores de uma quadrática negativa: P(x) = ?**-x²**. Em verdade, ele deduziu a solução para o problema das moedas (p = ½). A seguinte expressão geral descreve a probabilidade (**P(x**)) correspondente à curva que procuramos.

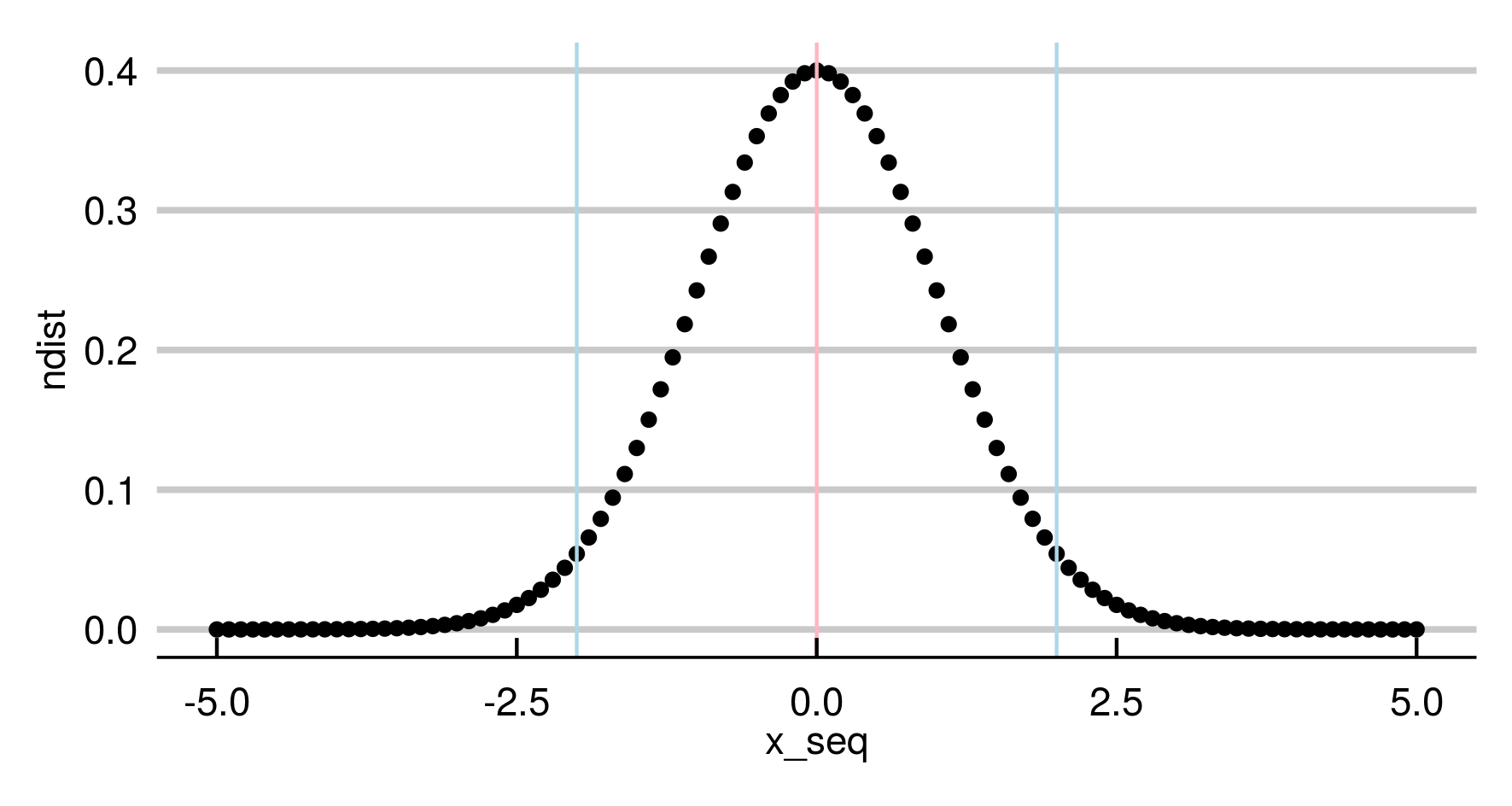


O número misterioso em questão é o número de *Euler (e ~ 2.72…)*, que será melhor explicado no capítulo 3*.* Há um fator, 1/√2π (aproximadamente 0.4), multiplicando o resultado. *Em R,* podemos definir a função

*>mgauss <- function(x) 0.4\*exp((-1)\*(x^2)/2)*

Em seguida, obter valores no intervalo [-5,5] e plotá-los:

*>library(ggplot2)  
>x\_seq <- seq(-5,5,by = 0.1)  
> ndist <- purrr::map(.f=mgauss,.x=x\_seq) %>% unlist  
> ggplot(data.frame(ndist,x\_seq),aes(x=x\_seq,y=ndist))+  
 geom\_point()+  
 geom\_vline(xintercept = 2,color="light blue")+  
 geom\_vline(xintercept = -2,color="light blue")+  
 geom\_vline(xintercept = 0,color="light pink")+  
 theme\_economist\_white(gray\_bg = F)*

**

Observamos como a distribuição gaussiana se dá a partir da equação.

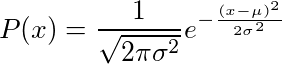
É evidente que -***x2***  sempre retorna valores negativos. Nossa função gera valores entre 0 e 1 exponenciando *e (e ~2.718...)* a um fator negativo quadrático (*y = 0.4*\**e-x²/2*).

Examinando o comportamento da equação, notamos que valores próximos à média **(x ≃ μ = 0)** fazem com que o expoente de ***e***, dado por ***-x2/2*** , se aproxime de 0, maximizando nossa função (ƒ**(0) = 0.4 \*** ***e-x²/2 = 0.4 \* e0 = 0.4***). O valor obtido (0.4) corresponde ao topo da curva no gráfico acima (linha rosa). Observamos a curva se aproximar do máximo simetricamente para valores próximos de 0.

Isso reflete diretamente o fato de que valores próximos à média serão mais prováveis e valores extremos menos prováveis. Para comparação: ƒ**(2) = 0.4** \* ***e-2²/2 =* 0.4** \* ***e-2* = 0.4 \* 0.135 ~ 0.05** (linha azul)**.** A probabilidade de se obter o valor médio (x = 0, p ~ 0.4) é oito vezes maior que a probabilidade de obter o valor 2 (x = 2; p = 0.05).

O termo quadrático torna a distribuição simétrica para valores opostos em relação à média. **P(x) = P(-x)**. Como calculamos P(2) antes, sabemos que: **P(-*2)* = P(*2) = 0.05***para *μ =* 0*.* É igualmente provável encontrar valores duas unidades maiores ou duas unidades menores que a média. Esses pontos estão marcados por uma linhas azuis na figura**.**

Podemos trabalhar com curvas normais com centros (média **μ**) deslocados para a esquerda (μ < 0) ou para a direita (μ > 0), subtraindo o termo de x em nosso expoente. Além disso, diferentes variâncias (**σ²**) refletem a frequência de valores longe da média e o quão distante dela eles são. Visualmente, determina o tamanho da base do sino na ilustração (Figura 3). Usamos a notação **N ~ (μ, σ²)** para descrever uma distribuição gaussiana arbitrária:



Poderíamos encontrar características desejáveis, como a simetria citada acima, em outras distribuições. Então, por que usamos uma equação mais complexa? Distribuições binomiais grandes e lançamentos de moedas são tão importantes?

## O Teorema do Limite Central

A razão é o Teorema do Limite Central.

S*e somarmos muitas distribuições de uma mesma família, a distribuição resultante se aproxima de uma normal*.

Exemplos ajudam a ganhar intuição. Ao lançar um dado justo de 6 faces, temos probabilidade de ⅙ em cada resultado.



Uma distribuição discreta uniforme, em que P(1) = P(2) = P(3) = P(4) = P(5) = P(6) e definida para números naturais entre 1 e 6: ***X* ~ Udiscr(1, 6)**.

A média para muitos lançamentos, ou valor esperado, é dado por:

**E(*X*) = E(U(1,6)) = (1+6) / 2 = 3.5**

Vamos fazer um experimento virtual usando 100 lançamentos de 11 dados.

O código em R para a seguir gera os dados e as visualizações de que precisamos:

>library(magrittr)

>library(ggthemes)

>library(ggplot2)

>source("multiplot.R")

>set.seed(2600)

>n\_plots <- 12

>dice\_fun <- function(n){runif(n, min=0, max=6) %>% ceiling} # Random samples

>data\_mat <- replicate(n=n\_plots-1,dice\_fun(100)) # Replicate

>data\_mat <- cbind(data\_mat,rowSums(data\_mat)) # Sum

>plot\_list <- vector("list", n\_plots) # Plot each distribution

>plot\_list <- apply(X=data\_mat, MARGIN=2, FUN=function(x)

ggplot(data.frame(obs=x),aes(x=obs)) +

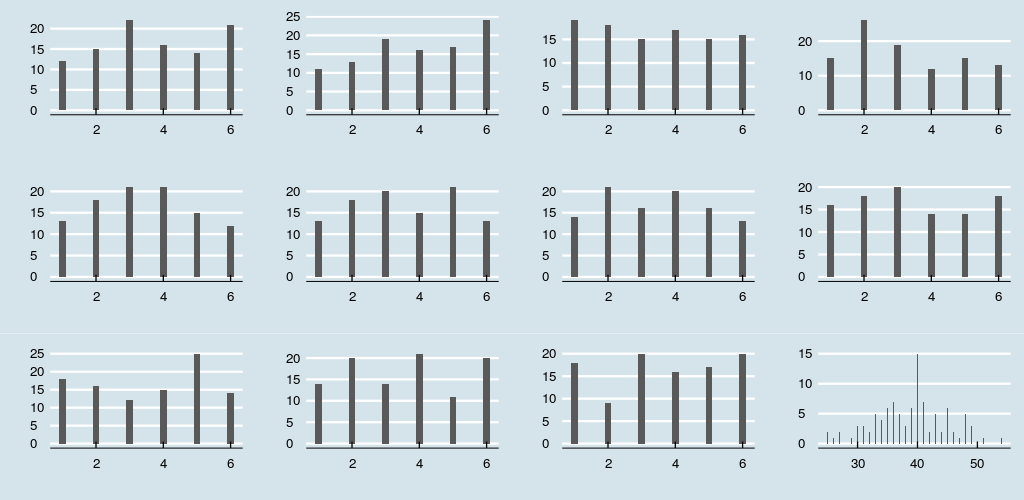
geom\_histogram(binwidth = 0.2)+

ylab("")+xlab("")+

theme\_economist())

>m\_plot <- multiplot(plotlist = plot\_list,cols=n\_plots/3)

# Multiplot function available at: http://www.cookbook-r.com/Graphs/Multiple\_graphs\_on\_one\_page\_(ggplot2)/



*Figura 2 . Soma de amostras (n=100) de 11 distribuições uniformes correspondentes ao lançamento de dados honestos de 6 faces. O resultado está na célula inferior à direita.*

Notamos que as barras estão distribuídas com alturas bastante parecidas nas 11 primeiras células. A frequência esperada para cada valor é ~1/6 do total de 100 lançamentos. *Freq* (X*i*) ~ 1/6\*100 ~ 16.66.

Algo interessante ocorre com a soma das distribuições (canto inferior direito).

O valor esperado é, como diz a intuição, a soma dos valores esperados em cada amostra:

***E* (*X*) = i=1∑11 *E* (Ui ~ (1,6) = 11\*3.5 = 38.5**.

O valor 38.5 corresponde aproximadamente ao centro da distribuição resultante (Figura 2, canto inferior direito)

Entretanto, a distribuição muda de forma! Sem muito esforço, é notável a semelhança com a curva normal, com valores extremos menos frequentes e simetricamente afastado da média (valor esperado).

É possível provar que a soma de muitas distribuições de uma mesma família converge para a distribuição normal em qualquer caso. Desde que estas sejam independentes. A esse resultado damos o nome de Teorema do Limite Central.[[3]](#footnote-2)

Este resultado tem uma sutil importância para o estudo dos fenômenos naturais através de experimentos.

## Ciência experimental e o Teorema do Limite Central

Muitos objetos de interesse para os cientistas são simplificações de fenômenos complexos. Um exemplo trivial está na cor da pele de seres humanos. Uma parte considerável depende do número de genes herdados relacionados à melanina. Eles se comportam de maneira aditiva. Assim, cada variante de gene extra pode contribuir para a cor final com X unidades na escala para medir pigmentação.

A cor de um indivíduo será influenciada pela soma dessas distribuições, o que é análogo à matemática descrita para os lançamentos de dados.



Podemos comparar grupos quanto a medidas fenotípicas finais (cor da pele) sem saber detalhes sobre as relações entre cada gene e seus mecanismos de expressão e regulação.

A distribuição final de melanina vem da soma distribuições individuais semelhantes e tenderá a ser normal.

Como vimos, o mesmo é válido para quaisquer distribuições subjacentes: se elas forem *gama*, uniformes ou de Poisson, a distribuição da soma ainda tenderá à normalidade.

A figura 2 mostra a soma de distribuições uniformes para dados honestos, evidenciando que esta se aproxima de uma normal.

X ~ U1(1,6) + U2(1,6) + … + U11(1,6) = X ~ N (38.5,σ²)

Vamos visualizar o mesmo processo para uma outra família de distribuições, gamma:

X ~ γ1(ɑ, 𝛽) +... + γn(ɑ, 𝛽) = X ~ N (μ’,σ’)

Para valores grandes de *n*:

>gamma\_fun <- function(n){rgamma(n,1)}

>data\_mat <- replicate(n=n\_plots-1,gamma\_fun(100))

>data\_mat <- cbind(data\_mat,rowSums(data\_mat))

>plot\_list <- vector("list", n\_plots)

>plot\_list <- apply(X=data\_mat, MARGIN=2, FUN=function(x)

ggplot(data.frame(obs=x),aes(x=obs)) +

geom\_histogram(binwidth = 0.2)+

ylab("")+xlab("")+

theme\_economist())

>m\_plot <- multiplot(plotlist = plot\_list,cols=n\_plots/3)

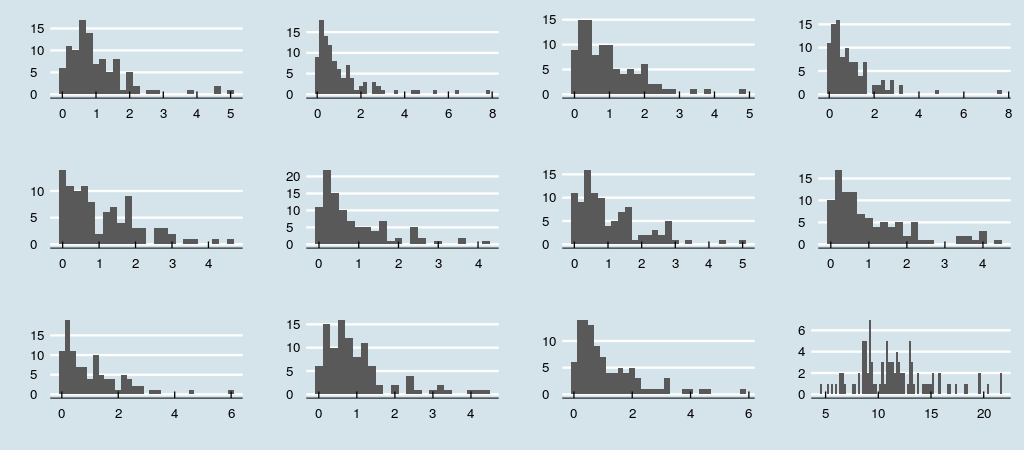


Figura 3. *Soma de amostras (n=100) de 11 distribuições gama. O resultado está na célula inferior à direita. Função de densidade de probabilidade para distribuição gama: ƒ(x) = 1/Γ(ɑ) · βa · xɑ-1 · e-βx, com ɑ = β= 1;*

Novamente, verificamos que a soma começa a ser simétrica em torno da média, com formato de sinos (base alargada).

Os fenômenos observáveis em nosso universo são naturalmente complexos. Especialmente em sistemas biológicos, há redundância de componentes e um objeto de interesse para cientistas é resultado da combinação de muitas variáveis subjacentes. O teorema do limite central permite que utilizemos distribuições normais para uma grande variedade de problemas. Ainda que as distribuições subjacentes sejam desconhecidas, a variável resultante de uma grande combinação será gaussiana em muitos casos.

A descoberta das equações que regem esses mecanismos de convergência foi um grande avanço para as ciências experimentais.

### Exercícios

1. Sobre a distribuição normal para uma variável aleatória, é verdadeiro (mais de uma possibilidade):
   1. A soma da probabilidade de todos os valores possíveis é 1.
      1. -Inf∫+Inf f(x) dx= 1.
   2. É simétrica em relação à moda.
   3. O valor esperado é dado por 1/σ√(2π).
   4. 95% dos valores estão próximos à média.
   5. Valores extremos são improváveis.
   6. É unicamente determinada por variância σ² e média μ.
   7. É contínua e diferenciável, facilitando cálculos.
   8. Amostras pequenas resultam em distribuições t.
2. Usando o comando “*?Distributions*” acesse algumas distribuições disponíveis na biblioteca de base do R.
   1. Plote o histograma da soma de 100 distribuições X² (*rchisq;*n = 60).
   2. Faça o mesmo procedimento para 100 distribuições de outra família e tamanho à sua escolha.
   3. Obtenha os valores de skewness e kurtosis para essas distribuições. Uma distribuição normal padrão (σ²=1;μ=0) possui skewness (assimetria) de 0 e kurtosis (volume dos valores mais extremos) de 3. Quais os encontrados por você?
   4. Cite dois fenômenos naturais cuja distribuição estatística é conhecida e qual a distribuição correspondente.

### 

### Darwins’s Finches

Mostraremos como a contribuição individual de genes com efeitos aditivo de distribuição uniforme resulta em medidas aproximadamente normais para os bicos das aves.

Vamos simular as medidas de bicos em 4 amostras (n=150) de pássaros.

O tamanho dos bicos é dado pelo efeito aditivo de muitos genes semelhantes, portanto esperamos que sua distribuição seja normal pelo Teorema do Limite Central.

Uma cópia do gene adiciona *x* milímetros ao tamanho final. O valor de x é sorteado de uma variável aleatória de distribuição uniforme, X ~ U (0,1).

Pássaros têm um número fixo de *n* de genes aditivos em cada amostra, sorteado no intervalo entre 80 e 100. A medida final dos bicos é dada pela soma efeitos dos *n* genes. Esse número é fixo em cada população e varia entre populações.

Para simular os dados com as condições acima:

>library(magrittr)

>library(ggthemes)

>library(ggplot2)

>set.seed(2600)

>n\_birds <- 150 # sample\_size

>genes\_low <- 80 # lower bound on number of genes

>genes <- 100 # upper bound on number of genes

>n\_islands <- 4 #samples

>unif\_sum <- function(genes){

replicate(n = genes,

expr = runif(100, min=0, max = 1)) %>%

rowSums

}

>generate\_pop <- function(n\_pop,n\_genes){

replicate(n=n\_pop,

expr = unif\_sum(n\_genes) %>% mean)

}

>galapagos\_birds <- purrr::map(.f = function(x) generate\_pop(n\_pop=n\_birds,

n\_genes = x),

.x = runif(n=n\_islands, genes\_low, genes) %>% ceiling) %>%

unlist %>% matrix(nrow=n\_birds,byrow=F) %>%

data.frame

Verificando o histograma das medidas finais, notamos, mais uma vez, que as medidas se aproximam de uma gaussiana.

>my\_alpha <- 0.5

>my\_bins <- 50

>ggplot(data=galapagos\_birds,aes(x=X1))+

geom\_histogram(alpha=my\_alpha,bins = my\_bins)+

geom\_histogram(data=galapagos\_birds,aes(x=X2),fill="dark blue",

alpha=my\_alpha,bins = my\_bins)+

geom\_histogram(data=galapagos\_birds,aes(x=X3),fill="dark red",

alpha=my\_alpha,bins = my\_bins)+

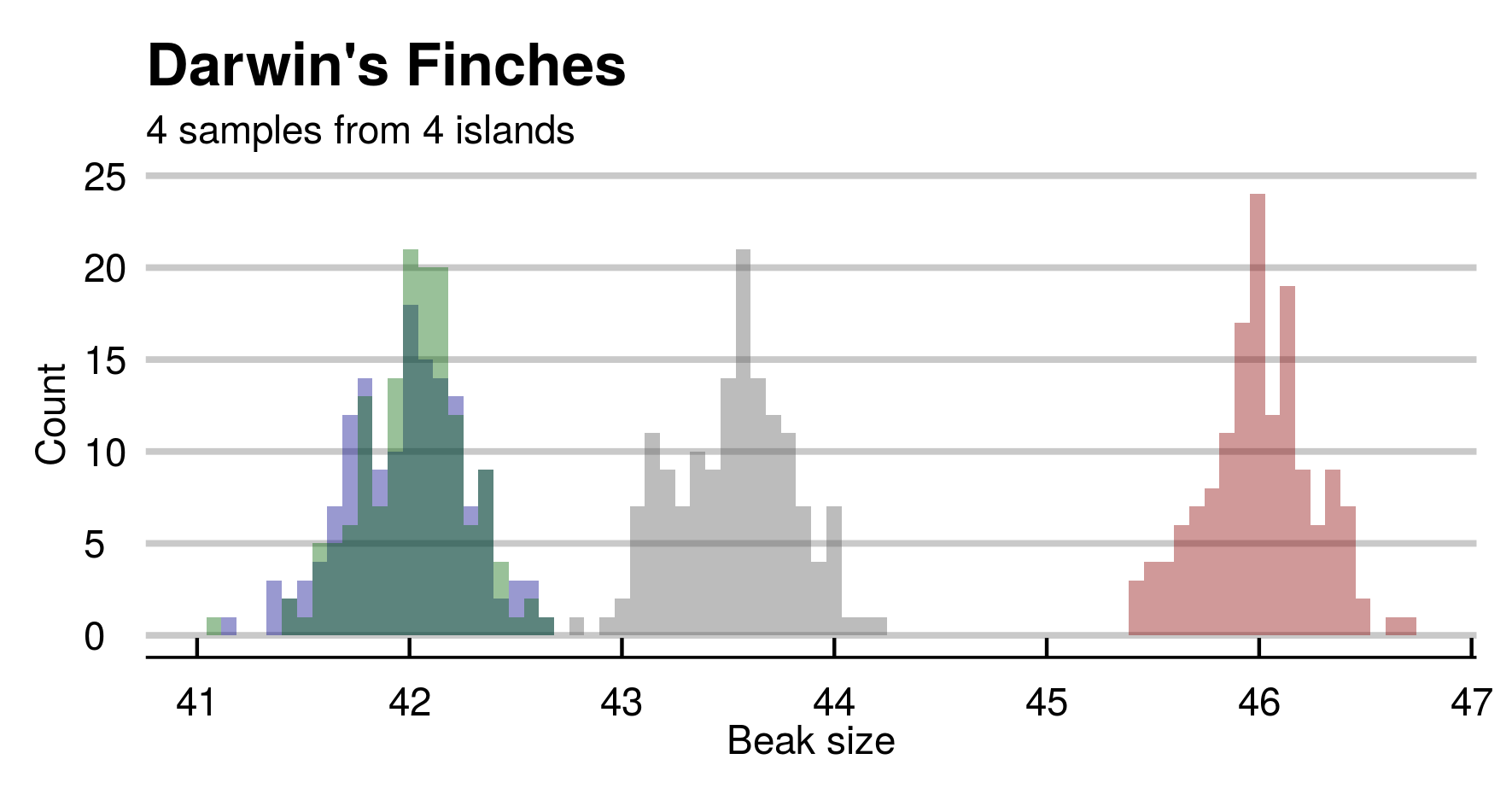
geom\_histogram(data=galapagos\_birds,aes(x=X4),fill="dark green",

alpha=my\_alpha,bins = my\_bins)+

xlab("Beak size")+ylab("Count")+

ggtitle("Darwin's Finches",subtitle = "4 samples from 4 islands")+

theme\_economist\_white(gray\_bg = F)



*Figura 4. Distribuição das medidas de bicos em populações simuladas para genes com efeito aditivo.*

Curiosamente, os números aleatórios gerados usando a semente sugerida (set.seed(2600) na linha 4) são semelhantes à suposição de Darwin: 4 ilhas (amostras) e três espécies (distribuições de bicos). Notamos que há duas amostras (verde, azul) de medidas bastante parecidas e outras duas separadas (cinza, vermelho).

Supondo que medimos os bicos de algumas aves, como saber se os grupos são diferentes? Calculando as diferenças entre distribuições, podemos inferir se duas amostras têm o mesmo número de genes subjacentes! Para isso, usaremos um racional e algumas ferramentas novas.

# Testes de hipótese

Filósofos da ciência estudam características no *modus operandi* de outros estudiosos. O que há em comum entre os procedimentos empregados por biólogos e geólogos? O que distingue Charles Darwin e Paul Dirac de John Dee e Edward Kelley? O que funciona em áreas distintas do conhecimento humano?

Adotamos a denominação coletiva de “ciências” para algumas áreas do conhecimento. Ainda, associamos a elas características em comum nos procedimentos e na estrutura interna. De alguma forma, cientificidade comunica credibilidade. Nas últimas décadas, filósofos discutiram a validade do problema de demarcar ciência de pseudociência e não-ciência.[[4]](#footnote-3) Neste capítulo, vamos nos ater a um paradigma conceitual mais antigo e indiscutivelmente influente.

O método hipotético-dedutivo foi popularizado no século XX como uma bandeira de identificação associada ao trabalho científico. Um ciclo que consiste em formular hipóteses, desenhar experimentos, testar hipóteses falseáveis, verificar resultados e repetir o processo de forma iterativa.

O racional em usar hipóteses testáveis é de que proposições válidas sobre um sistema contém informações que ajudam a prevê-lo. Assim, "faz sol ou não amanhã" é uma proposição inútil, enquanto "faz sol amanhã" é uma proposição útil. Note que "faz sol amanhã" é uma hipótese testável (falseável), enquanto "faz sol ou não amanhã" é uma hipótese verdadeira independente das observações.

O exemplo é grosseiro, porém alguns ramos do conhecimento humano produziram hipóteses não-falseáveis. K. Popper, líder da revitalização do método hipotético dedutivo no século passado, atacou severamente o materialismo dialético de Karl Marx, assim como a teoria de evolução por seleção natural de Charles Darwin. Por exemplo, Marx previu que a revolução aconteceria em um país industrializado e outros eventos que não se concretizaram. Seus seguidores justificaram com hipóteses ad-hoc. A teoria da evolução por seleção natural de Darwin era amparada em muitos exemplos de reprodução impossível (recomposição da trajetória evolutiva em fósseis). A psicanálise também sofreu duras críticas, em virtude da irrefutabilidade de seus pilares centrais.

Para Popper, a dificuldade em gerar hipóteses testáveis e falseáveis sinalizava uma evidente fragilidade nas teorias, as quais não empregariam métodos científicos em seus avanços.

Uma maneira de formalizar essa ideia, incorporando o uso de ferramentas quantitativas, é através de probabilidades. Calculamos a probabilidade associada a observações, considerando o cenário de uma hipótese (falseável).

Esse racional adequa ferramentas matemáticas robustas à plataforma epistemológica de Popper, sendo um modelo dominante de produção em ciências experimentais.

## Testes de hipótese

Em geral, os pesquisadores formulam uma hipótese base, chamada hipótese nula, que descreve o cenário menos interessante para o trabalho. Por exemplo, se estamos comparando dois grupos, A e B, quanto a uma intervenção, a hipótese nula costuma declarar que os grupos são iguais.

Queremos estudar o tamanho dos bicos de pássaros das ilhas A e B. A hipótese nula natural é: Não há diferença entre os bicos dos pássaros do tipo A e B.

Medimos o bico de alguns pássaros dos dois grupos e calculamos a probabilidades de encontrarmos essas medidas considerando que A e B são iguais. Se essa probabilidade for muito baixa, rejeitamos nossa hipótese.

Estruturando os passos:

1 . Definimos a hipótese nula (H0) e pelo menos uma hipótese alternativa(H1).

H0: Pássaros das ilhas A e B possuem bicos de tamanho igual.

H1: Os pássaros possuem bicos de tamanho diferentes.

Então, podemos fazer um experimento, coletando medidas experimentais para o comprimento dos bicos. Essas medidas, junto a premissas matemáticas razoáveis, permitem especular: qual a probabilidade *p* de obter nossas observações considerando distribuições iguais entre A e B? Isto é, considerando H0 verdade, nossos resultados seriam raros ou comuns?

Caso *p* seja menor que um limiar arbitrariamente pré-definido (convencionalmente, 0.05), rejeitamos H0. A probabilidade é pequena se H0 for verdade.

A interpretação do valor p não é muito intuitiva. Envolve mensurar quão improváveis são as observações em um cenário hipotético na vigência da hipótese nula. Sua tradução (errada) mais popular é de que representa "a chance de o resultado deste estudo estar errado".

A domínio dos procedimentos hipotético-dedutivos nas ciências produziu resultados interessantes. No sentido proposto por Thomas Kuhn de ciência normal como acúmulo de evidências e testagem de hipóteses. O fantasma de desenhar um experimento com possibilidade imparcial de falha aguçou a percepção de pesquisadores para a falibidade de ideias. O grau de sofisticação em reprodutibilidade de procedimentos foi amplificada.

Em contrapartida, o arcabouço acima é suficiente para produzir um trabalho científico críptico para leigos. Seguindo receitas pré-definidas, um texto parece estar em conformação com os padrões acadêmicos, ainda que a hipótese elementar em torno do objeto de pesquisa seja simplória. Outro efeito colateral é a busca por valores p baixos, desprezando precedentes teóricos e premissas probabilísticas (múltiplos testes).

## Um exemplo prático: Teste t para amostras independentes

Para testar estatisticamente se as medidas são diferentes, executaremos um teste t para comparação dos grupos.

A distribuição *t* surge quando queremos entender quão improváveis são nossas estimativas (μ’) supondo uma média real hipotética (μ) de origem em uma variável de distribuição normal desconhecida.

**Exemplo**: Medimos os bicos de 30 pássaros. Obtivemos média amostral μ’ = *38 mm* com desvio-padrão σ’ = 0.3 *mm*. Supondo que a média real (μ) é de 40 *mm*, qual é a probabilidade obtermos μ’ = *38 mm* em uma amostra aleatória (nossa situação)?

Entender a imprecisão da estimativa de uma média foi o eixo principal para a descrição dessa distribuição por William Gosset. Sob o pseudônimo *Student*, o estatístico, que trabalhava para a fábrica de cerveja Guiness, publicou na Biometrika (1908) o famoso texto *The probable error of a mean*.

Para entender a imprecisão, necessitamos de uma medida da dispersão dessas medidas.

Assumimos amostras retiradas de uma variável aleatória com distribuição normal e média μ. Podemos retirar *j* amostras de tamanho *N* e calcular a média dessas amostras μ’1,μ’2, …, μ’*j*.. As médias amostrais μ’*j* são estimativas da média real μ.

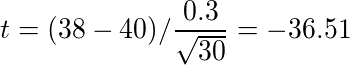
Qual a dispersão das estimativas μ’1,μ’2, …, μ’*j* ?

Para um conjunto de estimativas μ’1,μ’2, …, μ’*j* chamamos ela de ***erro padrão***(*standard error of the mean*), dado pelo **desvio-padrão populacional** dividido pela raiz quadrada do **tamanho da família de amostras** em questão (*std. err.* = σ/√n). Como não sabemos o desvio-padrão na população, usamos o valor de σ’j das amostras.

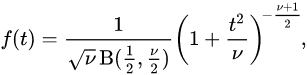
Student propôs o uso de uma quantidade para estimar a probabilidade de uma estimativa μ’ dado centro μ. Essa quantidade pivotal é a razão entre **(1)** distância das estimativas e média real, μ’ - μ, e **(2)** o erro padrão. A estatística **t**:



Assim, a estatística t para nosso exemplo (*μ’=38; μ= 40; n=30; σ’=0.3*) é:

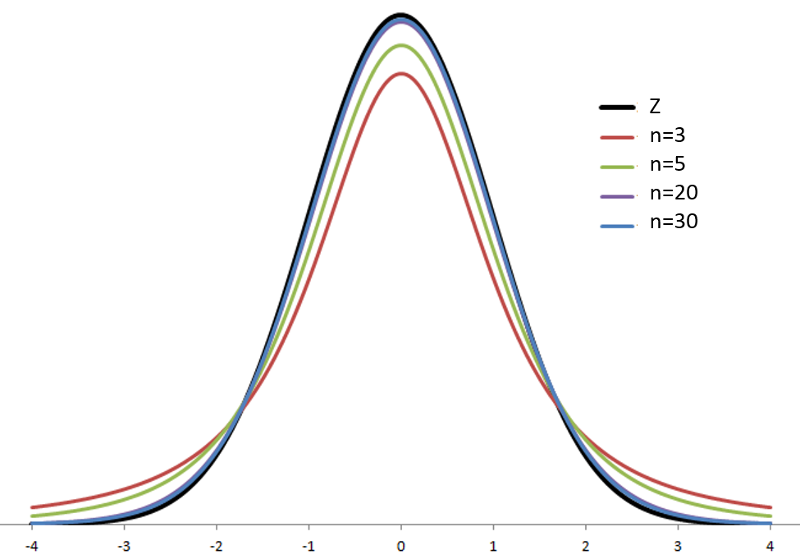


Student (Gosset) mostrou que essa estatística segue uma distribuição probabilística (t de *Student)* definida:



B é a função Beta e v são graus de liberdade. Possui densidade parecida com a da distribuição normal, porém com probabilidades maiores para valores extremos. O parâmetro v (graus de liberdade) expressa essa característica. Empiricamente é estimado pelo tamanho das amostras usadas na estimativa de μ’.

Associamos uma amostra (tamanho *n*) retirada de uma população normal (tamanho arbitrariamente alto, n → oo) a uma distribuição *t* com (*n-1)* graus de liberdade. Em nosso exemplo, n=30, então v = 29 (n-1).



Maiores valores correspondem a amostras maiores e fazem com que a distribuição t se aproxime de uma distribuição normal. Em um caso extremo, temos nsamples=npop. e as amostras são idênticas à distribuição de origem.

* Sabendo a estatística *t (t = -36.51)* e os graus de liberdade para nossa família de amostras (v = 29), podemos usar a expressão ƒ(t) para saber a probabilidade de obtermos nossa média *38 mm* se a média real for de *40 mm*.

Em R, a função pt retorna a probabilidade total de valores extremos menores que a estatística t fornecida.

>pt(36.51, df = 29)  
[1] 4.262e-26

A nossa hipótese considera valores t extremos maiores (direita) ou menores (esquerda), Então consideramos também o valor p associado aos extremos maiores que a estatística t simétrica positiva (tmin= -36.51; tmax= 36.61). É o que encontraríamos com a medida simétrica em relação à média de *42 mm.*

A seguinte expressão deveria funcionar.

>(1- pt(36.51, df = 29)) + pt(-36.51, df = 29)  
[1] 4.262e-26

Entretanto, o valor retornado parece idêntico ao de pt(-36.51, df = 29). Acontece que o R aproxima a soma das probabilidades de valores menores que t = 36.51 para 1, zerando o primeiro componente da soma (pt(36.51, df = 29)). Sabendo que a distribuição t é simétrica, podemos fazer então usar o seguinte truque:

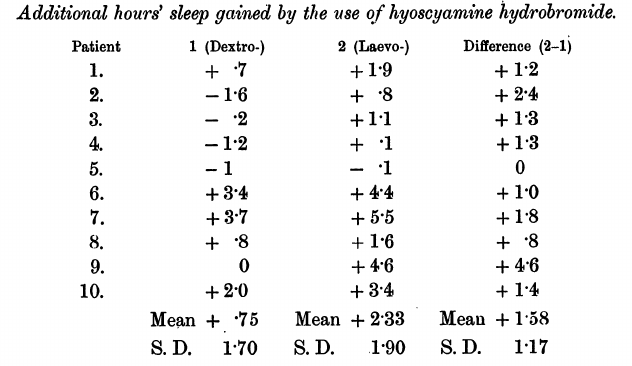
> 2\*pt(-36.51, df = 29)

[1] 8.524e-26 # valor p

#### Nota

Uma percepção errônea comum sobre a distribuição *t* é de que ela é resultado amostras pequenas retiradas de uma variável aleatória normal. Qualquer amostra retirada de uma variável de distribuição normal terá, por definição, distribuição normal, ainda que seja composta por 1 ou 2 observações. O que segue distribuição *t* é a quantidade pivotal descrita acima.

Na sessão IX do artigo, Student (Gosset) demonstra como seu insight pode ser usado para testar o efeito de isômeros da escopolamina como indutora do sono.[[5]](#footnote-4) São usadas duas amostras (*levo* e *dextro hidrobromido de* *hyoscyamina).*

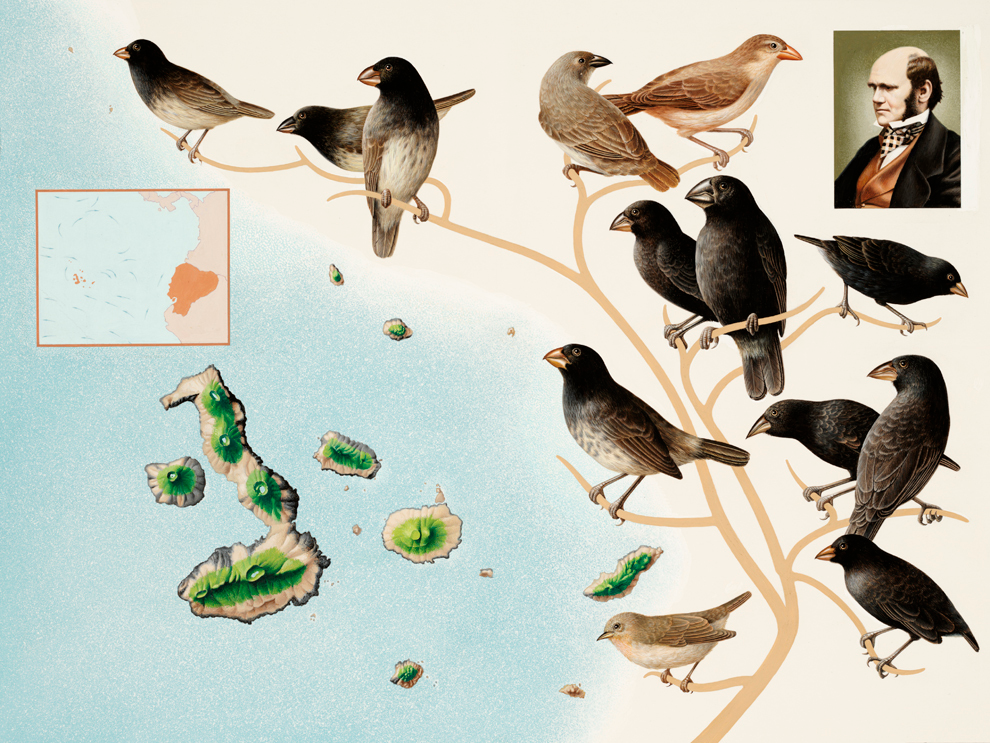


Retirado de The probable error of a mean, *pag. 20.* Is dados estão disponíveis na biblioteca de base do R, sob o nome '*school*’.

Usando dados de 10 pacientes que usaram ambas as substâncias e medidas da quantidade adicional de horas de sono observadas, “Student” calcula: (1) a probabilidade dos dados supondo média 0 em cada grupo e (2) a probabilidade dos dados supondo que a diferença das médias é 0.

O primeiro procedimento foi realizado com a medida dos bicos e é chamado teste t de amostra única. Hipotetizando um valor para a média (e.g. μbico= 40 *mm*; μsono adicional= 0 horas), calculamos as probabilidades de nossa estimativa.

O segundo procedimento é chamado de teste t de amostras independentes. Hipotetizamos um valor para diferença de médias entre duas populações (μa-μb= 0) e calculamos a probabilidade de nossa estimativa. *Exemplo prático: existe diferença de peso entre os bicos dos pássaros A e B?*



Aplicação

Retornando ao nosso exemplo de Galápagos, faremos um teste t de amostras independentes.

1. As medidas em A e B são amostras de variáveis aleatórias com distribuição normal.
2. Definimos a hipótese nula e pelo menos uma hipótese alternativa.
   1. H0: Pássaros das ilhas A e B possuem bicos de tamanho igual.
      1. μa - μb =0
   2. H1: Os pássaros possuem bicos de tamanho diferentes.

O procedimento é semelhante ao anterior. Calculamos uma quantidade intermediária que segue distribuição t usando a estimativa amostral da diferença e erro padrão associado. Então, podemos especular: qual a probabilidade *p* de alguém obter nossas observações considerando distribuições de médias iguais (μa = μb)? Esse teste infere a probabilidade para as populações de onde saíram as amostras.

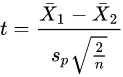
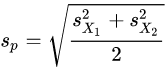
Caso *p* seja menor que um limiar arbitrariamente pré-definido (convencionalmente, 0.05), rejeitamos H0.A probabilidade de observarmos os dados é pequena se H0 for verdade.

Obtemos o valor p somando os valores de probabilidades correspondentes às diferenças obtidas ou valores mais extremos. Caso a diferença entre valores seja grande, o valor da estatística crescerá. Isso implica uma baixa probabilidade de observar aqueles resultados se as amostras fossem semelhantes (vindas da mesma distribuição).

### 

### Aplicação com software

Vamos computar um teste t para 2 amostras independentes. A estatística t é calculada com algumas mudanças: os graus de liberdade são somados e o erro padrão é balanceado entre amostras.

, sp corresponde ao desvio-padrão balanceado.

Considerando (n1-1) + (n2-1) graus de liberdade, calculamos a estatística *t* e o valor p correspondente para nossos graus de liberdade. Usando as amostras criadas anteriormente, correspondentes às barras cinza (A) e azul(B), vamos plotar os histogramas.

>ggplot(data=galapagos\_birds,aes(x=A))+

geom\_histogram(alpha=my\_alpha,bins = my\_bins)+

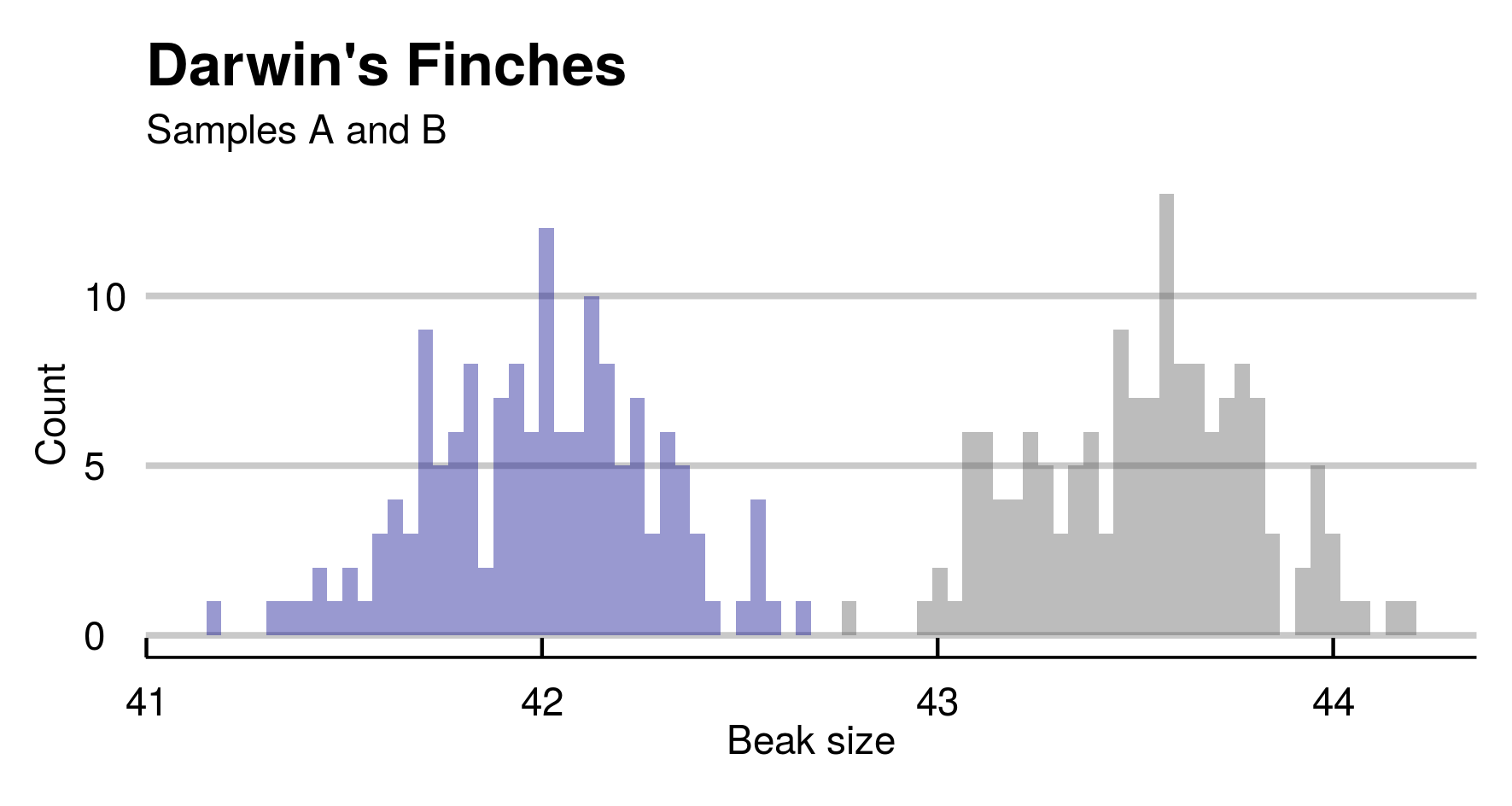
geom\_histogram(data=galapagos\_birds,aes(x=B),fill="dark blue",

alpha=my\_alpha,bins = my\_bins)+

xlab("Beak size")+ylab("Count")+

ggtitle("Darwin's Finches",subtitle = "Samples A and B")+

theme\_economist\_white(gray\_bg = F)



# Ajustes nos dados

>a <- galapagos\_birds$X1

>b <- galapagos\_birds$X2

>sd\_a <- sd(a) #desvio-padrao

>sd\_b <- sd(b)

Aqui, ao invés de comparar as estimativas das médias de distribuição t para amostras A e B.

Calculamos a (1) Diferença esperada na vigência da hipótese nula (diffH0 = 0), (2) estimativa da diferença (diff = μA-μB), graus de liberdade (*df*) e erro padrão balanceado ( *sepooled* ) para a distribuição das diferenças de médias:

>expected\_diff <- 0

>mean\_diff <- mean(a) - mean(b) #mean difference

>df\_pool <- length(a) + length(b) - 2 # degrees of freedom

>se\_pool <- sqrt(((length(a) - 1) \* sd\_a^2 + (length(b) - 1) \* sd\_b^2)/

df\_pool) # pooled std. error

A estatística t correspondente à diferença observada, considerando uma distribuição t com os parâmetros calculados acima.

>t <- (mean\_diff - expected\_diff)/ (se\_pool \* sqrt(1/length(a) + 1/length(b))) # t-statistic

Valor p para hipótese bicaudal (resultados extremos considerando a possibilidade de a diferença ser maior ou menor que 0):

>p <- 2\*pt(-abs(t), df = df\_pool) # two-sided p-value

Resultados

> result <- c(mean\_diff, t, p, mean(a), mean(b))

>names(result) <- c("Difference of means", "t", "p-value","Mean A","Mean B")

> result

Difference of means t

1.533321e+00 4.728513e+01

p-value Mean A Mean B

1.532661e-140 4.352244e+01 4.198912e+01

As funções *length*, *sqrt* (raiz quadrada), *sd*(desvio-padrão) e *mean*(média) são evidentes e comuns na biblioteca base de outras linguagens.

Obtivemos um valor p significativo (p < 0.001) usando n = 150. Os graus de liberdade são 149 (150 - 1) em cada amostra, sendo 298 ao total*.*

Alternativamente, para fins práticos, o R possui em sua biblioteca de base uma função para automatizar o processo em 1 linha:

> t.test(a,b,var.equal = T)

Two Sample t-test / data: a and b

t = 47.285, df = 298, p-value < 2.2e-16

Alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: 1.469506 1.597136

Sample estimates:

mean of x mean of y

43.52244 41.98912

Estatística t e graus de liberdade apresentados pela biblioteca base do R(base::t.test) são idênticos aos que encontramos realizando o procedimento ‘manualmente’.

Diante do valor p obtido, concluiríamos que a distribuição dos dados como observada é improvável se for verdade a hipótese nula H0 de que a diferença entre amostras seja 0.

Assim, **nosso procedimento sugere que há diferença entre o tamanho dos bicos nos pássaros A e B**.

Simplificando outras nuances, o método hipotético-dedutivo proposto por K. Popper tem como fundamentação a testagem de hipóteses. Assim, experimentos devem ser testados com o objetivo claro de testar teorias, confirmando-as ou falseando-as.

Na apresentação feita, usamos a estrutura para testar uma hipótese bem definida sobre uma entidade natural. Isto é, dadas medidas aleatórias feitas em pássaros das ilhas A e B, podemos afirmar que os bicos de um tipo de ave são maiores que os do outro tipo?

Usando premissas relacionadas ao Teorema do Limite Central, estudamos as distribuições Normal e T para embasar as inferências sobre a amostra.

### Exemplo de relatório

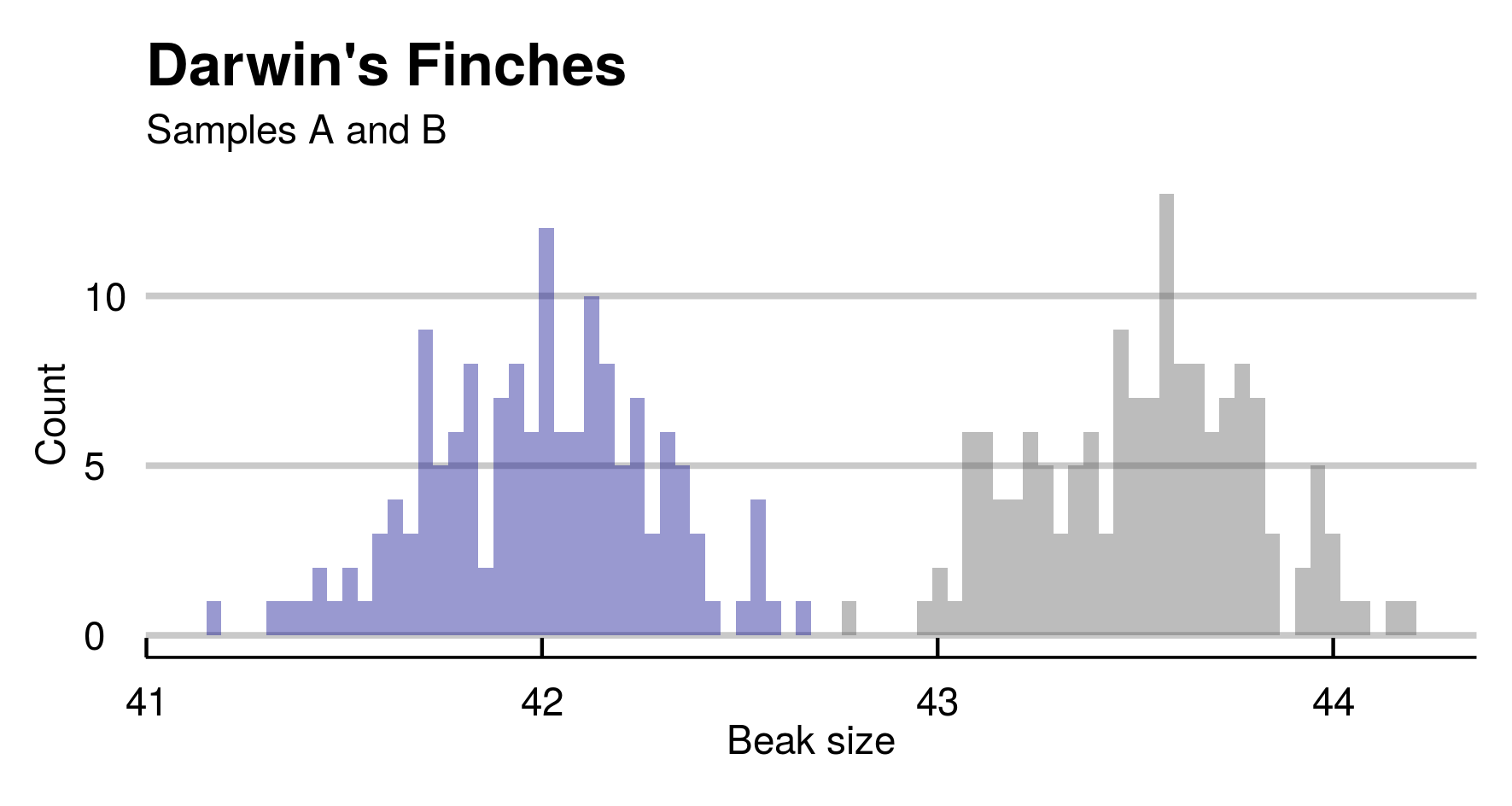
A diferença estimada entre tamanho médio dos bicos entre amostras A e B foi significativamente (p <0.05) diferente de 0 (t=47.28; df = 298).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Amostra A  (n= 150) | Amostra B  (n= 150) | valor p |
| Média (μ) | 43,52 | 41,99 | < 0,001 |
| Desvio-padrão (σ) | 0,28 | 0,28 |  |

### Report example

The estimated difference of beak mean sizes among samples A and B was significantly (p<0.05) different from zero (t = 47.28, df = 298)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sample A  (n= 150) | Sample B  (n= 150) | p value |
| Mean (μ) | 43.52 | 41.99 | < 0.001 |
| Std. Dev. (σ) | 0.28 | 0.28 |  |



### Exercícios

1. Usando o dataset simulado no capítulo:
   1. Execute teste T para cada par de amostras
   2. Quais testes apresentam p < 0.05?
      1. Descreva estatística t, graus de liberdade e valor p.
         1. Como são os graus de liberdade dos diversos testes?
         2. Esses valores eram esperados para nossas amostras?
      2. Plote histogramas para todos os pares comparados **em apenas um painel**. Dica: *grid.arrange*
      3. Plote boxplots para uma das comparações.
      4. Plote violin plots (geom\_violin) para uma das comparações.
2. Usando o dataset *iris*
   1. Escolha duas espécies e duas medidas.
   2. Execute testes t para ambas as medidas
   3. Reporte os resultados em uma tabela, incluindo média e desvio-padrão de ambas as medidas nas duas espécies.
3. Os dados usados por Student para escopolamina estão incluídas na biblioteca de base do R.
   1. Examine os dados invocando “sleep”: >*sleep*
      1. Plote histogramas para as medidas em ambos os grupos
      2. Execute um teste t supondo média população zero (μ = 0).
      3. Execute um teste t entre amostras, supondo têm a mesma média (H0:μ2 = μ2).
4. Gerando a distribuição t:
   1. Simule um conjunto de muitas medidas (sugestão: 100,000) a partir de uma distribuição normal(μ = 0, σ=1).
   2. Retire 200 amostras de n=30 e salve as 200 médias (função *sample*).
   3. Divida os valores por pelo erro padrão, σ/√n.
   4. Retire 200 amostras de uma distribuição t com 29 graus de liberdade (função *rt*)
   5. Plote o histograma superposto da distribuição obtida e da distribuição teórica

1. *É bastante notável que uma gradação quase perfeita na estrutura desse grupo possa ser traçada na forma do bico, desde um excedendo as dimensões do maior dos pardais bico-gordo, até outro diferindo pouco do papa-amoras.*  [↑](#footnote-ref-0)
2. *(...)[ao] ver esta gradação e diversidade em estrutura em um pequeno, intimamente relacionado grupo de pássaros, é possível imaginar que, a partir de poucos pássaros deste arquipélago, uma espécie foi escolhida e modificadas para certos fins. Darwin, Charles (1845), Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N (2nd. ed.), London: John Murray* [↑](#footnote-ref-1)
3. Prova formal: http://www.cs.toronto.edu/~yuvalf/CLT.pdf [↑](#footnote-ref-2)
4. Massimo Pigliucci - Philosophy of Pseudoscience: Reconsidering the Demarcation Problem [↑](#footnote-ref-3)
5. https://atmos.washington.edu/~robwood/teaching/451/student\_in\_biometrika\_vol6\_no1.pdf [↑](#footnote-ref-4)