

Inferência em Ciências e Aprendizagem de Máquina

Filosofia e aplicações com estatística e probabilidade.

Felipe Coelho Argolo felipe.c.argolo@protonmail.com

Londres, 8 de Julho de 2020

Página oficial: <https://www.leanpub.com/fargolo>

Volume 1 Segunda Edição

Prefácio

Lembre-se de que todos os modelos são errados; a questão prática é quão errados eles precisam ser para não serem úteis

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful

George Box & Norman R. Draper, Empirical Model-Building and Response Surfaces

Quando entramos no século XXI, os termos *inteligência artificial (artificial intelligence)*, *análise quantitativa (quantitative analysis)*, *aprendizagem de máquina (machine learning)*, *big data* e *ciência de dados* ganharam forte notoriedade em virtude de resultados inéditos em problemas de aplicação prática. Avanços técnicos em processamento de linguagem natural, visão computacional e outros campos foram rapidamente implementados por engenheiros e pesquisadores em finanças, indústria e ciências.

Estas soluções usam modelos estatístico-probabilísticos para modelar medidas empíricas. Um estudo sistemático do formalismo e das ferramentas envolvidas é volumoso (ver Lista de *Leituras recomendadas*).

Quod est inferius est sicut quod est superius. Et quod est superius est sicut quod est inferius, ad perpetranda miracula rei unius.

"O que é inferior é como aquilo que é superior. E o que é superior é como aquilo que é inferior, perpetuando os milagres de uma coisa só. ¹

O eixo filosófico deste texto está na investigação de fômenos naturais universais (ciência). Ele introduz tópicos em filosofia, estatística e probabilidade usados na investigação científica. Os temas de aprendizagem de máquina e inteligência artificial estão ligados aos métodos aprendidos. A difícil interpretabilidade de alguns desses modelos implica em questões de ordem prática (e ética) específicas.

Adinkras

As Adinkras, como a ave que ilustra a capa (Sankofa) são símbolos na cultura Akan. Eles representam conceitos e conhecimentos populares específicos, conectados à sua forma. Físicos teóricos também adotaram o nome para grafos representando as regras formais que regem partículas em modelo supersimétricos da gravidade.

Um dos problemas na primeira versão foi encontrar títulos e temas adequados, que sintetizassem cada capítulo. Encontrar os Adinkras e os conceitos que eles representam foi um fortuito acontecimento. Os textos passam a estar ancorados a estes símbolos.

1 . ADINKRA HENE

¹<http://catb.org/esr/writings/unix-koans/shell-tools.html>

Formado por círculos concêntricos, o Adinkra Rei é relacionado à inspiração e à criação dos outros Adrinkras.

São introduzidas intuições básicas em estatística descritiva e probabilidade (média e variância). Partindo do estudos de Arquimedes sobre alavancas, formas de descrever amostras e variáveis aleatórias usando intuições básicas. Aborda também a relação entre ciências empíricas e a distribuição normal.

2 . DWENNIMMEN

Representa visão de dois carneiros lutando. O Chifres de Carneiro simbolizam força e humildade, pois carneiros lutam ferozmente contra outros pares e predadores, porém aceitam o abate quando em cativeiro.

A identidade da ciência é fortemente ligada ao uso criterioso de experimentos para testar hipóteses. Elas abrem espaço para falhas.

O *segundo capítulo* acompanha Charles Darwin em Galápagos. Darwin esperou 20 anos entre a concepção da teoria e sua publicação. Trabalhou incansavelmente para investigar se suas impressões não eram falsas. Este capítulo ilustra como o racional hipotético-dedutivo funciona para estudar hipóteses científicas. O teste *t* de Student é aplicado para comparação dos bicos de aves em Galápagos.

3 . FUNTUNFUNEFU-DENKYEMFUNEFU

Os crocodilos que compartilham um estômago, mas que lutam por comida. Simbolizam unidade e cooperação. O *terceiro capítulo* destaca o papel descritivo e preditivo de teorias. Além de testar hipóteses, criamos modelos para as relações entre medidas. Aprenderemos correlações lineares (ρ de Pearson) e tamanho de efeito (*D de Cohen*). Também são introduzidas alternativas não-paramétricas: ρ de Spearman e teste U de Mann-Whitney).

Usamos regressão para fazer previsões usando *formas fechadas*. Resolvendo analiticamente as equações do modelo, encontramos uma estimativa única para os parâmetros envolvidos.

4 . AKOMA NTOSO Os corações conectados simbolizam concordância e entendimento mútuo. Com muitas variáveis (análise multivariada), grafos são a abstração base para relacionarmos conceitos. Estudamos regressão múltipla e sobre como lidar com covariáveis segundo um diagrama causal. O *quarto capítulo* introduz uma implementação formal do abrangente paradigma filosófico para **causalidade**. Colinearidade, confundidores, mediação e moderação. Para redução de dimensões e abordagem de medidas latentes, falaremos em análise factorial, análise de componentes principais (PCA) e sua generalização em equações estruturais (SEM).

5 . NEA ONNIM NO SUA A, OHU “Aquele que não conhece pode conhecer pela aprendizagem”. O *quinto capítulo* introduz redes neurais. Começamos da inspiração biológica dos neurônios artificiais e da primeira máquina inteligente da história: o *Mark I Perceptron*. Codificamos um Mark I virtual, que usa uma nova forma de estimar parâmetros: *gradient descent*. Ao invés de usar uma fórmula fechada, usamos derivativas para ‘caminhar’ em direção ao mínimo progressivamente.

Redes Neurais expandem o poder de um neurônio com múltiplos nodos para a construção de sistemas preditivos complexos. Redes profundas incluem camadas sucessivas, permitindo transformações em sequência para resolver classes mais gerais de problemas. Entendemos como os neurônios podem propagar erros aos outros, otimizando gradientes em conjunto com o mecanismo de *backpropagation*. Também codificaremos uma rede neural, Mark II.

6 . SANKOFA (San - Voltar; Ko - Ir ; Fa - Procurar, pegar)

O Adikra está ligado a retornar ao passado e aprender com ele. Modelos Bayesianos incorporam informações prévias (*prior*) em sua formulação. O *sextº capítulo* analise o suposto embate entre as escolas de probabilidade **frequencista** e **bayesiana**. O contexto é dado por alternativas ao método hipotético dedutivo: Carnap demonstra a dificuldade de refutações, Feyerabend propõe uma anarquia epistemológica amparada em fatos históricos e W. van Quine pinta um sistema entrelaçado para teorias, hipóteses e observações. Reabordamos alguns exemplos anteriores usando Stan para inferência bayesiana.
Exploramos uma terceira forma de estimar parâmetros. Sem fórmulas fechadas, usamos o poder das simulações estocásticas (*Markov Chain Monte Carlo*).

Prefácio à segunda edição

Aproximadamente um ano se passou desde o lançamento da 1^a edição. Algumas modificações importantes foram incorporadas.

Julia foi incluída como uma linguagem alternativa a R. É uma linguagem com uma comunidade menor, mas bastante promissora. Além de oferecer maior velocidade de execução, ela oferece uma sintaxe mais concisa para os exemplos.

Entrei em contato com o trabalho de Richard McElreath (Statistical Rethinking), o que resultou em frutos positivos: O capítulo 1 inclui uma segunda perspectiva (máxima entropia) para a utilização da distrição normal em ciências naturais. O capítulo 3 foi reestruturado para incluir uma abordagem mais geral do estudo de causalidade com grafos direcionados, usando o pacote/software **dagitty**. O capítulo 5 tem trechos relacionados à escolha de priors e avaliação de performance.

Na primeira edição, o trabalho de Ron Eglash em etnomatemática já havia influenciado no uso das cores verde e amarelo, associadas à Orumla e à divinação Iorubá, que usam números binários. Os Adinkras trazem uma nova camada à estética do livro.

Sumário

Capítulo 1 - ADINKRAHENE - Média, variância

- Centro e dispersão
 - Média e variância
- Distribuição normal
- Ciência experimental e o Teorema do limite central
- Momentos

Capítulo 2 - DWENNIMMEN - Os pássaros de Darwin e o método hipotético dedutivo

- Pássaros em Galápagos
- Método hipotético-dedutivo e Testes de hipótese
 - Valor p
 - Distribuição t de Student e teste t

Capítulo 3 - FUNTUNFUNEFU-DENKYEMFUNEFU - Sobre a natureza das relações

- Prelúdio: Quem precisa do valor p?
- Tamanho de efeito: D de Cohen
- Correlações lineares
 - Coeficiente de correlação ρ de Pearson
 - Predições com regressão linear
- Correlações e testes não paramétricos
 - ρ de Spearman
 - Teste U de Mann Whitney

Capítulo 4 - AKOMA NTOSO - Análise multivariada, grafos e inferência causal

- Regressão múltipla
 - Colinearidade
- Grafos e trajetórias causais
 - Mediação e moderação
 - Análise fatorial
 - Equações estruturais

Capítulo 5 - NEA ONNIM NO SUA A, OHU - Neurônios

- Regressão logística
- Um neurônio artificial: O perceptron
 - História e implementação do zero : Mark I
- Redes Neurais e Deep learning (múltiplas camadas)
- Gradient Descent
- Backpropagation

Capítulo 6 - SANKOFA - Contexto e inferência Bayesiana

- Probabilidades

- Frequencistas e Bayesianos
- Muitos métodos científicos: Feyerabend, Carnap e Quine
- Inferência Bayesiana
 - Teorema de Bayes
 - Intuições: prior, likelihood, posterior e probabilidades marginais
 - Comparação de amostras com distribuição normal
 - Correlação linear
- Estimadores e Métodos Markov Chain Monte Carlo
 - Soluções fechadas, Gradient Descent e MCMC

Pré-requisitos

Todos os exemplos podem ser reproduzidos usando software livre.

Leitura recomendada:

Filosofia e divulgação científica

- Surely You're Joking, Mr. Feynman
- O mundo assombrado pelos demônios - Carl Sagan
- A lógica da pesquisa científica - K. Popper
- A estrutura das revoluções científicas - Thomas Kuhn
- Contra o Método - Paul Feyerabend
- Dois dogmas do empiricismo - Willard van Quine
- Stanford Encyclopedia of Philosophy - <https://plato.stanford.edu/>
- The Open Handbook of Formal Epistemology - <https://jonathanweisberg.org/post/open-handbook/>

Neurociências

- Principles of neural science - Eric Kandel

Matemática/computação

- Coleção '*Fundamentos da matemática elementar*'
- Statistical Rethinking. A Bayesian Course with Examples in R and Stan, Richard McElreath.
- Bioestatística sem segredos. Annibal Muniz.
- What is mathematics - Courant & Robbins
- Better Explained (<https://betterexplained.com/>)
- <http://material.curso-r.com/>
- R Graphics Cookbook
- R Inferno
- Learn you a Haskell for Great Good
- Layered Grammar of Graphics - Hadley Wickham.
- Algorithms unlocked
- Online: Statsexchange, stackoverflow, mathexchange, cross-validated.

Machine Learning

- An Introduction to Statistical Learning: with Applications in R
- Neural Networks and Learning Machines - Simon Haykin
- Stanford (computer vision): <http://cs231n.stanford.edu/>
- Oxford 2015 (Deep learning): (<https://www.youtube.com/watch?v=dV80NAIEins&list=PLE6Wd9FR--EfW8dtjAuPoTuPcqmOV53Fu>)

Agradecimentos

Minha família, Suzana, Paulo, Isaac e Chris. Amigos Gabriel, Guilherme, Wei.

Aos professores: Carla Daltro, Anibal Neto, Lucas Quarantini, Luis Correia, Rodrigo Bressan, Ary Gadelha.

Aos colegas Fatori, Luccas, Macedo, Walter, Rafael, Sato, Hiroshi, Lais, Luci, Davi, n3k00n3 (Fernando), Loli (Lorena).

Para comentários, críticas, sugestões, ou simplesmente dizer *oi*: felipe.c.argolo@protonmail.com.

ADINKRAHENE - Media, variância + Momentos <http://www.adinkra.org/htmls/adinkra/adin.htm>
DWENNIMMEN - Hipóteses <http://www.adinkra.org/htmls/adinkra/dwen.htm>
FUNTUNFUNEFU-DENKYEMFUNEFU - Correlações e regressão <http://www.adinkra.org/htmls/adinkra/funt.htm>
AKOMA NTOSO - Múltiplas variáveis <http://www.adinkra.org/htmls/adinkra/akon.htm>
NEA ONNIM NO SUA A, OHU - Aprendizagem de máquina <http://www.adinkra.org/htmls/adinkra/neao.htm>
SANKOFA - Inferência Bayesiana <http://www.adinkra.org/htmls/adinkra/sank.htm>

Chap 1: Funções, Muitas formas de calcular a variância, momentos, Distribuição normal.



Capítulo 0 : Ferramentas

Programação com estatística básica

Master Foo and the Shell Tools²

Um aprendiz do caminho Unix veio ao Mestre Foo e disse: “Estou confuso. Não é o caminh Unix que cada programa deve se concentrar em uma coisa e fazê-la bem?

Mestre Foo assentiu.

O aprendiz continuou: “Também não é do caminho Unix que a roda não deve ser reinventada?

Mestre Foo assentiu novamente.

“Então, por que existem diversas ferramentas com capacidades similares em processamento de texto: sed, awk e Perl? Com qual delas posso praticar melhor o caminho Unix?”

Mestre Foo perguntou ao aprendiz: “Se você tem um arquivo de texto, qual ferramenta usaria para produzir uma cópia com algumas palavras trocadas por uma string de sua escolha?”

O aprendiz torceu o nariz e disse: “As expressões regulares de Perl seriam um excesso para tarefa tão simples. Eu não conheço awk, e venho escrevendo scripts sed nas últimas semanas. Como tenho experiência com sed, eu preferiria ele no momento. Mas se o trabalho precisa ser feito apenas uma vez, um editor de textos funcionaria.”

Mestre Foo assentiu e respondeu: “Quando você estiver com fome, coma; quando estiver com sede, beba; quando estiver cansado, durma.”

E, ao ouvir isso, o aprendiz foi iluminado.

²<http://catb.org/esr/writings/unix-koans/shell-tools.html>

Computadores

Ao longo do texto, usaremos exemplos com software. Computadores são úteis para acelerar os cálculos necessárias para nossos objetivos.

Há milênios, o homem usa instrumentos, como ábacos e tabelas, para fazer operações extensas e precisas envolvendo grandes números. Dado um problema ou dado a ser computado, esses mecanismos automatizam partes do processo devido à maneira como foram construídos. A principal diferença destas ferramentas para os computadores de hoje é que nossas máquinas podem ser programadas para fazer computações arbitrárias.

Ada Lovelace (*10 December 1815 – 27 November 1852*) foi a primeira a descobrir essa possibilidade. Estudando a Máquina Analítica de Charles Babbage, Ada concebeu uma maneira de realizar computações para as quais a máquina não havia sido desenhada originalmente. O programa concebido calculava os Números de Bernoulli. Discutivelmente, alterar a estrutura de máquinas mais simples também consiste em reprogramá-las.

Máquinas desse tempo pesavam toneladas e eram muito mais lentas. O avançar dos anos tornou a tecnologia mais acessível, ao ponto de possibilitar computadores pessoais de alta potência e baixo-custo. Além disso, ao invés de operações mecânicas complexas, podemos usar linguagens de programação que traduzem comandos baseados no inglês para instruções de máquina.

Os programas aqui apresentados são escritos em Julia. É uma linguagem com compilador em tempo real (JIT) voltada à computação estatística, possuindo ferramentas úteis em sua biblioteca de base. Entre estas, funções para gerar e manipular distribuições probabilísticas.

Sendo uma linguagem de ‘alto nível’, não temos sobrecarga cognitiva no programador com manejo de memória e hardware no código. A abstração de detalhes físicos, como registradores da CPU, são feitas automaticamente pelo interpretador. O ecossistema para visualização de dados possui poder e flexibilidade. A comunidade cresce rápido e fluência nessa linguagem dá acesso a ferramentas muito diversas com bases grandes de suporte. Há suporte para estilo funcional e orientado a objetos.

Curso rápido

Códigos são importantes ao longo dos próximos capítulos para realizar cálculos, gerar dados e visualizações.

Felizmente, os programas que escreveremos são simples, de forma que não precisamos conhecer todos os recursos e características da linguagem. O capítulo 0 apresenta instrumentos básicos (R e Julia) para caminharmos.

Veremos diversas maneiras de escrever um programa para calcular a variância σ^2 de um conjunto de medidas.

Capítulo 0 : Julia

Instalação

Julia Instruções para download e instalação podem ser encontradas em:
<https://julialang.org/> Para Linux, envolve baixar o binario/código-fonte/tarball diretamente do website.

IDE Com Julia instalado, recomendo o uso do ambiente de desenvolvimento Juno (<http://docs.junolab.org/stable/>) para obter algumas facilidades. Entre elas: atalhos, editor com highlight de sintaxe, autocompletar, renderização em tempo real de animações e plots, visualização de datasets, ambiente de desenvolvimento, logs, suporte a markup languages.³

Capítulo 0 : Ferramentas (R)

Programação com estatística básica

³Este texto é escrito em Markdown e o código-fonte pode ser encontrado em <https://github.com/fargolo/stat-learn>

R: Curso rápido

Códigos são importantes ao longo dos próximos capítulos para realizar cálculos, gerar dados e visualizações.

Felizmente, os programas que escreveremos são simples, de forma que não precisamos conhecer todos os recursos e características da linguagem R. Neste capítulo, entenderemos os instrumentos básicos para caminharmos.

Veremos diversas maneiras de escrever um programa para calcular a variância σ^2 de um conjunto de medidas.

Instalação

R Instruções para download e instalação podem ser encontradas em:

<https://cloud.r-project.org/>

Em Windows, o processo costuma consistir em clicar no executável de instalação e concordar com os prompts.

Para Linux, envolve adicionar o CRAN à lista de repositórios e baixar o pacote *r-base* ou o código-fonte/tarball diretamente do website.

Rstudio Com o R instalado, recomendo o uso do ambiente de desenvolvimento RStudio (<https://www.rstudio.com/>) para obter algumas facilidades. Entre elas: atalhos *vim*, editor com highlight de sintaxe, autocompletar, renderização em tempo real de animações e plots, visualização de datasets, ambiente de desenvolvimento, logs, suporte a markup languages, como Markdown, RMarkdown e Latex.⁴

Tipos

Primeiro, vamos conhecer as entidades básicas do R. Lidamos rotineiramente com vetores, que são células contíguas contendo dados. Os dados podem ser de tipos: lógico (verdadeiro/falso), caracteres, números inteiros, reais e complexos:

“logical”: a vector containing logical values (TRUE/FALSE) “integer”: a vector containing integer values (1,2,3,4...,23,26...)

“double”: a vector containing real values (3.14...)

“complex”: a vector containing complex values (2 +2i)

“character”: a vector containing character values (“string”)

Para saber o tipo de um objeto em R, use `typeof(objeto)`. Podemos acessar elementos de um vetor pelo seu índice, independente do tipo. Declaramos dois vetores, `character` e `double`.

⁴Este texto é escrito em Markdown e o código-fonte pode ser encontrado em <https://github.com/fargolo/stat-learn>

```

>a <- c("banana", "terracota", "pie")
>b <- c(2.2, 4.4, 5.5)
> typeof(a)
[1] "character"
> typeof(b)
[1] "double"

```

A função *combine*: `c(arg1,arg1,...)` combina argumentos em um vetor. Para nossas aplicações, vamos usar números reais (`double`) na maioria dos casos. Os tipos `integer`, `double` e `complex` fazem parte da classe dos números (*numeric*)

```

>class(b)
[1] "numeric"

```

Operadores

Além dos operadores clássicos (+,-,/,-, ...), usamos constantemente dois operadores pouco comuns: O operador “`<-`” atribui o valor da expressão a sua direita ao objeto à sua esquerda. É preferível ao operador “`=`” para evitar confusão ao passar argumentos de funções e fazer comparações lógicas.

```

>a <- 3
>a
[1] 3

```

O operador “`%>%`” da biblioteca **magrittr** fornece o resultado da expressão à sua esquerda como argumento para a expressão à sua direita. Evita aninhamento de expressões, tornando fluxos de computações mais legíveis.

As expressões a seguir são equivalentes.

```

>library(magrittr)
>result <- 3 %>% exp %>% exp
>result
[1] 528491311
>result == exp(exp(3))
[1] TRUE

```

Onde $\exp(a) = e^a$, $e \sim 2.72\dots$. A expressão “`3 %>% exp %>% exp`” equivale a “ $\exp(\exp(3))$ ”, ou e^{e^3} . Usando parênteses, partimos da última computação. Usando o pipe (`%>%`), começamos com a primeira operação. Para a maioria das abstrações, é uma boa maneira de encadear funções.

Notem que para usar um recurso da biblioteca **magrittr**, carregamos usando o comando `library(magrittr)`. Para instalar uma biblioteca do repositório oficial (CRAN), usamos o comando `install.packages("magrittr")`.

Matrizes e data frames

R possui estruturas que ajudam a manipulação de dados estruturados como os que vemos comumente em ciências.

A mais simples é a lista. Uma lista é um conjunto de objeto de quaisquer tipos. Podemos ter uma lista contendo vetores, doubles, matrizes e gráficos.

```
>mlist <- list(a = c(1,5,6,7), b = c("a","b","c","d"))
>mlist
$a
[1] 1 5 6 7
$b
[1] "a" "b" "c" "d"
>class(mlist)
[1] "list"
```

Podemos acessar estruturas internas da maioria dos objetos em R pelo nome usando o operador

```
>typeof(mlist$a)
[1] "double"
>typeof(mlist$b)
[1] "character"
```

Outro tipo útil é composto pelas matrizes, que correspondem às matrizes da matemática, podendo também conter caracteres em suas células.

```
>matrix(data=c(mlist$a, mlist$b), ncol=2)
[,1] [,2]
[1,] "1"  "a"
[2,] "5"  "b"
[3,] "6"  "c"
[4,] "7"  "d"
```

Podemos conduzir multiplicação de matrizes facilmente.

```
>mat_example <- matrix(c(.5,.25,.25,.5,0,.5,.25,.25,.5), nrow=3, byrow=TRUE)
>mat_example
[,1] [,2] [,3]
[1,] 0.50 0.25 0.25
[2,] 0.50 0.00 0.50
[3,] 0.25 0.25 0.50
>mat_example %*% c(1,0,1)
[,1]
[1,] 0.75
[2,] 1.00
[3,] 0.75
```

Por fim, data.frames são extensões das matrizes:

```
>mat_example %>% data.frame
      X1    X2    X3
1 0.50 0.25 0.25
2 0.50 0.00 0.50
3 0.25 0.25 0.50
```

Data frames são os objetos mais comumente tratados em R e seguem o formato tidy.

Cada variável corresponde a uma coluna.

Cada observação corresponde a uma linha.

Cada tipo de unidade observacional forma uma tabela.

Um exemplo visual torna as coisas mais fáceis. A seguir, temos uma variável categórica (País) e duas numéricas (Número de médicos por 1.000 habitantes em 2011 e Expectativa de vida ao nascer) em formato tidy:

Country	Doctors 2011	Life Expectancy at Birth
Aruba	NA	NA
Andorra	NA	83
Afghanistan	0.23400000	61
Angola	NA	52
Albania	1.11300000	74
Arab World	1.52685042	NA
United Arab Emirates	NA	77
Argentina	NA	76
Armenia	2.84500000	71

Figure 1: País, Número de médicos a cada 1000 habitantes em 2011 e Expectativa de vida ao nascer. “NA” corresponde a dados faltantes no R. Layout do RStudio
Fonte: WHO

Note que cada linha corresponde a apenas um país (observação) e cada coluna representa uma variável. Se queremos ver a observação 9, vamos à linha correspondente e podemos encontrar os valores: “Armenia” (País), “2.845” (Médicos/1.000 hab. em 2011) e “71” (Expectativa de vida ao nascer).

Para acessar o valor correspondente, usamos índices separados por vírgula. O

primeiro espaço é reservado às linhas selecionadas e deve ser um vetor de números (linhas selecionadas) ou vetor com valores lógico do tamanho do dataset (valores com índices TRUE serão incluídos). O segundo espaço corresponde às colunas e deve conter índices numéricos ou nomes das variáveis.

```
# primeiras 5 linhas com variaveis species e sepal.length'
>iris[1:5,c("Species",'Sepal.Length')]
  Species Sepal.Length
1   setosa      5.1
2   setosa      4.9
3   setosa      4.7
4   setosa      4.6
5   setosa      5.0
```

Gramática dos gráficos e ggplot

Uma das ferramentas de destaque no ecossistema R é a **ggplot**. Ela provê uma sintaxe bastante poderosa e flexível para plotar visualizações. O segredo está em seu design, que utiliza gramática de gráficos (**Grammar of GraphicsPlot**).

Bertin⁵ delineou essa abordagem, que consiste em mapear características dos dados a elementos visuais seguindo uma sintaxe consistente. A lib ggplot implementa uma gramática em camadas, possibilitando superposições para gráficos complexos.

```
>head(sleep)
  extra group ID
1   0.7     1  1
2  -1.6     1  2
3  -0.2     1  3
```

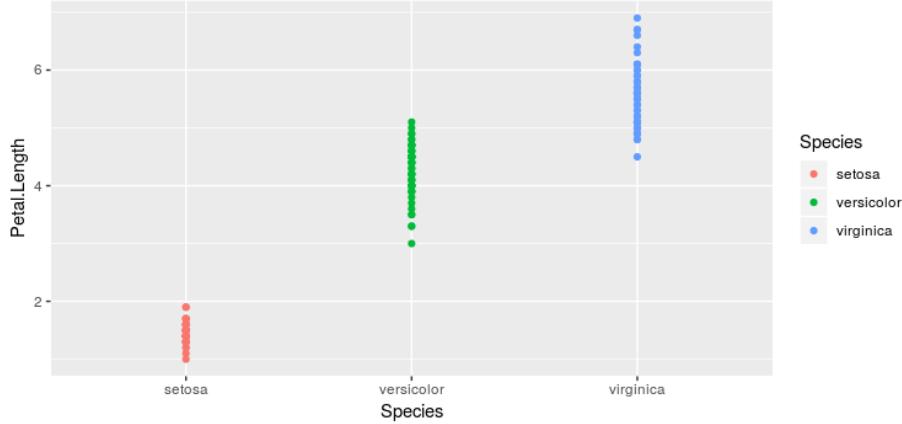
Para usarmos o ggplot, podemos declarar (1) o dataframe usado, (2) a relação entre medidas e parâmetros estéticos e (3) objetos geométricos. Parâmetros opcionais podem ser usados, aumentando o número de camadas ou criando transformações.

Assim, podemos plotar um histograma das medidas dos dois grupos com (1) dataset iris; (2) dimensão y: tamanho da pétala, cores:espécie, dimensão x: espécie; e (3) objeto geométrico: ponto.

Assim, teremos pontos com a altura (dimensão y) correspondente à medida da pétala e separados ao longo do eixo x por espécies. O ggplot automaticamente discretiza o eixo x.

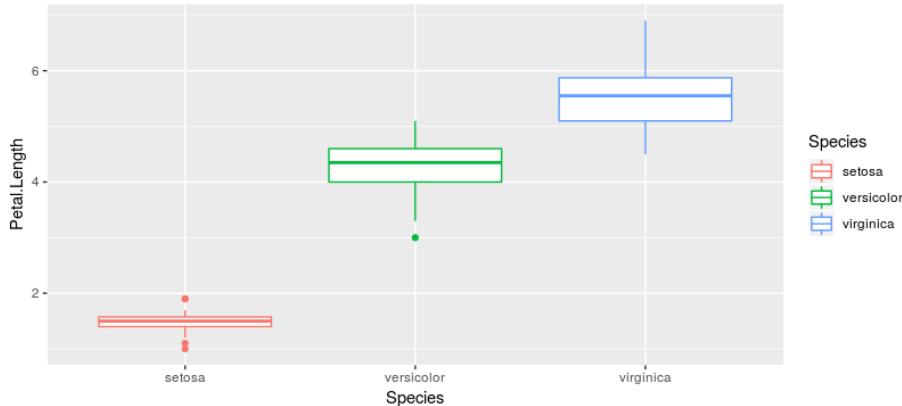
```
>library(ggplot2)
>ggplot(data=iris,aes(y=Petal.Length,x=Species,color=Species))+
  geom_point()
```

⁵Bertin, J. (1983),Semiology of Graphics, Madison, WI: University of Wisconsin Press



Para ilustrar a flexibilidade da biblioteca, note que mudando apenas o objeto geométrico (geom), obtemos um gráfico diferente, mantendo dados e relações (mappings) iguais :

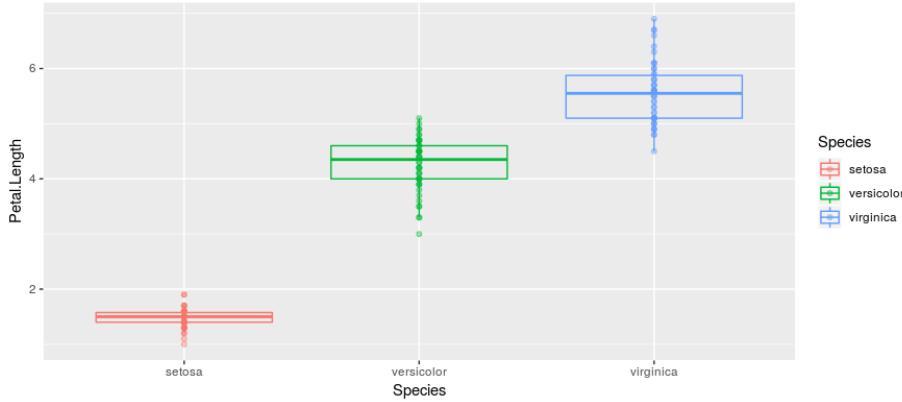
```
>ggplot(data=iris,aes(y=Petal.Length,x=Species,color=Species))+  
  geom_boxplot()
```



As figuras acima são conhecidas como boxplots. O centro correspondente à mediana (percentil 50), as bordas correspondem aos percentis 25 (inferior) e 75 (superior). Os fios, conhecidos como “bigodes”, estendem-se até $1,5 * \text{IQR}$ (onde $\text{IQR} = \text{Percentil 75} - \text{Percentil 25}$).

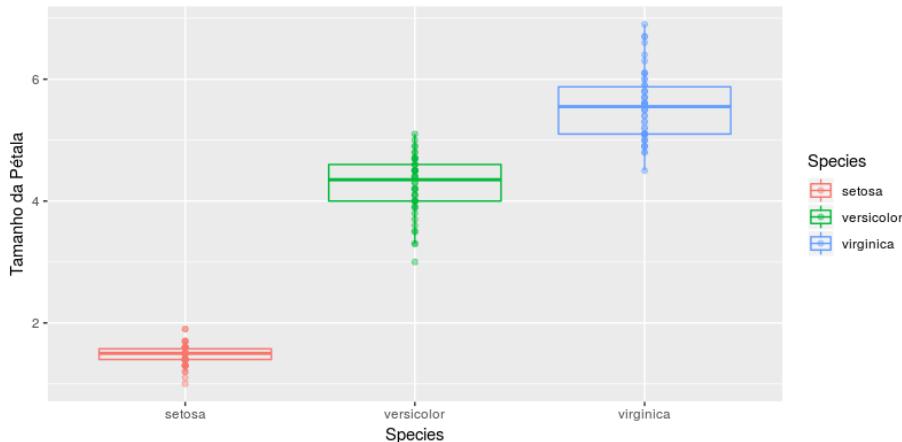
É possível adicionar camadas e estas podem sobreescrivere informação de camadas anteriores. Isso torna a sintaxe do ggplot altamente modular. A seguir, superponemos pontos e boxplot:

```
>ggplot(data=iris,aes(y=Petal.Length,x=Species,color=Species))+  
  geom_point(alpha=0.4)+ # camada 1  
  geom_boxplot(alpha=0) # camada 2
```



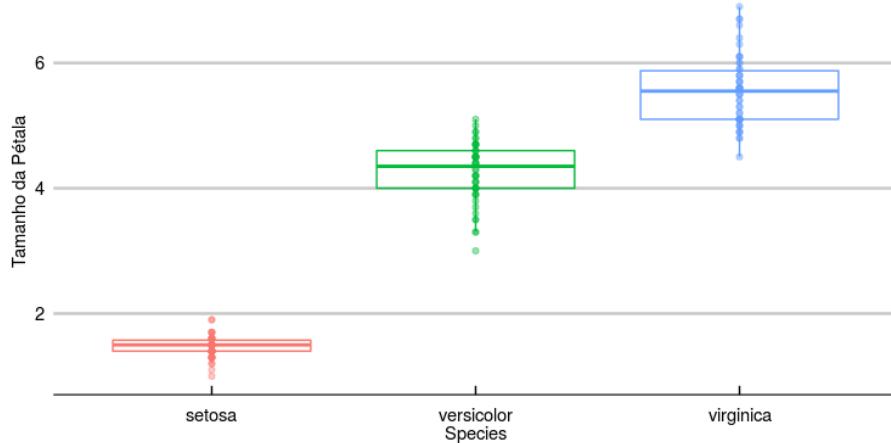
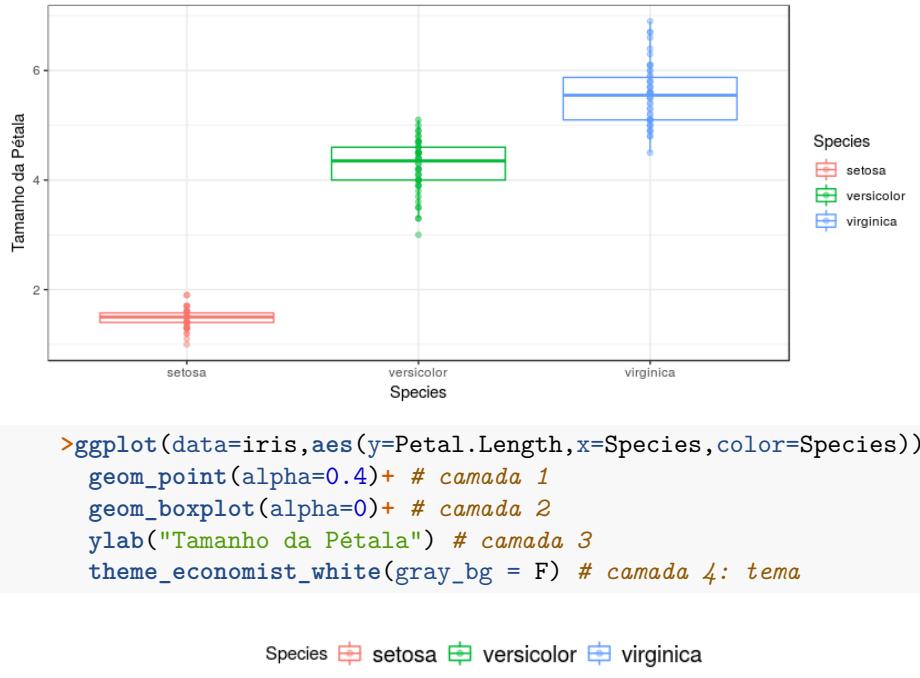
O parâmetro *alpha* regula a transparência dos objetos. Colocamos os boxplot com transparência total (*alpha=0*), dando visibilidade aos pontos (*alpha=0.4*). Adicionamos algum grau de transparência para que pontos superpostos sejam mais escuros que pontos individuais. Adicionaremos uma terceira camada, que substitui o rótulo do eixo y para uma legenda em português:

```
>ggplot(data=iris,aes(y=Petal.Length,x=Species,color=Species))+  
  geom_point(alpha=0.4)+ # camada 1  
  geom_boxplot(alpha=0)+ # camada 2  
  ylab("Tamanho da Pétala") # camada 3
```



Ainda, existem temas prontos para mudar o estilo geral da imagem:

```
>ggplot(data=iris,aes(y=Petal.Length,x=Species,color=Species))+  
  geom_point(alpha=0.4)+ # camada 1  
  geom_boxplot(alpha=0)+ # camada 2  
  ylab("Tamanho da Pétala") # camada 3  
  theme_bw() # camada 4: tema
```



Funções

Uma das formas de escrever programas é através de funções.

Podemos declarar funções que (1) aceitam argumentos de entrada, (2) executam computações com esses argumentos e (3) devolvem resultados na saída.

Assim, podemos criar a função soma2, que recebe dois argumentos numéricos e

retorna a soma de ambos.

```
>soma2 <- function(argumento1,argumento2){  
  return(argumento1+argumento2)  
}
```

Ao invocarmos soma2 com os argumentos 2 e 5, recebemos soma2(argumento1=2, argumento2=5) = $2+5 = 7$.

```
>soma2(argumento1=2,argumento2=5)  
[1] 7
```

Podemos omitir o nome dos argumentos. Assim os objetos são passados na ordem de entrada.

```
>soma2(2,3)  
[1] 5
```

Por padrão, o valor retornado é mostrado no console.

R aceita em sua sintaxe que uma função seja argumento de outra numa mesma instrução:

```
>soma2(2, soma2(3,2) )  
[1] 7
```

A expressão acima é equivalente a $(2 + (3 + 2)) = 7$.

Podemos definir a função de média para um vetor de números, dado pela (1) soma dividida pelo (2) tamanho do vetor:

```
>mean_vec <- function(x){  
  sum(x)/length(x)  
}  
>mean_vec(b) # Anteriormente definido por b <- c(2.2, 4.4, 5.5)  
[1] 4.033333
```

`sum(x)` retorna a soma de todos os elementos do vetor `x`. `length(x)` retorna o tamanho (número de células) do vetor `x`.

A média é uma medida de tendência central para um conjunto de observações. É o ponto mais perto de todos os outros.

Muitas formas de calcular a variância

Também podemos calcular uma medida relacionada ao quanto nossos valores se afastam do centro.

Primeiro, calculamos uma distância entre cada elemento `x` e a média das observações μ . A noção de distância implica que ela deve ser um valor positivo. Supondo que `x` e μ são medidas num espaço ordenado, podemos usar o módulo

da diferença entre os valores: $\|x - \mu\|$. Ainda, podemos usar o quadrado da diferença: $d_i = (x_i - \mu)^2$.

A variância σ^2 das observações é uma medida da dispersão de toda a amostra. Para calcular σ^2 , somamos todas as distâncias d_i e dividimos o resultado por $n - 1$.

```
>var_2 <- function(x) sum((x - mean(x))^2) / (length(x) - 1)
>var_2(b)
[1] 2.823333
```

Sendo proporcional às distâncias dos valores em relação à média, a variância σ^2 tende a ser maior quando os valores são muito distintos entre si:

```
>c <- c(100, 200, 1, 45, -24)
>var_2(c)
[1] 7966.3
```

Outra medida de dispersão, dada nas unidades originais da medida observada, é o desvio-padrão σ , dado pela raiz da variância σ^2 .

```
>var_2(b) %>% sqrt
[1] 1.680278
```

O R possui funções embutidas para muitas aplicações estatísticas: `sd` (desvio-padrão), `var` (variância), `mean` (média)... Em especial, temos funções prontas para trabalhar com diversas distribuições probabilísticas de variáveis aleatórias. Para sortear 10 números de uma distribuição normal:

```
>rnorm(n=10, mean=0, sd=1)
[1] 0.2874490 0.2931469 3.1897423 1.7445002 3.3998010 -0.1482911
[7] 2.0257046 -0.6002109 -0.2840376 -0.7715565
```

Distribuição gamma.

```
>rgamma(n=10, shape=1)
[1] 1.1183441 1.2770135 1.0972053 1.4820536 2.3542620 0.8231831 0.5535210
[8] 5.0481559 0.2853060 0.1623315
```

Exponencial:

```
>rexp(n=10, rate = 1)
[1] 0.31657586 0.26676766 0.02288276 0.92801416 0.44006133 0.05238540
[7] 1.10213153 0.91931786 2.58807134 0.41825081
```

Vetores, loops e recursões

Anteriormente, definimos a função para calcular variância como:

```
>var_2 <- function(x) sum((x - mean(x))^2) / (length(x) - 1)
```

Isso só é possível porque o R aplica funções a vetores de maneira automática. Assim, a expressão $(x - \text{mean}(x))^2$ subtrai a média de cada elemento do vetor x.

Normalmente, é necessário usar estruturas recursivas para isso. O laço for (for loop) define uma sequência de tamanho n definido e repete um bloco de comandos n vezes. Se queremos imprimir números entre 1 e 10:

```
>for (i in 1:10) print(i)
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
[1] 10
```

A instrução avalia print(i) para valores i=1,2,3..,10 de forma repetida.

Vamos reescrever nossa função para calcular variância σ^2 usando um loop. Podemos definir um loop com o tamanho do vetor x e calcular o quadrado da diferença em cada elemento.

Assim,

```
var_3 <- function(x){
  accumulator <- numeric() #armazena distâncias
  for (i in 1:length(x)) # loop começa em 1 segue até o tamanho do vetor
    accumulator[i] <- (x[i] - mean(x))^2 # calcula e armazena distâncias.
  return (sum(accumulator) / (length(x) - 1)) #calcula media
}
```

Ambas definições apresentam o mesmo resultado que a implementação nativa do R:

```
> var(b)
[1] 2.823333
> var_2(b)
[1] 2.823333
> var_3(b)
[1] 2.823333
```

Ainda, uma maneira de manipular muitos elementos é através de funções de alta ordem. Estas funções recebem outras funções como argumentos. Um exemplo é a função map da lib purrr. Definimos uma função para a distância, $f(y) = (y - \mu)^2$, e aplicamos em todos os elementos. Só então, somamos os resultados e dividimos

por n-1.

Tudo pode ser feito em apenas um pipe:

```
>map(.f = function(y) (y - mean(arg))^2, .x = arg) %>% # Define e aplica função  
  unlist(.) %>% sum(.)/(length(arg) - 1) # Soma as distâncias e divide por n-1
```

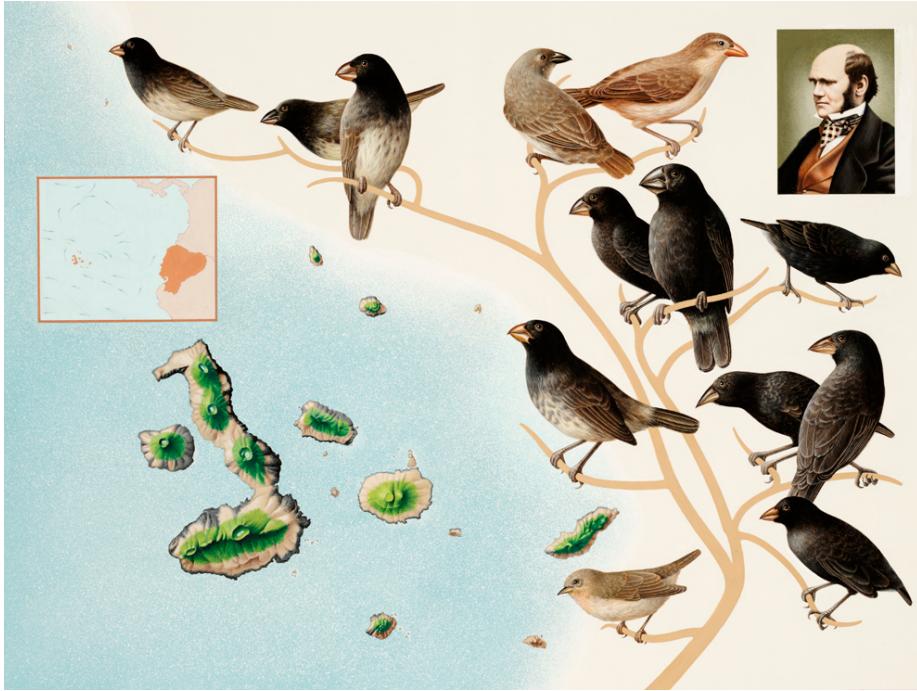
Quando usamos o pipe, o caractere de ponto (.) se refere ao valor fornecido como entrada pela pipe anterior. Assim, sum(.), no exemplo acima soma os valores passados pela função unlist(.), que por sua vez, transforma em vetor uma lista de valores passada pela função *map*. Nossa função pode ser escrita:

```
var_4 <- function(arg){  
  purrr::map(.f = function(y) (y - mean(arg))^2, .x = arg) %>%  
    unlist %>% sum(.)/(length(arg) - 1)  
}  
> var_4(b)  
[1] 2.823333
```

Exercícios

1. Qual a diferença entre linguagens compiladas e interpretadas?
2. Um programa escrito em R pode ser escrito em qualquer outra linguagem. Esta afirmação é verdadeira? Por quê?
3. Cite 3 recursos que uma IDE fornece ao programador.
4. Modifique o tema de fundo do RStudio para um de cor escura (menos luz para os olhos :)).
5. Usando o operador `<-`, produza:
 - Um vetor com componentes do tipo logical
 - Dois vetores de 5 elementos do tipo double
 - A soma dos elementos nos vetores do item b.
 - A divisão entre elementos dos vetores do item b.
 - Aplique as funções sd, mean e var em amostras normais aleatórias de n = 10, 30, 100 e 300. A função rnorm (n,mean,sd) pode ajudar. Compare os valores da distribuição de origem com os obtidos.
7. *UnLISP it!* Transforme as seguintes expressões, substituindo parênteses aninhados pelo operador pipe (`%>%`) quando julgar conveniente:
 - `round (mean (c(10 , 2, 3)))`
 - `round (mean (rnorm (n = ceiling (runif (1,0,10))))`
 - `paste("a",seq(1:max(sample(1:10))))`
 - `round(nrow(iris) + exp(1), digits = ceiling(runif(1,0,10)))`
8. Usando o código das funções var_2 (vetorizado), var_3 (for loop) e var_4 (função de alta ordem map)
 - Escreva as funções correspondentes (sd_2, sd_3, sd_4) para desvio-padrão e compare com a função padrão do R (sd). Dica: Basta aplicar raiz quadrada ao valor final retornado anteriormente!
9. Usando o dataset iris

- Selecione apenas os exemplos com tamanho de pétala maior que 4.
 - Selecione os 10 maiores exemplares. Suponha que o tamanho é dado pela média das 4 medidas fornecidas.
 - Calcule a média e o desvio-padrão para duas medidas em cada espécie.
 - Faça um scatterplot entre duas medidas
 - Adicione cores de acordo com a espécie
 - Adicione o rótulo de texto a um dos pontos
 - Mude títulos (principal, eixos x e y, legenda)
 - Mude o tema de fundo. Dica: experimente os temas da lib *ggthemes*
10. Usando loops, escreva uma função que retorna uma aproximação de e .
- Lembre-se de que $e = \lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n$.



Capítulo 1 : Os pássaros de Darwin e o método hipotético-dedutivo.

Testes estatísticos e distribuições probabilísticas

Parte 1 - Introdução

Charles Darwin observou que os pássaros tentilhões 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 em “On the Origin of Species” (1859).

Neste capítulo, simularemos dados para uma abordagem quantitativa do problema. Estudaremos medidas de bicos dos tentilhões em pequenas amostras de cada ilha e faremos inferências sobre as populações de origem (espécies diferentes).

Também estudaremos a relação natural entre a distribuição normal e a distribuição t , ligadas entre si. A adoção da distribuição normal em trabalhos científicos é bastante popular. Para entender os motivos, o Teorema do Limite Central e o conceito de entropia são fundamentais.

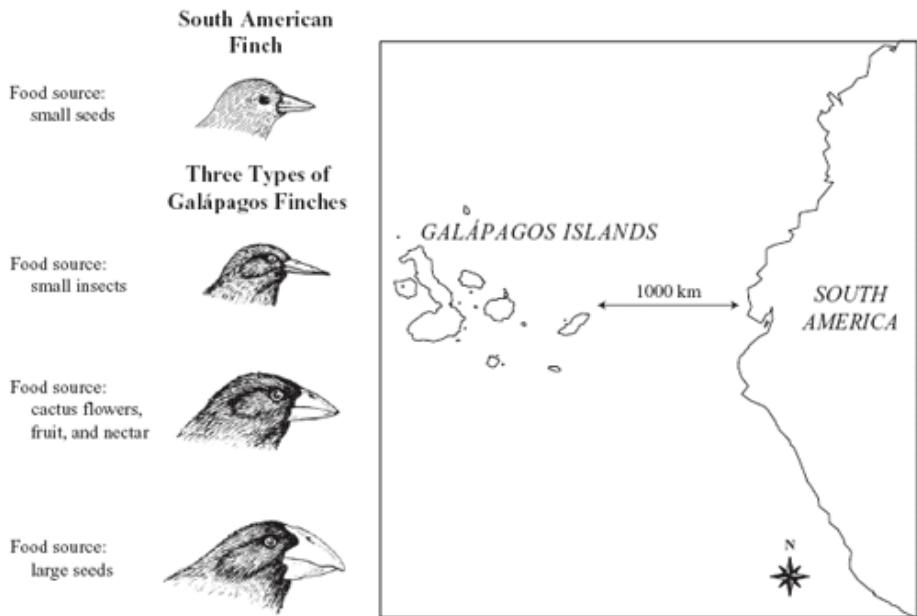


Figure 2: Fringilídeos em Galápagos.

Ilhas Galápagos

Em sua viagem a bordo do Beagle, 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: *É bastante notável que uma graduação 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.*⁶

Ele 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 semente do cacto, enquanto os de bico curto extraçalham a base do cacto e comem sua polpa.

Antes da publicação de *A origem das Espécies*, o caso dos tentilhões (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 ambientes específicos:

(...) [ao] ver esta graduação e diversidade em estrutura em um pequeno, intima-

⁶ “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.”^[^4] Tradução livre. The Voyage of the Beagle (1839).

*mente relacionado grupo de pássaros, é possível imaginar que, a partir de poucos pássaros deste arquipélago, uma espécie foi escolhida e modificadas para certos fins.*⁷

Dúvidas - Hipóteses e observações

Darwin levou aproximadamente 20 anos entre a concepção inicial da ideia (1838) e a publicação da obra (1859). Ciente de que propostas semelhantes foram ridicularizadas, ele foi metílico na defesa de sua teoria sobre a origem das espécies.

A observação dos pássaros na ilha era uma evidência, porém não confirmava a teoria. Darwin então traçou um plano de investigação para testar diversas consequências distintas da teoria.

Distribuição geográfica, variabilidade fenotípica (hibridização e fertilização cruzada), variação sob domesticação... Será que experimentos nessas áreas obedeceriam as previsões?

As duas décadas foram dedicadas a contatar e interagir com especialistas de diferentes áreas (da botânica à criação de pombos e coelhos). As evidências acumuladas falaram fortemente em favor da explicação darwiniana, que descrevia campos diferentes num modelo abrangente e simples. O trabalho de formiga consistia em explorar dados e em convencer outros cientistas a aceitarem a ideia. Isto durou até que Alfred Wallace antecipou algumas das consequências mais contundentes em 1855, as quais Darwin tinha evitado atacar diretamente. (“On the Law which has Regulated the Introduction of New Species”, Annals and Magazine of Natural History).

Charles Lyell era um geólogo, amigo de Darwin, e foi quem incentivou fortemente a publicação de uma exposição sólida da teoria. A teoria concebida 1938 para a origem das espécies poderia estar errada, ainda que as evidências do Beagle fossem promissoras. O estudo das hipóteses secundárias esclareceria a veracidade da teoria. As confirmações experimentais deram segurança para uma defesa convincente.

Probabilidades

É interessante notar que a linguagem usada para denotar diferenças é eminentemente quantitativa (no trecho acima: *dimensions, largest, differing*).

Darwin observou a adequação dos bicos à dieta através de sua intuição, sem realizar medidas.

A inspeção visual de um naturalista treinado foi capaz detectar essas nuances.

⁷ “*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.*”^[^5] Tradução livre. 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

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. O especialista (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**.

Assim como o naturalista, examinaremos as diferenças para grupos distintos. Porém, usaremos estatística e probabilidades (distribuição normal e Student's t) para testar hipóteses e fazer conclusões mais acuradas sobre as medidas.

Falseabilidade e hipóteses

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, científicidade comunica credibilidade. Nas últimas décadas, filósofos discutiram a validade do problema de demarcar ciência de pseudociência e não-ciência.⁸ 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 teorias, 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 faz sol 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 faz sol amanhã” é uma hipótese verdadeira independente das observações.

Theophrastus(c. 371–c. 287 AC) e Eudemus(400 AC) originalmente delinearam o *modus tollens*.

- 1 . Se a teoria é verdadeira, o fato X acontecerá,
- 2 . X não aconteceu, *logo* a teoria é falsa.

K. Popper foi um líder da revitalização do método hipotético dedutivo no século passado. Para ele, a dificuldade em gerar hipóteses testáveis e falseáveis sinalizava uma evidente fragilidade nas teorias. A liberdade para testar a veracidade de teorias é uma característica marcante da ciência em oposição a práticas esotéricas e/ou baseadas em autoridade.

Popper atacou severamente o materialismo dialético de Karl Marx, assim como a teoria da evolução por seleção natural de Charles Darwin e a psicanálise.

Estes ramos do conhecimento humano encontraram dificuldades com o critério de demarcação proposto. Marx previu que a revolução aconteceria em uma nação industrializada, através da classe operária e outros eventos que não se concretizaram. Seus seguidores usaram hipóteses *ad-hoc* para justificar as observações mantendo as previsões feitas à luz do materialismo dialético.

A teoria da evolução por seleção natural de Darwin era amparada em muitos exemplos de reprodução impossível (e.g. recomposição da trajetória evolutiva em

⁸Massimo Pigliucci - Philosophy of Pseudoscience: Reconsidering the Demarcation Problem

fósseis). A psicanálise também sofreu duras críticas, em virtude da irrefutabilidade de seus pilares centrais.

Como discutiremos nos próximos capítulos, hipóteses não são essenciais na vida do cientista. Entretanto, previsões falseáveis são extremamente úteis para evidenciar a utilidade de uma teoria. Um especialista tem sua credibilidade extremamente aumentada quando costuma acertar palpites em situações incertas e o mesmo vale para teorias científicas.

Testes de hipóteses podem ser formalizados através de probabilidades e estatística, que incorporam aspectos quantitativos. Calculamos a probabilidade associada a observações, considerando o cenário de uma hipótese (falseável).

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

Testes de hipótese Costumamos partir de uma hipótese base, chamada hipótese nula, que descreve o cenário menos interessante, isto é, a inexistência dos fenômenos propostos pelo cientista.

É comum comparar dois grupos, A e B, quanto ao resultado de uma intervenção. A hipótese nula costuma assumir que os grupos apresentam resultados iguais.

Queremos estudar o tamanho dos bicos de pássaros das ilhas A e B. A hipótese nula natural assume que as espécies são iguais: 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 vêm de populações iguais.

Se as diferenças forem grandes, a probabilidade é muito baixa. Rejeitamos nossa hipótese e aceitamos a alternativa (há diferença entre A e B).

Estruturando os passos:

1. Definimos a hipótese nula (H_0) e pelo menos uma hipótese alternativa(H_1).
 - H_0 : Pássaros das ilhas A e B possuem bicos de tamanho igual.
 - H_1 : 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 H_0 verdade, nossos resultados seriam raros ou comuns?

Caso p seja menor que um limiar pré-definido (convencionalmente, 0.05), rejeitamos H_0 . A probabilidade é muito pequena para H_0 ser verdade.

A domínio dos procedimentos hipotético-dedutivos nas ciências produziu resultados interessantes. Especialmente no eixo de trabalho denominado por Thomas Kuhn de “ciência normal”, focada no acúmulo de evidências e testagem de

hipóteses. O ideal de desenhar um experimento imparcial, com possibilidade de fracasso, aguçou a percepção de pesquisadores para a falibilidade de ideias. O grau de sofisticação em reproduzibilidade de procedimentos foi amplificada.

Nota *Usamos o limite inferior de 0.05 como critério para rejeitar a hipótese nula, o que pode parecer arbitrário. E é. Os valores p eram interpretados de acordo com sua magnitude e estatística com base em que foram calculados. Foi Ronald Fisher, em *Statistical Methods for Research Workers* (1925), quem propôs (e posteriormente popularizou) o número: “The value for which $p = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not.⁹”*

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 secundários. 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 ($p(\text{cara}) = p(\text{coroa}) = 0.5$), aproximam-se de uma curva suave (contínua) à medida em que a quantidade de eventos aumenta.

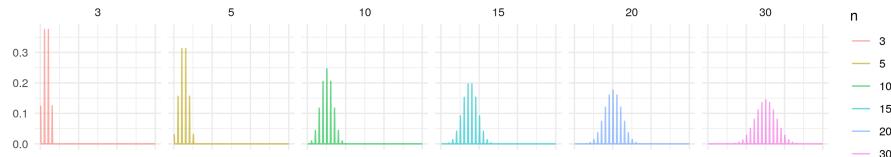


Figure 3: Distribuições binomiais para diferentes números de lançamentos com $p = 0.5$ (e.g: lançamento de uma moeda). Para $n > 1$, valores extremos indicam resultados com apenas caras (cauda à esquerda, 0000...) ou coroas (cauda à direita, 1111...)

A distribuição de Bernoulli descreve a possibilidade de dois eventos, como o lançamento de moedas. Tomando os valores discretos de caras (0) e coroas (1), a observação é 1 com probabilidade p e 0 caso contrário ($1 - p$). Para uma moeda honesta, temos uma distribuição probabilística uniforme sobre o domínio, $X = 0, 1$: $P(1) = P(0) = 0.5$.

⁹O valor [da estatística z em uma curva normal] para o qual $p = 0.05$, ou 1 em 20, é de 1.96 ou aproximadamente 2; é conveniente pegar esse ponto como um limite ao julgar quando um desvio deve ser considerado significante ou não.

Se somarmos distribuições de Bernoulli, obtemos a distribuição binomial. Cada observação é um conjunto de lançamentos. Tomando $p = 0.5$, resultados mais frequentes são números parecidos de caras (0) e coroas (1).

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 com que a distribuição se aproxime da seguinte curva contínua:

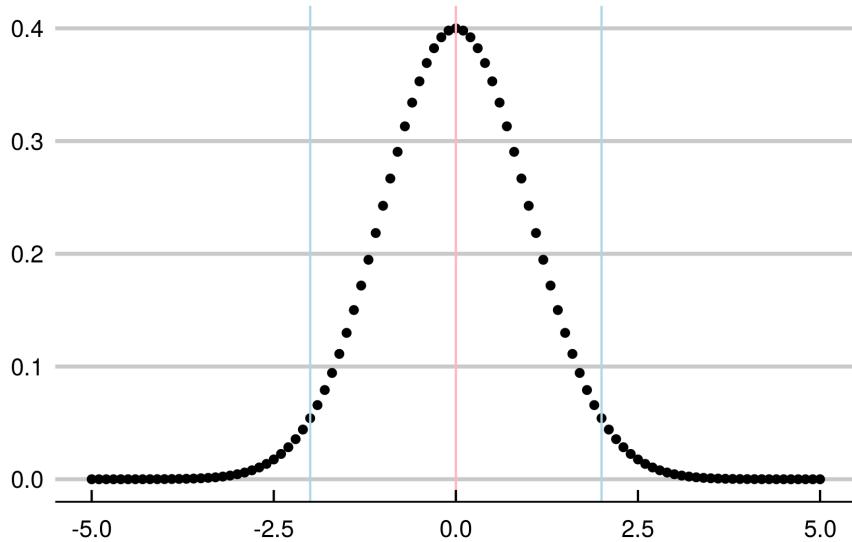


Figure 4: Distribuição normal (gaussiana), cujo formato lembra o de um sino

De Moivre intuiu que a distribuição de binomiais com muitos lançamentos aproximava o de uma função suave. Ele buscava uma aproximação em termos da função exponencial [natural] e^x .

Mas quais os parâmetros da curva?

Primeiro, de Moivre deduziu a solução para o problema das moedas ($p = \frac{1}{2}$). A seguinte expressão geral descreve a probabilidade $P(x)$ correspondente à curva que procuramos, conhecida como *gaussiana*.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Em que e é número de Euler ($e \sim 2.72\dots$).

O valor $\frac{1}{\sqrt{2\pi}}$ surge como normalizador para avaliarmos a função como densidade de probabilidade (A integral de $-\infty$ a $+\infty$ deve ser 1). O valor π surge da integral de Gauss para e^{-x^2} e decorre do fato de $2\pi i$ ser período da função e^x :

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

Intuições A definição possui um fator constante $\frac{1}{\sqrt{2\pi}}$ (aproximadamente 0.4), multiplicando o resultado da exponencial no formato e^{-x} . Em Julia, podemos definir a função e observar a probabilidade associada a alguns pontos em torno do máximo ($f(0) = 0.4$):

```
mgauss(x) = 0.4*exp((-1)*(x^2)/2)
mgauss(-2), mgauss(-1), mgauss(0) , mgauss(1) , mgauss(2)
(0.054134113294645084, 0.2426122638850534,
 0.4, 0.2426122638850534, 0.054134113294645084)
```

Em seguida, obter alguns valores no intervalo $[-5, 5]$ e plotá-los, dando origem à curva gaussiana anterior.

```
using Plots, Distributions
gauss_values = map(mgauss, -5:0.1:5)
plot(gauss_values, xaxis=("Gaussian"), leg=false)
```

Observamos como a distribuição se dá a partir da equação.

Já que x^2 retorna apenas valores positivos, $-x^2$ sempre retorna negativos. A função retorna valores entre 0 e 1 exponenciando ($e \sim 2.718\dots$) a um fator negativo quadrático ($y \sim 0.4 * e^{-x^2/2}$).

Notamos também que valores próximos ao centro ($x \sim \mu = 0$) fazem com que o expoente de se aproxime de 0, maximizando nossa função: $f(0) = 0.4 * e^{-x^2/2} = 0.4 * e^0 = 0.4$). O valor obtido (0.4) corresponde ao topo da curva no gráfico acima (linha rosa).

Observamos a curva se aproximar simetricamente do máximo em valores próximos de 0.

Isso reflete diretamente o fato de que valores próximos à média serão mais prováveis e valores extremos menos prováveis. A rigor, a probabilidade para qualquer valor dentre os infinitos possíveis é zero.

É possível avaliar a probabilidade de evento um relacionado ao intervalo entre os pontos a e b pela integral de $f(x)$ sobre o intervalo $[a, b]$:

$$P(A_{a,b}) = \int_a^b f(x)dx$$

Por exemplo, um evento (A) relacionado a ‘valores menores ou iguais a zero’ em uma escala estão no intervalo $[-\infty, 0]$:

$$P(A) = \int_{-\infty}^0 f(x)dx$$

O termo quadrático torna a distribuição simétrica para valores opostos em relação à média. $P(A) = P(-A)$. Como calculamos $f(2)$ antes, sabemos que: $f(-2) = f(2) = 0.05$ para $\mu = 0$. 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 ($\mu < 0$) ou para a direita ($\mu > 0$), subtraindo o termo de x em nosso expoente. Além disso, diferentes variâncias (σ^2) refletem a frequência de valores longe da média e o quanto distante dela eles são. Visualmente, determina o tamanho da base do sino na ilustração (Figura 3).

Usamos a notação $N \sim (\mu, \sigma^2)$ para descrever uma distribuição gaussiana com média μ e variância σ^2 arbitrárias:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Por que usamos a distribuição normal?

Distribuições binomiais grandes e moedas são tão importantes? Os lançamentos são exemplo de uma classe maior de fenômenos. Cada série de resultados é composta por muitos eventos quase idênticos (lançamentos individuais).

Entropia Em ciências naturais, raramente conhecemos de antemão os mecanismos pelos quais as observações são geradas. Consequentemente, não sabemos a distribuição probabilística que elas obedecem. Um dado justo tem probabilidades equivalentes entre os valores possíveis (distribuição uniforme). Um dado viciado tende a cair com mais frequência em determinados valores (picos e vales).

Como veremos no capítulo a seguir, podemos descrever uma distribuição através das relações entre os valores possíveis e o centro. São os *momentos*. O primeiro momento reflete a posição relativa do centro (média), enquanto o segundo reflete a dispersão dos valores (variância). Conhecer o mecanismo natural de origem permitiria especificar distribuições diretamente, entretanto precisamos enfrentar as limitações do mundo real.

Se sabemos apenas o centro (média) e a dispersão (variância) de uma distribuição, qual o palpite mais conservador possível?

Considerando números reais (domínio em $[-\infty, +\infty]$), a distribuição normal é aquela com máxima entropia em relação às outras. A grosso modo, isso quer dizer

que é a descrição usando menos informação quando comparada com qualquer outra distribuição obedecendo essas restrições (média e variância definidas).

A Gaussiana é aquela que introduz menos informação extra em relação às possíveis distribuições verdadeiras. Pelo **Princípio da Máxima Entropia**, é a que melhor descreve observações *a priori*, quando apenas temos idéia da média e da variância. Essa é uma justificativa razoável para adotarmos gaussianas como ferramentas. A prova é razoavelmente complexa, envolvendo cálculo de variações para otimizar a expressão:

$$H(x) = - \int_{-\infty}^{+\infty} p(x) - \log p(x) dx$$

O Teorema do Limite Central Outra ligação entre a distribuição normal e as ciências naturais se dá pelo teorema do limite central.

Se somarmos muitas distribuições de uma mesma família, a distribuição resultante se aproxima de uma normal. Sem muitas explicações, assumimos que isso era verdade para moedas.

Exemplos ajudam a ganhar intuição. Ao lançar um dado justo de 6 faces, temos probabilidade de $\frac{1}{6}$ 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: $X \sim U_{discr}(1, 6)$.

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 Julia a seguir gera os dados e as visualizações de que precisamos:

```
using Distributions, Plots, StatsPlots
dice_fun(x) = rand(DiscreteUniform(1,6),x)
data_mat = [dice_fun(100) for _ in 1:11]
p1 = map(x->histogram(x,bins=6,legend=false),data_mat)
p2 = histogram(sum(data_mat,dims=1),bins=30,legend=false)
plot(p1...,p2,layout=(4,3))
```

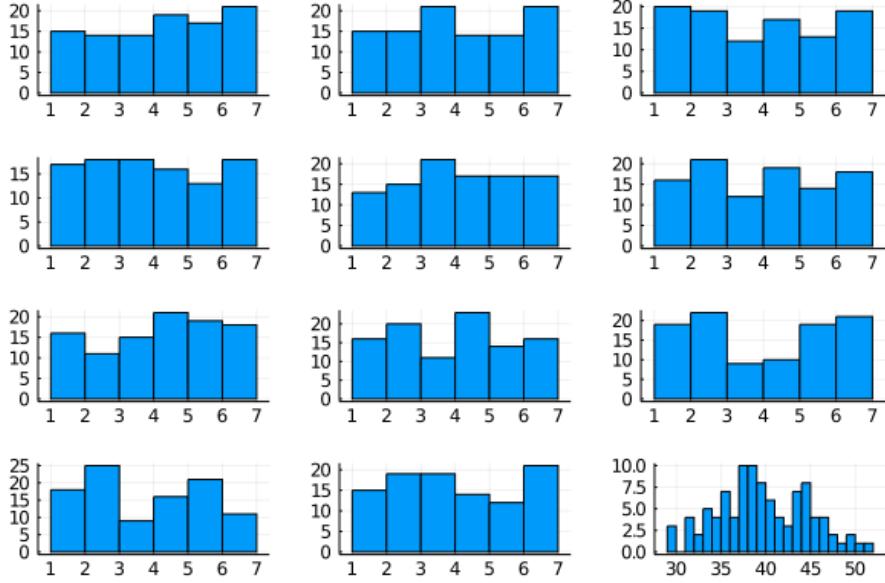


Figure 5: Soma de amostras ($n=100$) de 11 distribuições uniformes correspondentes ao lançamento de dados honestos de 6 faces. O resultado está na célula inferior à direita.

Notamos que as barras estão distribuídas com alturas bastante parecidas nas 11 primeiras células. A frequência esperada para cada valor é $\sim 1/6$ do total de 100 lançamentos. $Freq(X_i) \sim \frac{1}{6} * 100 \sim 16.66$ 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: $E(X) = \sum_{i=1}^{11} E(U_i \sim (1, 6)) = 11 * 3.5 = 38.5$

O valor 38.5 corresponde aproximadamente ao centro da distribuição resultante (Figura 2, canto inferior direito) É notável a semelhança com a curva normal, com valores extremos menos frequentes e simetricamente afastados da média (valor esperado), que define o valor máximo.

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. A prova formal pode ser consultada em outro local,¹⁰ mas voltaremos a ela. Este resultado tem uma util importância para o estudo dos fenômenos naturais através de experimentos.

¹⁰Yuval Filmus. 2010. Two Proofs of the Central Limit Theorem <http://www.cs.toronto.edu/~yuvalf/CLT.pdf>. Ela se dá mostrando a convergência de momentos entre a soma e gaussiana, um conceito que entenderemos no capítulo a seguir.

Muitos objetos de interesse para os cientistas são manifestações de fenômenos envolvendo múltiplos elementos. 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 de 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.

$$X \sim U_1(1, 6) + U_2(1, 6) + \dots + U_{11}(1, 6) = X \sim N(38.5, \sigma^2)$$

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

$$X \sim \gamma_1(\alpha, \beta) + \dots + \gamma_n(\alpha, \beta) = X \sim N(\mu', \sigma')$$

Para valores grandes de n:

```
gamma_fun(x) = rand(Gamma(1),x)
data_mat = [gamma_fun(100) for _ in 1:11]
data_mat <- cbind(data_mat, rowSums(data_mat))
append!(data_mat, sum(data_mat, dims=1))
p1 = map(x->histogram(x, legend=false), data_mat)
```

```
p2 = histogram(sum(data_mat,dims=1),bins=30,legend=false)
plot(p1...,p2,layout=(4,3))
```

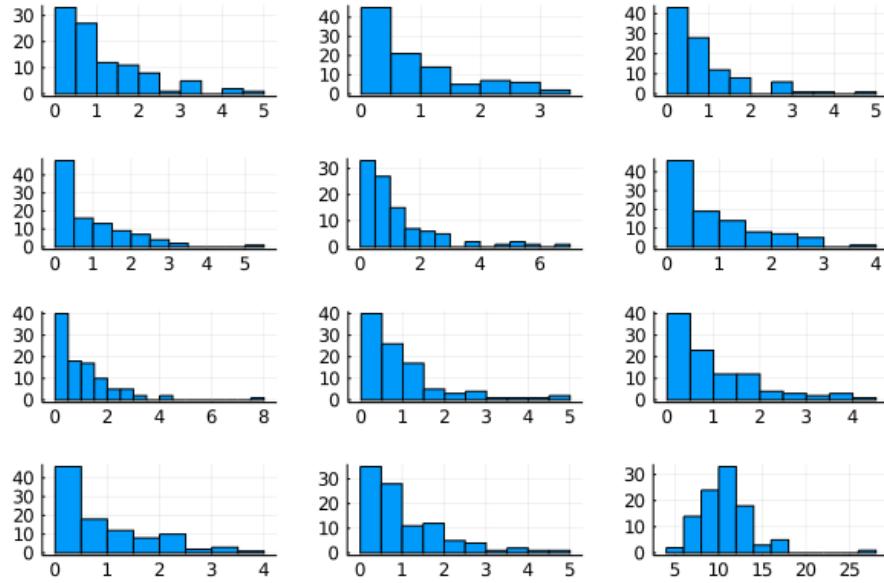


Figure 6: Soma de amostras ($n=100$) de 11 distribuições gama. O resultado está na célula inferior à direita. Função de densidade de probabilidade para distribuição gama: $f(x) = 1/\Gamma(\alpha) * \beta^\alpha * x^{\alpha-1} * e^{-\beta x}$, com $\alpha = \beta = 1$

Novamente, verificamos que a soma começa a ser simétrica em torno da média, com formato de sinos (base alargada). Muitos fenômenos observáveis em nosso universo são naturalmente compostos por múltiplos elementos semelhantes. Especialmente em sistemas biológicos, há redundância de componentes e um objeto de interesse para cientistas é resultado da combinação de muitas variáveis subjacentes.

Exercícios

1. Sobre a distribuição normal para uma variável aleatória, é verdadeiro (mais de uma possibilidade):
 - a. A soma das probabilidades de todos os valores possíveis é 1.
 - i. $\int_{-\infty}^{+\infty} f(x)dx = 1$.
 - b. É simétrica em relação à moda.
 - c. O valor esperado é dado por $1/\sigma\sqrt{2\pi}$.
 - d. 95% dos valores estão próximos à média.
 - e. Valores extremos são improváveis.
 - f. É unicamente determinada por variância σ^2 e média μ .
 - g. É contínua e diferenciável.
 - h. Amostras pequenas resultam em distribuições t.
2. Usando o comando “?Distributions” acesse algumas distribuições disponíveis na biblioteca de base do R.
 - a. Plote o histograma da soma de 100 distribuições X^2 (função rchisq; use n = 60).
 - b. Faça o mesmo procedimento para 100 distribuições de outra família e tamanho à sua escolha.
 - c. Obtenha os valores de skewness e kurtosis para essas distribuições. Uma distribuição normal padrão ($\sigma^2 = 1; \mu = 0$) possui skewness (assimetria) de 0 e kurtosis (frequência de valores mais extremos) de 3. Quais os encontrados por você?
 - d. Cite dois fenômenos naturais cuja distribuição estatística é conhecida e qual a distribuição correspondente.

Parte 2 - Darwins's Finches e um teste paramétrico

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, $X \sim U(0, 1)$. 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:

```
using Distributions , DataFrames , Random
Random.seed!(5)
n_birds = 150 # sample_size
genes_low = 80 # lower bound on number of genes
genes_hi = 100 # upper bound on number
n_islands = 4 #samples
function unif_sum(n_genes)
    gene_samples = [rand(Uniform(0,1),100) for _ in 1:n_genes]
    effects_sums = sum(gene_samples)
    return effects_sums
end
function generate_pop(;n_pop,n_genes)
    population = [unif_sum(n_genes) |> mean for _ in 1:n_pop]
end
galapagos_birds = map(x -> generate_pop(n_pop=n_birds, n_genes=x) ,
rand( DiscreteUniform(genes_low,genes_hi), n_islands)) |> DataFrame
```

Como esperado, verificamos que o histograma das medidas finais se aproximam de uma gaussiana.

```
using StatsPlots
@df stack(galapagos_birds) groupedhist(:value, group = :variable,
    bar_position = :dodge,bins=50,title=("Darwin Finches"),
    xlabel="Break Size",ylabel="Count",legend=false)
```

Os números aleatórios gerados usando a semente sugerida (`Random.seed!(5)`, linha 4 do código acima) são semelhantes à suposição de Darwin: 4 ilhas (amostras) e três espécies (distribuições de bicos). Notamos que há duas amostras (roxo, vermelho) de medidas bastante parecidas e outras duas separadas (verde, azul). Supondo que medimos os bicos de algumas aves, como saber se os grupos

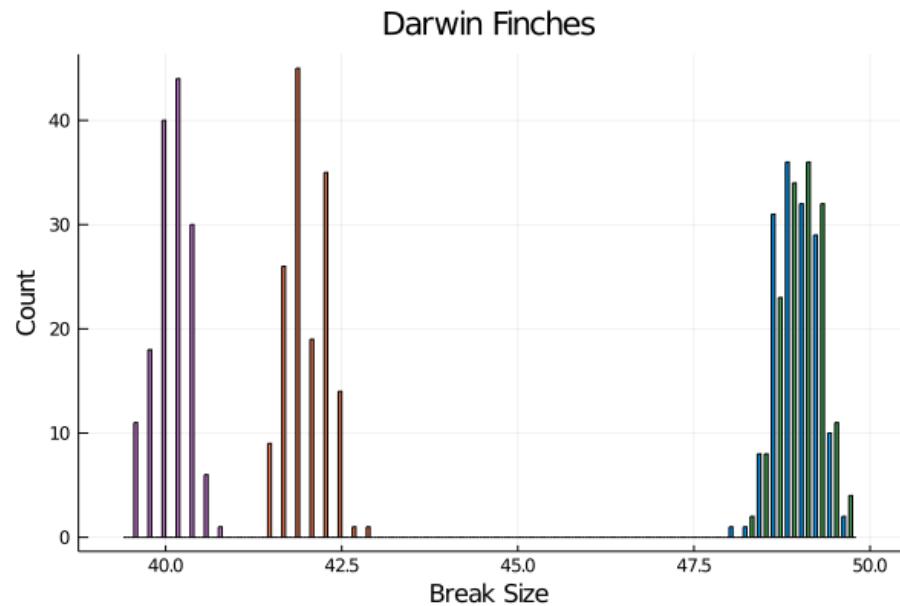


Figure 7: Figura 4. Distribuição das medidas de bicos em populações simuladas para genes com efeito aditivo.

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.

Teste t e distribuição t de Student: Um exemplo prático

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 $\mu' = 38$ mm e desvio-padrão $\sigma' = 0.3$ mm. **Problema:** Supondo que a média real (μ) da população é de 40 mm, qual é a probabilidade de obtermos $\mu' = 38$ mm em uma amostra aleatória, como aconteceu em nosso experimento?

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 Guinness, publicou na Biometrika (1908) o famoso artigo *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 com média μ e desvio-padrão σ . Podemos retirar j amostras de tamanho n e calcular a média dessas amostras $\mu'_1, \mu'_2, \dots, \mu'_j$. As médias amostrais μ' são estimativas da média real μ .

Qual a dispersão das estimativas $\mu'_1, \mu'_2, \dots, \mu'_j$?

Para um conjunto de estimativas $\mu'_1, \mu'_2, \dots, \mu'_j$, chamamos de **erro padrão (standard error of the mean)** o desvio-padrão populacional σ dividido pela raiz quadrada do tamanho da família de amostras em questão ($std.err. = \sigma/\sqrt{n}$). Como não sabemos o desvio-padrão na população, aproximamos usando o valor do desvio-padrão σ' amostral.

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

$$t = \frac{Z}{s} = (\mu' - \mu) / \frac{\sigma}{\sqrt{n}}$$

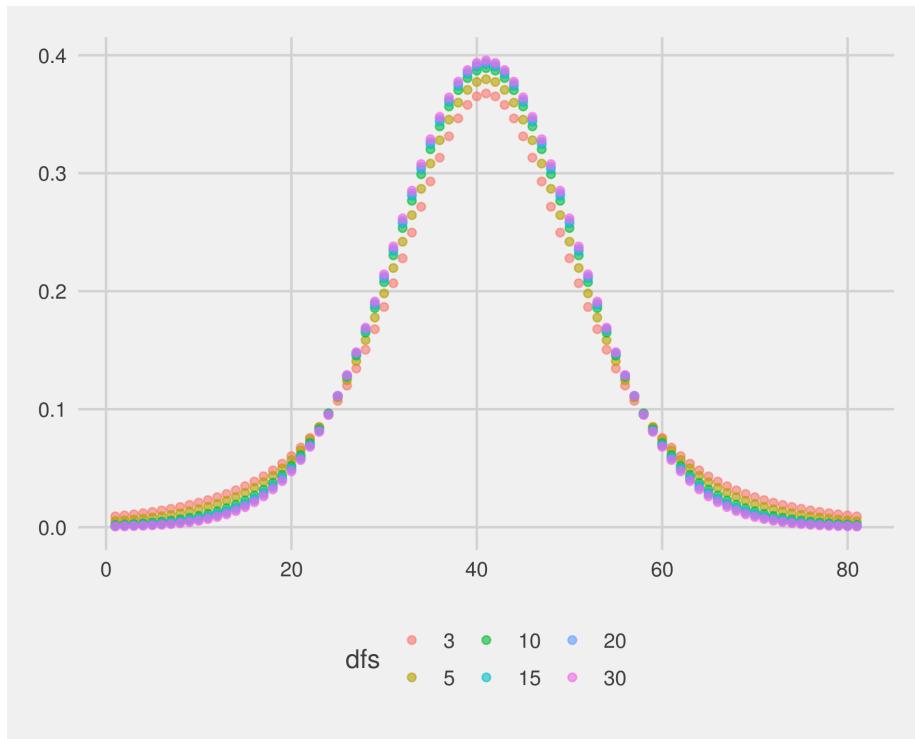
Assim, a estatística t para nosso exemplo ($\mu'=38$; $\mu= 40$; $n=30$; $\sigma'=0.3$) é:

$$t = \frac{(38 - 40)}{\frac{0.3}{\sqrt{30}}}$$

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

$$f(t) = \frac{1}{\sqrt{\nu} B(\frac{1}{2}, \frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

B é a função Beta¹¹ e v são graus de liberdade. Possui densidade parecida com a da distribuição normal, porém com probabilidades maiores para valores extremos. O parâmetro ν (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, $n \rightarrow \infty$) a uma distribuição t com $n - 1$ graus de liberdade. Em nosso exemplo, $n = 30$, então $\nu = n - 1 = 29$.



Maiores valores correspondem a amostras maiores e fazem com que a distribuição t se aproxime de uma distribuição normal. Em um caso extremo, temos $n_{samples} = n_{pop}$ e as amostras são idênticas à distribuição de origem.

Sabendo a estatística t (-36.51) e os graus de liberdade para nossa família de amostras ($\nu = 29$), podemos usar a expressão $f(t)$ para saber a probabilidade de obtermos nossa média 38 mm numa amostra ($n = 30$) se a média populacional for de 40 mm.

¹¹A função Beta é aceita dois argumentos(x, y) e seu resultado é a razão é entre (1) produto das funções $\Gamma(x)\Gamma(y)$ e (2) função gama da soma $\Gamma(x+y)$. A função Γ generaliza o conceito de fatoriais (produto dos antecessores).

Para tanto, somamos as probabilidades de valores extremos menores que a estatística t fornecida.

$$\int_{-\infty}^{-36.51} f(t)dt$$

Em julia, a função nativa *cdf* faz o trabalho sujo de calcular a integral:

```
using Distributions
cdf(TDist(29), -36.51)
4.262182718504655e-2
```

Esse valor reflete a probabilidade de valores t negativos mais extremos (menores) que os nossos ($t < -36.51$).

Teste bicaudal Parece ser nosso valor p, porém precisa de um ajuste: queremos saber a probabilidade associada a obter valores tão extremos em geral, não nos restringindo a valores extremamente menores.

Uma vez que a distribuição é simétrica, a cauda à esquerda (negativos) é idêntica à cauda à direita (positivos). Valores extremos (negativos ou positivos) em relação à média são duas vezes mais prováveis que valores negativamente extremos. Consideramos significativos valores t muito maiores (direita) ou menores (esquerda) que a média. Então, nosso limiar deve ser robusto à possibilidade de extremos maiores que a estatística t simétrica positiva.

O valor $t = 36.51$ seria a estatística resultante de uma amostra com média simétrica (42 mm) em relação à média (40 mm). Recorde-se de que a medida original foi 38 mm.

$(t_{min} = -36.51; t_{max} = 36.51)$. Ao fazer esse ajuste, chamamos o teste de bicaudal.

Sabendo da simetria na distribuição t, podemos fazer então usar o seguinte truque:

```
2*cdf(TDist(29), -36.51)
8.52436543700931e-26
```

Não é possível calcular diretamente as probabilidades para $t = 36.51$, pois Julia aproxima a integral acima ($p \sim 1 - 4.262^{-26} \sim 1$).

```
cdf(TDist(29), 36.51)
1.0
```

Nota *Uma percepção errônea comum sobre a distribuição t é de que ela descreve amostras pequenas retiradas de uma população com distribuição 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.¹² São usadas duas amostras (levo e dextro hidrobromido de hyoscyamina).

Additional hours' sleep gained by the use of hyoscyamine hydrobromide.

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2-1)
1.	+ .7	+ 1.9	+ 1.2
2.	- 1.6	+ .8	+ 2.4
3.	- .2	+ 1.1	+ 1.3
4.	- 1.2	+ .1	+ 1.3
5.	- 1	- .1	0
6.	+ 3.4	+ 4.4	+ 1.0
7.	+ 3.7	+ 5.5	+ 1.8
8.	+ .8	+ 1.6	+ .8
9.	0	+ 4.6	+ 4.6
10.	+ 2.0	+ 3.4	+ 1.4
Mean	+ .75	Mean + 2.33	Mean + 1.58
S. D.	1.70	S. D. 1.90	S. D. 1.17

Figure 8: Retirado de The probable error of a mean, pag. 20. Os 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 é idêntico ao que realizamos com a medida dos bicos e é chamado teste t de amostra única (*one sample t-test*). Hipotetizando um valor para a média (e.g. $\mu_{bico} = 40mm$; $\mu_{sono\,adiconal} = 0\,horas$), 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 ($\mu_a - \mu_b = 0$) e calculamos a probabilidade de nossa estimativa. Exemplo prático: existe diferença de peso entre os bicos dos pássaros A e B?

¹²https://atmos.washington.edu/~robwood/teaching/451/student_in_biometrika_vol6_no1.pdf

Aplicações

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.
 - H_0 : Pássaros das ilhas A e B possuem bicos de tamanho igual.
 - $\mu_a - \mu_b = 0$
- b. H_1 : 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 ($\mu_a = \mu_b$)? 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 H_0 . A probabilidade de observarmos os dados é pequena se H_0 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).

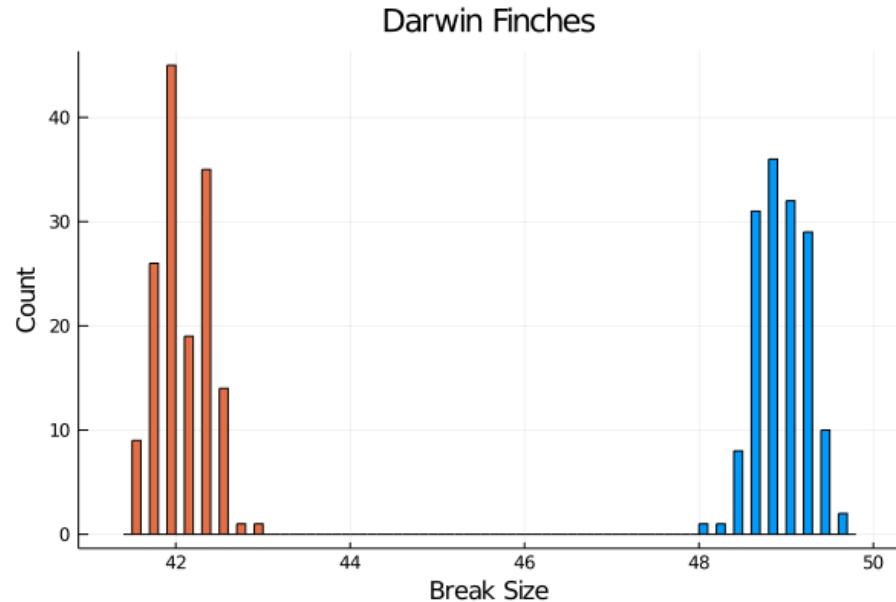
Teste t de Student com Julia Vamos computar um teste t para 2 amostras independentes. A estatística t é calculada com algumas mudanças. Os graus de liberdade são somados e o erro padrão (dispersão das estimativas) é balanceado através da média ponderada (pelos graus de liberdade, n-1) entre amostras.

$$t = \frac{X_1 - X_2}{\sigma_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\sigma_{pooled} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Considerando $(n_1 - 1) + (n_2 - 1)$ 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.

```
@df stack(galapagos_birds[:, [:x1, :x2]]) groupedhist(:value, group = :variable,
  bar_position = :dodge, bins=50, title="Darwin Finches",
  xlabel="Break Size", ylabel="Count", legend=false)
```



```
# Ajustes nos dados
a = galapagos_birds[:, :x2]
b = galapagos_birds[:, :x4]
sd_a = std(a)
sd_b = std(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 ($diff_{H_0} = 0$), (2) estimativa da diferença ($diff = \mu_A - \mu_B$), graus de liberdade (df) e erro padrão balanceado (se_{pooled}) para a distribuição das diferenças de médias.

```
expected_diff = 0
mean_diff = mean(a) - mean(b)
6.963886183171148
# graus de liberdade balanceados
df_pool = length(a) + length(b) - 2
# desvio padrao balanceado
sd_pool = sqrt((length(a) - 1) * sd_a^2 + (length(b) - 1) * sd_b^2)/df_pool)
```

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

```
# Diferenca dividida por erro padrao
# t-statistic
t = (mean_diff - expected_diff)/ (sd_pool * sqrt(1/length(a) + 1/length(b)))
```

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

```
p = 2*cdf(TDist(df_pool),-abs(t))
```

Finalmente, agregando o sumário dos resultados (médias A e B, diferença verificada, estatística t resultante, valor p):

```
results = Dict(
    "Mean Difference" => mean_diff,
    "t"=>t, "p value" => p,
    "Mean in A" => mean(a), "Mean in B" => mean(b))
Dict{String,Float64} with 5 entries:
"Mean Difference" => 0.46412
"t" => 16.7569
"p value" => 7.28163e-45
"Mean in A" => 42.9969
"Mean in B" => 42.5328
```

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. Podemos automatizar o processo em 1 linha:

```
using HypothesisTests
UnequalVarianceTTest(a, b)
Two sample t-test (unequal variance)
-----
Population details:
```

```

parameter of interest: Mean difference
value under h_0: 0
point estimate: 0.4641197814036815
95% confidence interval: (0.4096, 0.5186)

Test summary:
outcome with 95% confidence: reject h_0
two-sided p-value: <1e-44

Details:
number of observations: [150, 150]
t-statistic: 16.756889937632724
degrees of freedom: 297.5604548062057
empirical standard error: 0.02769725069097449

```

A estatística t e graus de liberdade apresentados pela implementação são idênticos aos que encontramos realizando o procedimento passo a passo. Ao invés do valor exato ($p = 1.23^{-179}$), recebemos a informação de que $p < 1e^{-99}$. Diante do valor p obtido, concluiríamos que a distribuição dos dados como observada é improvável se for verdade a hipótese nula H_0 de que a diferença entre amostras é 0.

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

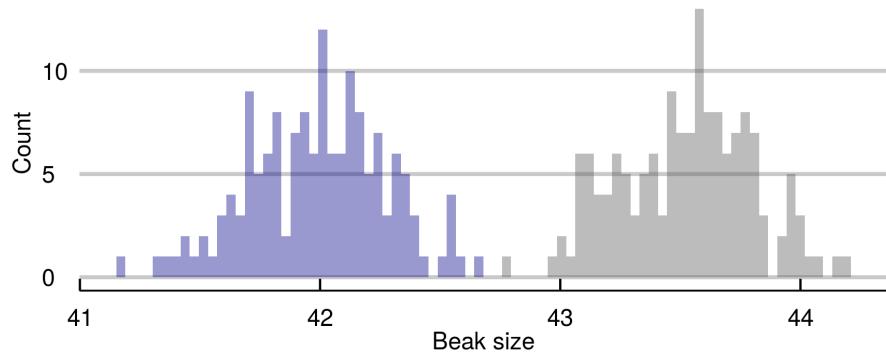
	Amostra A	Amostra B	valor p
Média(μ)	43,52	41,99	<0,001
Desvio-padrão(σ)	0,28	0,28	

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

	Amostra A	Amostra B	valor p
Mean (μ)	43,52	41,99	<0,001
Std. Dev. (σ)	0,28	0,28	

Darwin's Finches

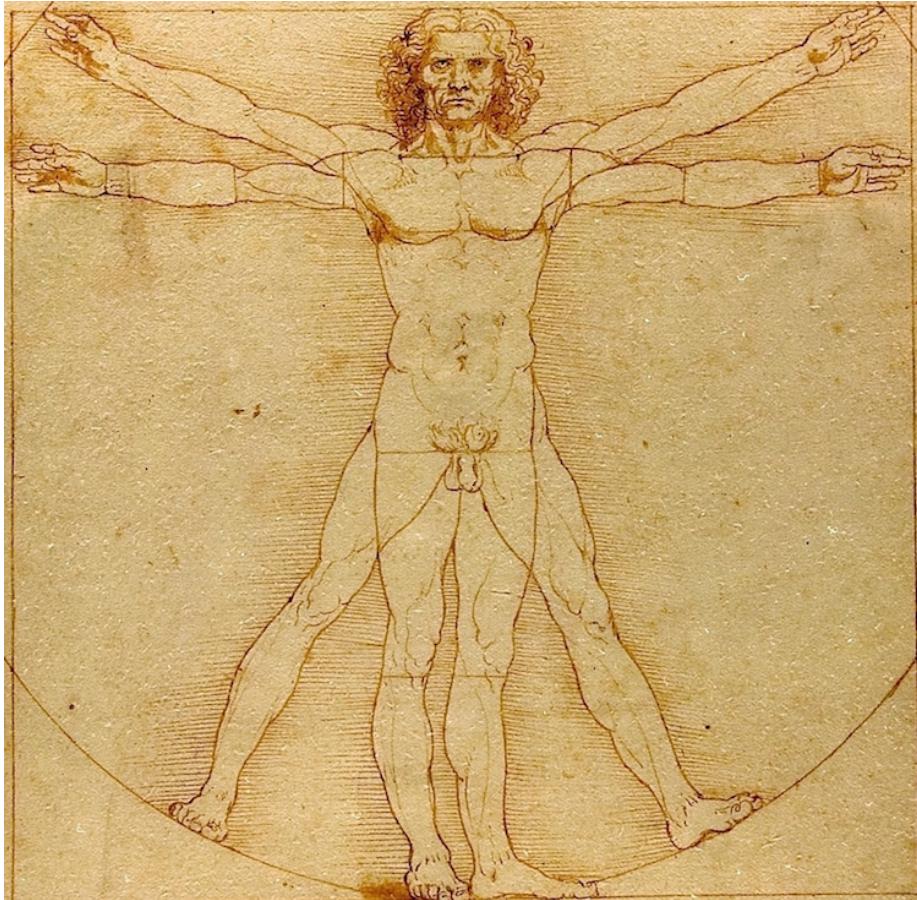
Samples A and B



Nota

Exercícios

1. Usando o dataset simulado no capítulo:
 - a. Execute teste T para cada par de amostras
 - b. Quais testes apresentam $p < 0.05$?
 - i. 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?
 - ii. Usando ggplot, plote histogramas para todos os pares comparados em apenas um painel. Dica: grid.arrange
 - iii. Plote boxplots para uma das comparações.
 - iv. A partir do gráfico anterior, adicione uma camada com violin plots (geom_violin) transparentes ($\text{alpha}=0$).
2. Usando o dataset iris
 - a. Escolha duas espécies e duas medidas.
 - b. Execute testes t para ambas as medidas
 - c. 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.
 - a. Examine os dados invocando “sleep”: >sleep
 - i. Plote histogramas para as medidas em ambos os grupos
 - ii. Execute um teste t supondo média populacional zero ($\mu = 0$).
 - iii. Execute um teste t entre amostras, supondo a mesma média ($H_0 : \mu_1 = \mu_2$).
4. Gerando a distribuição t:
 - a. Simule um conjunto de muitas medidas (sugestão: 100,000) a partir de uma distribuição normal ($\mu = 0, \sigma = 1$).
 - b. Retire 200 amostras de $n=30$ e salve as 200 médias (função sample).
 - c. Divida os valores por pelo erro padrão, $\frac{\sigma}{\sqrt{n}}$.
 - d. Retire 200 amostras de uma distribuição t com 29 graus de liberdade (função rt)
 - e. Plote o histograma superposto da distribuição obtida e da distribuição teórica



Capítulo 2 : Sobre a natureza das relações

Prelúdio: *Hypotheses non fingo?*

Eu ainda não fui capaz de descobrir a razão para essas propriedades da gravidade, e não faço hipóteses. Tudo aquilo que não é deduzido do fenômeno pode ser chamado de hipótese; e hipóteses, sejam metafísicas ou físicas, ou baseadas em qualidades ocultas, ou mecânicas, não têm lugar na filosofia experimental. Nesta filosofia, as proposições particulares são inferidas a partir do fenômeno, e então generalizadas por indução.

O racional apresentado no capítulo anterior é diretamente relacionado ao método hipotético-dedutivo e seus princípios filosóficos. Apesar de adequado a este cenário, 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*”.

O arcabouço descrito no capítulo anterior é suficiente para produzir um trabalho científico críptico para leigos.

Ao seguir receitas pré-definidas (formulação de H_0 e H_1 , cálculo de estatísticas e valores p), um texto parece estar em conformação com os padrões acadêmicos, mesmo que a hipótese elementar em torno do objeto de pesquisa seja simplória. Assim, inadvertidamente, priorizamos a forma e relegamos a segundo plano o miolo de propostas científicas.

Outro efeito colateral é a busca por valores p que rejeitem H_0 , desprezando precedentes teóricos e premissas probabilísticas (múltiplos testes).

A difícil interpretabilidade do valor p e as armadilhas frequentes envolvidas no processo de inferência levaram a comunidade científica a questionar a hegemonia desse parâmetro. Há uma presente tendência a abandonar o valor p e o limite $p < 0.05$ como critérios canônicos.

Vamos conhecer argumentos formais contra o método hipotético dedutivo nas ciências. Por enquanto, basta sabermos que é sempre vantajoso obter outras informações, complementares ou alternativas.

Neste capítulos, vamos aprender a estimar (1) a magnitude da diferença entre duas amostras e (2) quão relacionados são valores pareados (e.g. peso e altura).

I have not as yet been able to discover the reason for these properties of gravity from phenomena, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction.
Isaac Newton (1726). Philosophiae Naturalis Principia Mathematica, General Scholium. Third edition, page 943 of I. Bernard Cohen and Anne Whitman's 1999 translation, University of California Press ISBN 0-520-08817-4, 974 pages.

Tamanho de efeito

O tamanho de efeito nos ajuda a expressar magnitudes.

Retomando o exemplo anterior, de que adianta uma diferença significativa entre o tamanho dos bicos dos pássaros, se ela for de 0.00001 mm?

Ainda, existem casos em que estudos pequenos sugerem efeitos importantes, porém o tamanho amostral não fornece poder estatístico suficiente para rejeição da hipótese nula.

Além de saber quão improvável é a diferença observada, é natural imaginarmos o quão grande ela é.

Uma medida bastante popular é o *D de Cohen (Cohen's D)*.

É um parâmetro que expressa a magnitude da diferença sem usar unidades de medida.

Uma torcedora de futebol conta (feliz) a um amigo que seu time favorito venceu com placar de 4×1 (gols). Porém, esse amigo acompanha basquetebol e está acostumado a placares como 102×93 (cestas).

Como é possível comparar gols com cestas? Qual vitória representa pontuações mais discrepantes: 4×1 ou 102×93 ?

O problema aqui é que as pontuações se comportam de maneiras diferentes entre os esportes. Os placares no basquete possuem médias e dispersões muito maiores.

O D de Cohen consiste em expressar essa diferença em desvios-padrão. Bastante simples:

$$D_{cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

Usando a biblioteca *effects*, podemos calcular diretamente:

```
library(effects)
# O dataset galapagos_birds foi criado no capítulo 1
>cohen.d(galapagos_birds$X1,galapagos_birds$X2)

Cohen's d

d estimate: -5.460017 (large)
95 percent confidence interval:
      lower      upper 
-5.954047 -4.965987
```

Cohen propôs algumas faixas para classificar a magnitude desses efeitos:

	Pequeno	Médio	Grande
Cohen's D	0-0.2	0.2-0.5	0.5 - 0.8

Assim, podemos atualizar nossos resultados anteriores, reportando também o tamanho de efeito da diferença e seu intervalo de confiança. Se as distribuições forem da mesma família, temos uma estimativa comparável entre contextos.

Correlações

Na empreitada científica, não nos atemos apenas a comparações. Um objetivo mais nobre é descrever exatamente como se dá a relação entre entidades estudadas.

Como sabemos, existem muitas classes de funções para expressar relações entre variáveis/conjuntos. Nos capítulos anteriores, usamos algumas funções, como $y = \sqrt{x}$ e $y = e^x$.

Diversas leis naturais tornaram-se particularmente conhecidas, como a relação entre força, massa e aceleração, elucida por Newton:

$$\vec{F} = m\vec{a}$$

E a relação entre massa e energia para um objeto em repouso, descoberta por Einstein:

$$E = mc^2; c^2 \sim 8.988 * 10^{16} \frac{m^2}{s^2}$$

As equações acima descrevem uma relação linear entre grandezas.

Relações lineares

Uma relação linear entre duas variáveis indica que elas estão correlacionadas em uma proporção constante para qualquer intervalo.

Isto é, valores maiores de massa correspondem a um aumento proporcional em energia. O valor de c^2 expressa essa proporção constante.

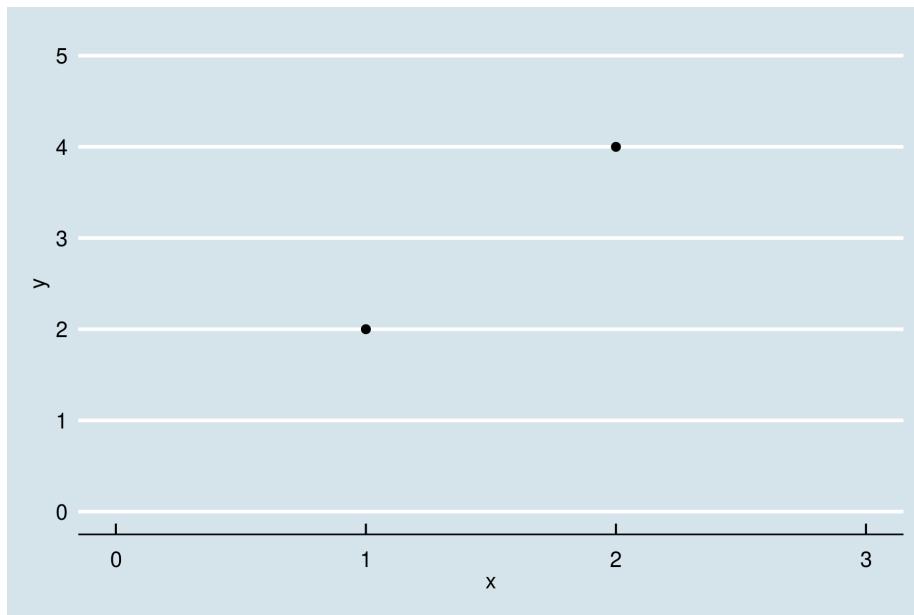
Exemplo: uma molécula de água pesa aproximadamente $m_{H_2O} = 2.992 \times 10^{-23} g$. Portanto, a energia associada é $E_{H_2O} = 2.992 \times 10^{-23} \times 8.988 \times 10^{16} \sim 2.689^{-6} J$. Se triplicarmos o número de moléculas de água, o mesmo acontecerá com a energia associada: $E_{3H_2O} = 3 \times E_{H_2O}$.

Se a correlação é positiva, incrementos em x serão proporcionais a incrementos em y . Se a correlação é negativa, incrementos em x serão proporcionais a decréscimos em y .

Num cenário perfeito, se sabemos que há uma relação linear entre variáveis, precisamos de apenas duas observações para descobrir proporção entre elas. Esse

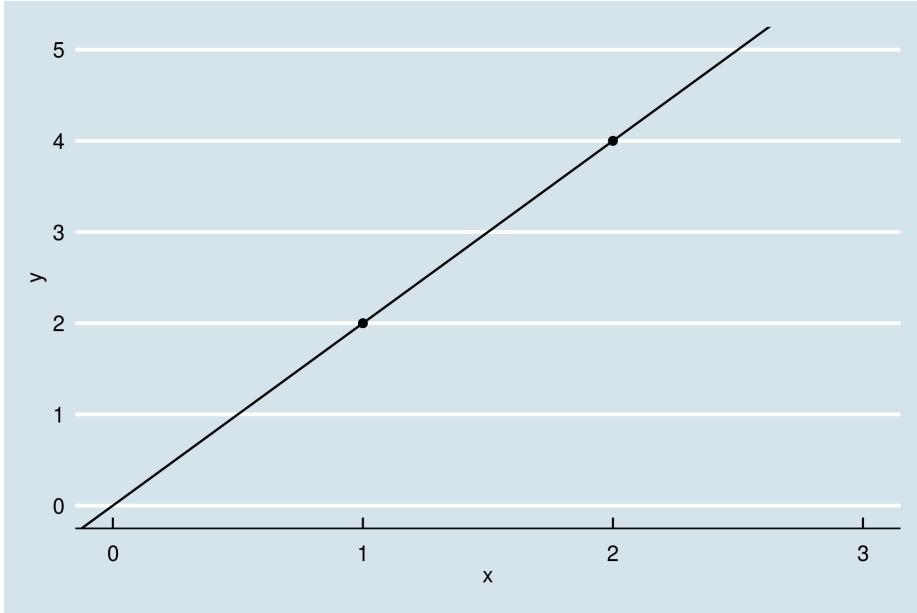
problema é idêntico ao de encontrar a inclinação da reta que passa por dois pontos. É de fácil resolução usando técnicas elementares.

```
>library(ggplot2)
>ggplot()+
  geom_point(mapping=aes(x=1,y=2))+
  geom_point(mapping=aes(x=2,y=4))+
  xlim(0,3)+ylim(0,5)+
  theme_economist()
```



$$y = \beta * x$$
$$a = (1, 2); b = (2, 4) \rightarrow \beta = 2$$

```
>ggplot()+
  geom_point(mapping=aes(x=1,y=2))+
  geom_point(mapping=aes(x=2,y=4))+
  xlim(0,3)+ylim(0,5)+
  geom_abline(slope = 2)+
  theme_economist()
```



Erros e aleatoriedade

Controlando fatores experimentais, as relações descritas são bastante precisas. Em um cenário sem atrito com superfícies e com o ar, os erros de medida obtidos com $\vec{F} = m\vec{a}$ são muito baixos.

Entretanto, nem sempre isso é verdadeiro.

Primeiro, podemos sofrer interferência de variáveis desconhecidas.

Imaginemos um conjunto de medidas antropométricas, com altura e peso e indivíduos.

É esperado que a altura de um ser humano esteja relacionada com seu peso. Entretanto, outras características não medidas, como percentual de gordura total, podem interferir nos valores finais. Normalmente, tratamos essas flutuações como erros aleatórios¹³.

Podemos simular este cenário partindo de variáveis idênticas e adicionando ruído aleatório.

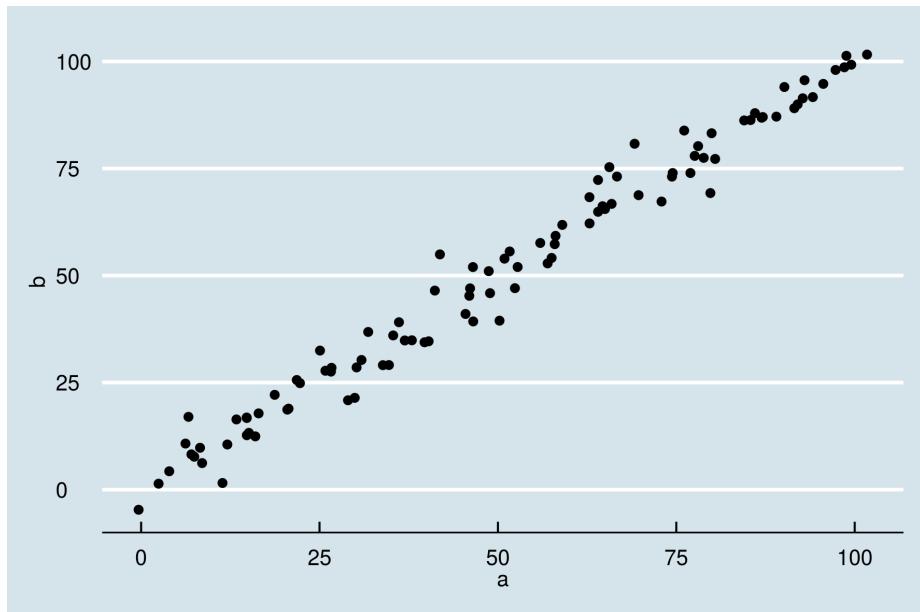
```
>set.seed(2600)
>a <- seq(1:100)+rnorm(n=100, sd=3)
>b <- seq(1:100)+rnorm(n=100, sd=3)
```

¹³A natureza da aleatoriedade é uma questão filosófica. Em última instância, podemos imaginar que seria possível explicar flutuações randômicas através de variáveis desconhecidas (*hidden variables*). Isso é verdade para a maioria dos fenômenos naturais. Entretanto, descobertas experimentais recentes em física quântica (*Bell's inequality experiment*) sugerem que variáveis ocultas não podem explicar a natureza probabilística das observações.

```

>cor_data <- data.frame(a,b)
>ggplot(cor_data,aes(x=a,y=b))+
  geom_point() + theme_economist()

```



O resultado sugere que há uma forte relação linear entre x e y . Por outro lado, notamos que é impossível para uma reta cruzar todos os pontos. A seguir, vamos investigar como quantificar a correlação linear, assim como encontrar a reta que minimiza a distância para todas as observações.

Com essas ferramentas, podemos estender nossas inferências. Além de comparações, teremos noções sobre a magnitude de uma relação, assim como poderemos prever o valor esperado para novas observações.

O coeficiente de correlação produto-momento de Pearson, ou, simplesmente, ρ de Pearson.

O coeficiente de correlação (ρ) de Pearson é um número real garantidamente¹⁴ entre -1 e 1. Expressa a magnitude e o sentido de uma relação linear, sendo -1 uma relação inversa perfeita e 1 uma relação direta perfeita.

Para os dados que geramos, a correlação é quase perfeita: $\rho = 0.989$.

O coeficiente possui *produto-momento* em seu nome, pois usa uma abstração originalmente empregada na física: o momento.

¹⁴Inequação de Cauchy–Schwarz

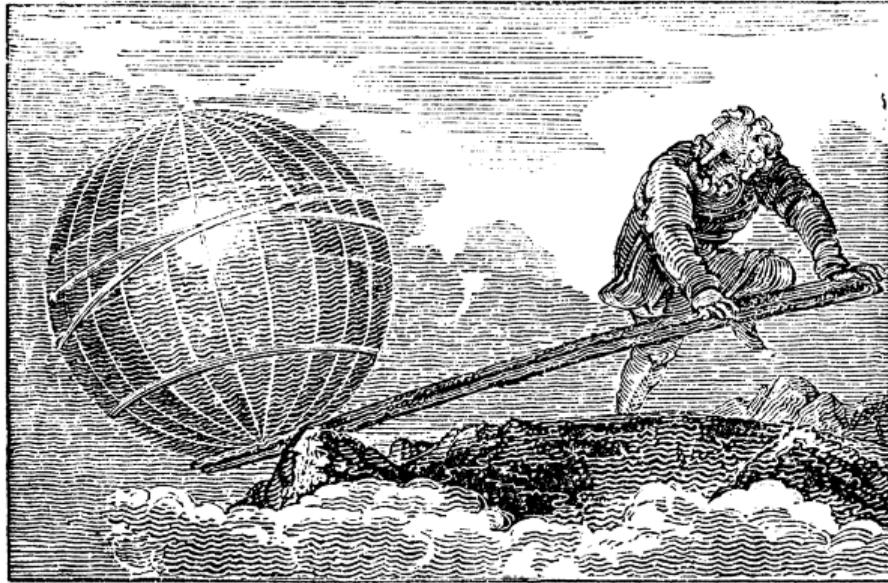


Figure 9: Dê-me um ponto de apoio e eu moverei a Terra

Um breve mergulho na física: Momentos

15

Para adquirir uma intuição sobre o coeficiente, é interessante resgatar o conceito físico de momento, originalmente concebido por Arquimedes. Embora não tenha inventado a alavanca, ele descreveu os princípios matemáticos por trás dela.

Em *Sobre o equilíbrio dos planos*, Arquimedes declara que *Magnitudes ficam em equilíbrio quando em distância reciprocamente proporcional aos seus pesos*.

Essa é a conhecida Lei da alavanca. Dado um ponto de apoio e um plano sobre ele, aplicamos uma força em qualquer local do plano. O momento (torque) resultante é o resultado da multiplicação da grandeza física (F) pela distância até o ponto fixo (d).

$$M = F * d$$

Supondo uma força constante, quanto mais nos afastamos do ponto fixo, maior o momento resultante. Posteriormente, os físicos estenderam o conceito para outros domínios. Por exemplo, um objeto com cargas opostas $-q$ e $+q$ separados por uma distância d possui momento (momento dipolar elétrico) análogo: $M = q * d$. De uma maneira geral, *falamos em momento ao multiplicarmos uma grandeza física por uma distância*.

¹⁵Pappus de Alexandria, Synagogue, Livro VIII

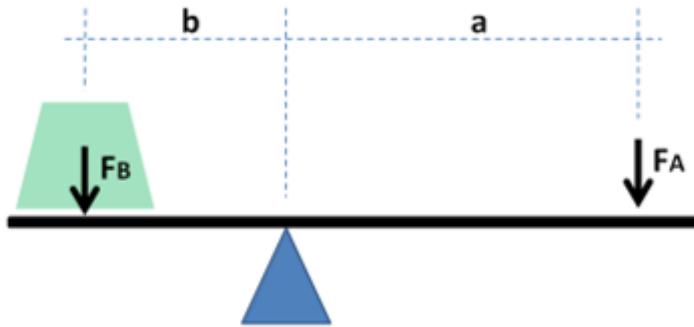


Figure 10: .

Momento resultante No caso da alavanca, vimos que cada força aplicada sobre o objeto está associada a um momento(torque). Sabemos que a gravidade atua sobre cada pedaço com massa compondo o todo. Podemos então calcular o momento resultante somando os momentos de todos os N pontos. Seja F_i a função descrevendo a força em cada i-ésimo:

$$M = \sum_{i=1}^N F_i d_i$$

Um sistema, como o pássaro apoiado sobre o dedo, está em equilíbrio quando a soma dos momentos em relação ao ponto fixo é zero. Para cargas elétricas, o sistema é apolar quando o momento é zero. Na figura abaixo, vemos como a molécula de CO_2 é apolar, enquanto a molécula de água é polar:

Os momentos descritos acima são expressões do *primeiro momento*, uma vez que a grandeza é multiplicada pela distância com expoente 1: $d = d^1$.

Podemos calcular outros momentos, exponenciando o componente espacial (distância). Vamos estudar agora momentos de massa de um objeto unidimensional:

O **momento zero** de massa para um objeto é $M_0 = \sum_{i=1}^N m_i d_i^0$. Como $d^0 = 1$, temos $M_0 = \sum_{i=1}^N m_i$, que é simplesmente a soma das massas de todos os pontos. O momento zero é a **massa total**.



Figure 11: Como o brinquedo acima fica equilibrado sobre apenas um ponto?

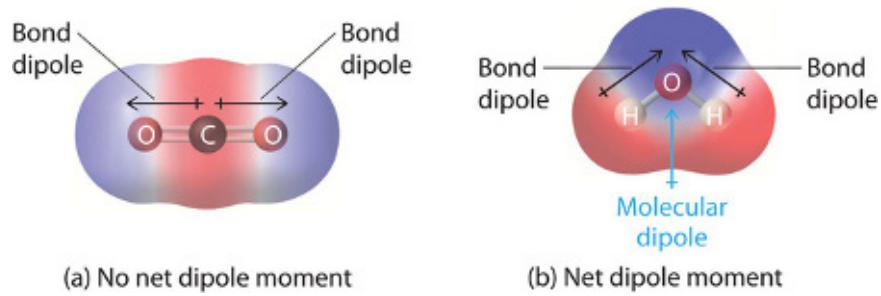


Figure 12: .

$$M_0 = m$$

O **primeiro momento** de massa para um objeto é $M_1 = \sum_{i=1}^N m_i d_i^1$ e determina o **centro de massa** em relação à dimensão d . É o ponto em que está o dedo em que se equilibra o pássaro da foto.

$$M_1 = C_m$$

O **segundo momento** de massa é $M_2 = \sum_{i=1}^N m_i d_i^2$ e é o **momento de inércia**. Corresponde à resistência do sistema a rotações. Perceba que os termos d_i^2 estariam presentes na área de um círculo com centro idêntico ao do objeto e raio igual à distância para o centro: πd^2 . A resistência total a rotação é análoga à resistência oferecida pelos raios destes círculos imaginários¹⁶.



O n-ésimo momento é dado por

$$M_n = \sum_{i=1}^N m_i d_i^n$$

¹⁶<https://physics.stackexchange.com/a/371165/218274>

Generalizando momentos

Podemos generalizar ainda mais a abstração e calcular momentos de entidades abstratas, como variáveis aleatórias. **Melhor: já fizemos isso anteriormente!**

Seja $f(x)$ a função que descreve uma distribuição de probabilidades para a variável,

Assim como o **momento zero** representa a soma da contribuição de cada ponto para a massa (massa total), aqui ele representa a soma das probabilidades possíveis, a probabilidade total (1).

O **primeiro momento** corresponde ao centro de massa na mecânica estática. Para probabilidades, é o centro, a **média**.

O **segundo momento** corresponde ao momento inercial e é a **variância**.

Os momentos **terceiro** e **quarto** normalizados informam sobre assimetrias (*skewness*) e peso de valores extremos (*kurtosis*).

Formalmente, seja $d(x, x_0)$ o valor da distância ao centro x_0 de referência ($x - x_0$), o n-ésimo momento μ_n é definido por:

$$\mu_n = \int_{-\infty}^{\infty} d(x, x_0)^n f(x) dx$$

A integral acima corresponde à versão contínua da soma de partes discretas apresentadas antes para uma grandeza física, como a massa: $M_n = \sum_{i=1}^N d_i^n m_i$

Momento zero:

$$\mu_0 = \int_{-\infty}^{\infty} d(x, x_0)^0 f(x) dx$$

A soma de todas probabilidades de uma distribuição deve somar 1.

$$= \int_{-\infty}^{\infty} f(x) dx = 1$$

Primeiro momento: $\mu_1 = \int_{-\infty}^{\infty} d(x, x_0) f(x) dx$, supondo centro em 0 ($x_0 = 0$), temos a média,

$$\mu_1 = \int_{-\infty}^{\infty} x f(x) dx$$

, também chamado valor esperado $E[X]$. Estende a intuição de somar as medidas e dividir pelo número de observações ao passo em que usamos uma integral para somar as infinitesimais possibilidades para $f(x)$.

Segundo momento:

$$\mu_2 = \int_{-\infty}^{\infty} d(x, x_0)^2 f(x) dx$$

. Como vimos no capítulo introdutório, a soma dos quadrados dos desvios, nossa variância,

$$\sigma^2 = E[(x - \mu)^2]$$

Notas finais sobre o Teorema do Limite Central

Podemos entender melhor o teorema do limite central. As informações fornecidas pelos momentos são valiosas: uma função de probabilidade é totalmente definida por seus momentos.

O Teorema do Limite Central, de que falamos antes, é provado mostrando equivalência entre momentos da curva normal e da soma de n distribuições idênticas através de outras ferramentas.

Podemos criar uma *Função geradora de momentos*, $M_X(t) = E[e^{tX}]$ em que t é um valor fixo. Chamamos ela assim, pois sua forma polinomial via expansão de Taylor corresponde à uma série que contém todos os momentos M_n : $1 + tX + \frac{t^2 M_2}{2!} + \frac{t^3 M_3}{3!} + \dots$, já que $\frac{de^x}{dx} = e^x$ e a derivada de ordem n multiplica a de ordem $n - 1$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

$$\begin{aligned} E[M_X(t)] &= 1 + tE[X] + \frac{t^2 E[X^2]}{2!} + \frac{t^3 E[X^3]}{3!} + \dots \\ &= 1 + tM_1 + \frac{t^2 M_2}{2!} + \frac{t^3 M_3}{3!} + \dots \end{aligned}$$

A *Função característica* é a transformada de Fourier da função de densidade, associando valores a componentes periódicos no plano imaginário. Envolve multiplicar t pela unidade na definição da função geradora de momentos $M_X(t) = E[e^{tX}]$, $\phi_X(t) = M_X(it) = E[e^{itX}]$. É possível usar a função característica para mostrar que os momentos na soma de distribuições semelhantes convergem para os momentos de uma distribuição gaussiana. Isto é: $\phi_{\sum X_n}(t) \sim \phi_{N(\mu, \sigma)}(t)$ para X_n semelhantes¹⁷.

Com os conceitos adquiridos em mãos, é fácil entender o ρ de Pearson.

¹⁷As primeiras provas assumiam X_n idênticas, porém versões mais gerais foram demonstradas. Two Proofs of the Central Limit Theorem, Yuval Filmus, 2010. <http://www.cs.toronto.edu/~yuvalf/CLT.pdf>

Calculando correlações lineares

A noção de **distância** ou **desvio** se repetiu muitas vezes.

De fato, o coeficiente de correlação linear nasceu quando Francis Galton (1888) estudava numericamente dois problemas aparentemente distintos em antropometria¹⁸:

1. **Antropologia:** Se recuperássemos de um túmulo antigo apenas um osso da coxa (fêmur) de um indivíduo, o que poderíamos dizer sobre sua altura?
2. **Ciência forense:** Com o intuito de identificar criminosos, o que pode ser dito sobre medidas diferentes de uma mesma pessoa?

Galton percebeu que, na verdade, estava lidando com o mesmo problema. Dadas medidas pareadas, (x_i, x'_i) , o que o desvio de x_i informa sobre o desvio de x'_i ?

O fêmur recuperado do esqueleto de um faraó é 5 cm maior que a média. Quão distante da média esperamos que seja sua altura? Ingenuamente, podemos pensar que se uma das medidas é 1% maior que a média, a outra também será 1% maior. Galton percebeu que havia um armadilha nesse pensamento.

Apesar de haver uma relação entre as medidas, há também flutuações aleatórias: parte do desvio é resultante disso. Precisamos entender o grau de correlação pra fazer um bom palpite.

Então, propôs um coeficiente mensurando a relação entre desvios de variáveis. Se tamanho do fêmur e altura estão muito relacionadas, um fêmur grande sugere indivíduo igualmente alto. Caso contrário (baixa correlação), um fêmur grande (desvio alto) não implica grande estatura.

Para quantificar a relação, multiplicamos os desvios de cada par de medidas:

$$Cov(X, X') = \sum_{i=1}^N (x_i - \mu_x)(x'_i - \mu_{x'})$$

A expressão acima expressa a **covariância** entre X e X' e será útil em outros contextos. A expressão lembra o cálculo do primeiro momento, porém cada desvio é multiplicado pelo desvio correspondente da medida pareada. Daí o nome coeficiente de correlação *produto-momento*.

Note que, se ambos os desvios concordam em sentido (sinal), o resultado da multiplicação será positivo. Pares consistentemente concordantes aumentam o valor da soma final. Se ambos os desvios discordam em sentido (sinal), o resultado será negativo. Pares consistentemente discordantes diminuem o valor da soma final.

Assim, podemos ter variáveis altamente correlacionadas positiva ou negativamente, desde que o sentido da associação seja constante. Em contrapartida, se as

¹⁸Francis Galton's account of the invention of correlation. Stephen M. Stigler. Statistical Science. 1989, Vol. 4, No. 2, 73-86.

medidas são ora discordantes e ora concordantes, os valores tendem a se anular na soma e o resultado se aproxima de zero.

Observar apenas a covariância é perigoso, pois os valores dependem da unidade de medida e da dispersão dos dados.

Calculamos o coeficiente de correlação de Pearson, normalizando¹⁹ a covariância ao dividí-la pelo produto dos desvios-padrão:

$$\rho_{XX'} = \frac{cov(X, X')}{\sigma_X \sigma_{X'}}$$

De forma extensa:

$$\rho_{XX'} = \frac{\sum_{i=1}^N (x_i - \mu_x)(x'_i - \mu_{x'})}{\sqrt{\sum_i^N (x_i - \mu_x)^2} \sqrt{\sum_i^N (x'_i - \mu_{x'})^2}}$$

Uma boa notícia: ρ segue uma distribuição conhecida, a distribuição t, com $n-2$ graus de liberdade. Podemos usar as ferramentas anteriores para testar hipóteses.

Exemplo prático

O exemplo a seguir foi um feliz achado. Na época, o governo brasileiro discutia a necessidade da ampliar número de médicos para melhorar a assistência à saúde. Alguns defendiam ser uma decisão acertada, enquanto outros advogavam que os investimentos deveriam ser feitos em outras áreas da saúde.

Por curiosidade, accesei os dados da WHO (World Health Organization) e do banco mundial (World Bank) sobre quantidade de médicos por país e indicadores de saúde. Minha expectativa era encontrar pelo menos uma tímida relação entre indicadores. Mais do que isso, entender qual a localização do Brasil em relação a outros países. Fui surpreendido por uma forte correlação, que exploraremos a seguir.

Adotamos países como unidade observacional com medidas x , o número de médicos 1,000 habitantes, e y , a expectativa de vida saudável ao nascer.

Usando dados obtidos dos portais da WHO e do World Bank, plotamos os pontos no plano cartesiano.

```
# http://apps.who.int/gho/data/view.main.HALEXv
# https://data.worldbank.org/indicator/SH.MED.PHYS.ZS
>library(magrittr)
>library(ggplot2)
>library(dplyr)
```

¹⁹Aqui, normalização tem o sentido de ajustar a escala das medidas. Não confundir com transformações para que os dados passem a ter distribuição gaussiana.

```

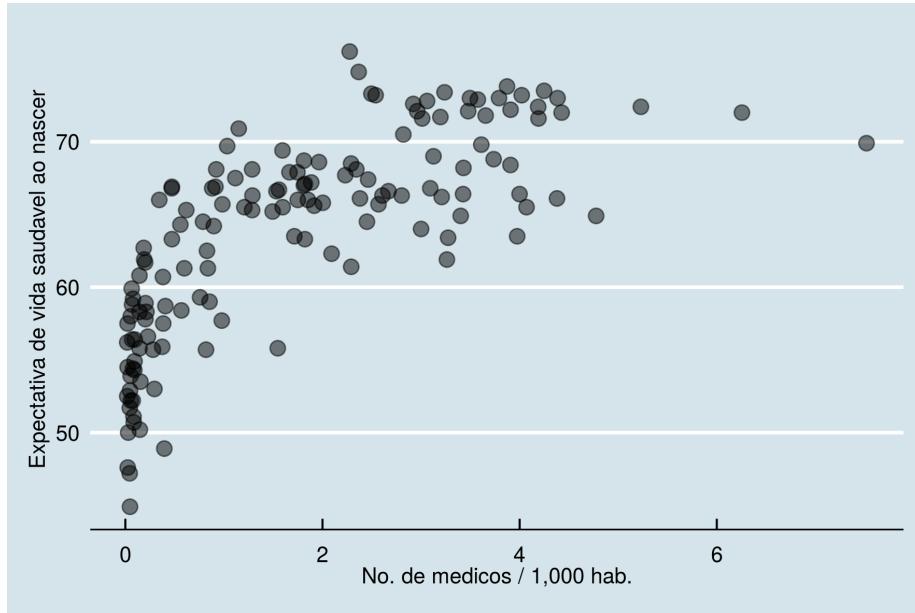
>worldbank_df <- read.csv("data/API_SH.MED.PHYS.ZS_DS2_en_csv_v2_10227587.csv",
                           header = T, skip = 3)
>colnames(worldbank_df)[1] <- "Country"

>worldbank_df$n_docs <- sapply(split(worldbank_df[,53:62], #lists of values
                                         seq(nrow(worldbank_df))),
                                 function(x) tail(x[!is.na(x)],1)) %>% #ultimos valores não nulos
                                         as.numeric

>who_df <- read.csv("data/who_lifeexpect.csv",skip=2)
>who_df$hale <- who_df$X2016
>uni_df <- left_join(worldbank_df[,c("Country", "n_docs")],
                      who_df[,c("Country", "hale")], by="Country")

>ggplot(uni_df, aes(x=n_docs, y=hale))+
  geom_point(alpha=0.5, size=3) +
  xlab("No. de médicos / 1,000 hab.")+
  ylab("Expectativa de vida saudável ao nascer")+
  theme_economist()

```



É evidente que o padrão não é aleatório. Visualmente, notamos que o valor da expectativa de vida aumenta com maior N° de médicos. Ainda, notamos um aumento inicialmente rápido até atingir um platô. O padrão é semelhante ao de uma curva logarítmica.

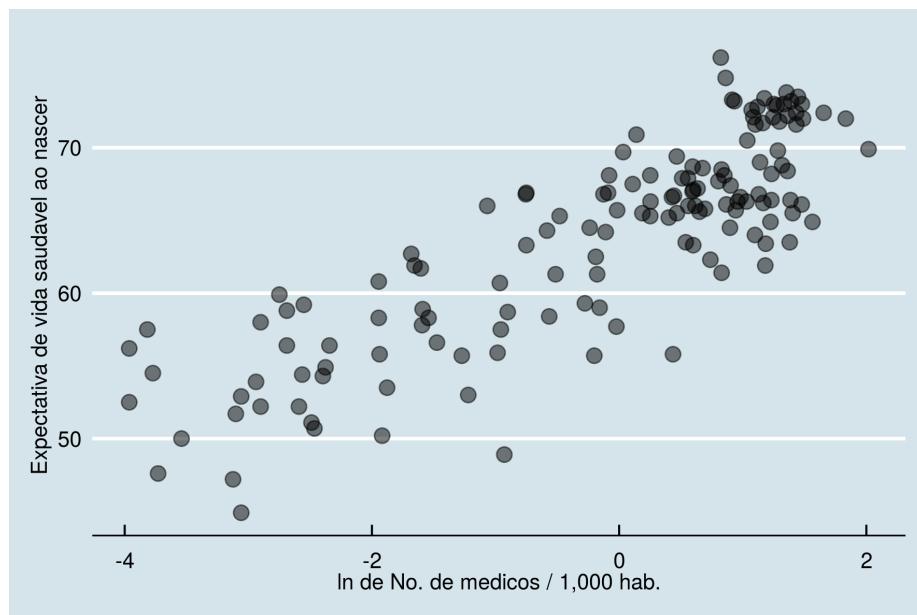
$$y = \log(x) \text{ ou } HALE = \log(N_{\text{médicos}})$$

Se essa hipótese for verdade, transformar o número de médicos usando função logarítmica tornará a relação linear com a variável transformada:

Se $y = \log(x)$, fazemos a substituição $x' = \log(x)$ para obtermos $y = x'$.

Então a expectativa de vida se torna linearmente correlacionada ao logaritmo do número de médicos.

```
>uni_df$log_docs <- log(uni_df$n_docs)
>ggplot(uni_df,aes(x=log_docs,y=hale))+
  geom_point(alpha=0.5,size=3) +
  xlab("ln de No. de medicos / 1,000 hab.")+
  ylab("Expectativa de vida saudavel ao nascer")+
  theme_economist()
```



De fato, verificamos uma notável tendência linear para os pontos.

Usando a implementação nativa em R para o coeficiente de Pearson:

```
>cor.test(uni_df$log_docs,uni_df$hale)
Pearson's product-moment correlation
data: uni_df$log_docs and uni_df$hale
t = 18.572, df = 143, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7854248 0.8828027
sample estimates:
cor
0.8407869
```

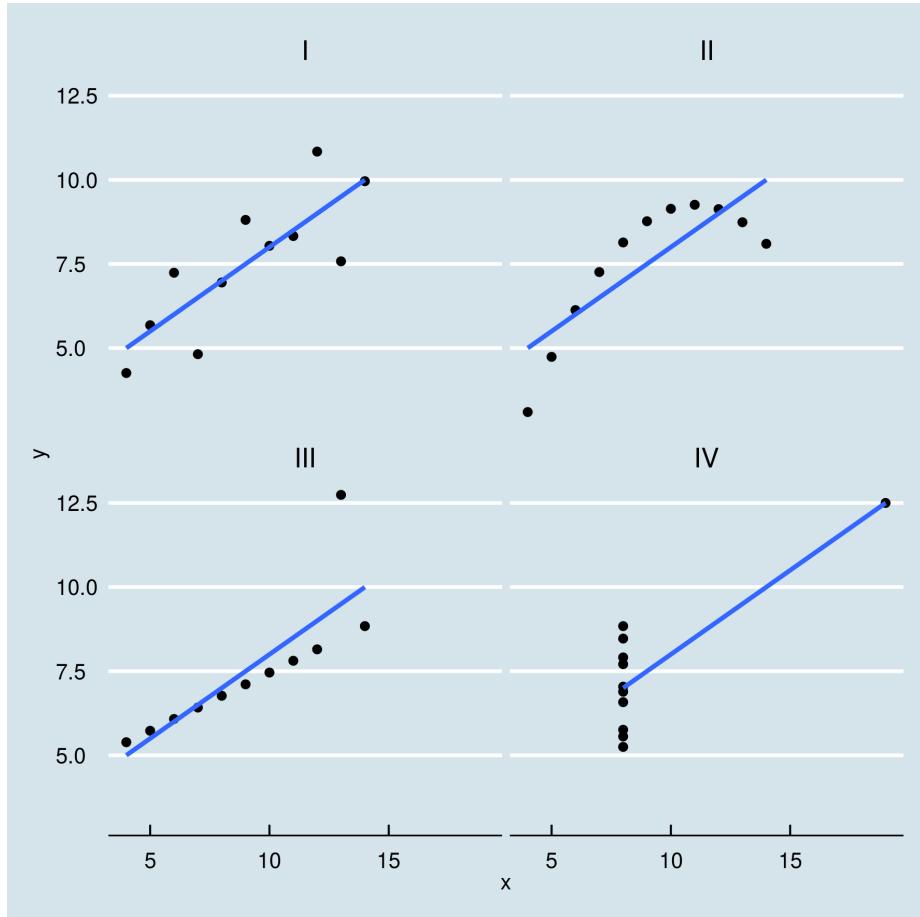
A correlação linear obtida para nossa amostra de países é surpreendentemente grande, como sugeria a visualização ($\rho \sim 0.841$).

O valor p é baixo ($p < 0.001$) considerando a hipótese nula H_0 de $\rho = 0$. Concluímos então que há uma relação linear significativa de forte magnitude entre o logaritmo do número de médicos e a expectativa de vida dos países em nossa amostra.

É realmente curioso que exista uma relação matemática tão evidente entre construtos tenuamente conectados. O tempo médio que um organismo leva entre nascimento e morte e o número de profissionais atuantes. É virtualmente impossível explicitar cada relação causal por trás dessa relação, que se manifesta de forma robusta através da soma de muitos fatores relacionados.

Nota *É costumaz afirmar que não existe relação entre variáveis caso o coeficiente de relação não se mostre importante. Como vimos, esse indicador informa apenas sobre relações lineares entre variáveis. A visualização dos dados pode ser de grande ajuda na inferência sobre a natureza de relações.*

Dados com distribuições bastante diferentes podem resultar em coeficientes iguais, como mostra o clássico quarteto de Anscombe. As 4 amostras abaixo apresentam o mesmo coeficiente de correlação.



Previsões

Agora, sabemos que é razoável assumir uma relação linear entre essas variáveis. Como dito antes, podemos então encontrar a reta que minimiza a distância para as observações.

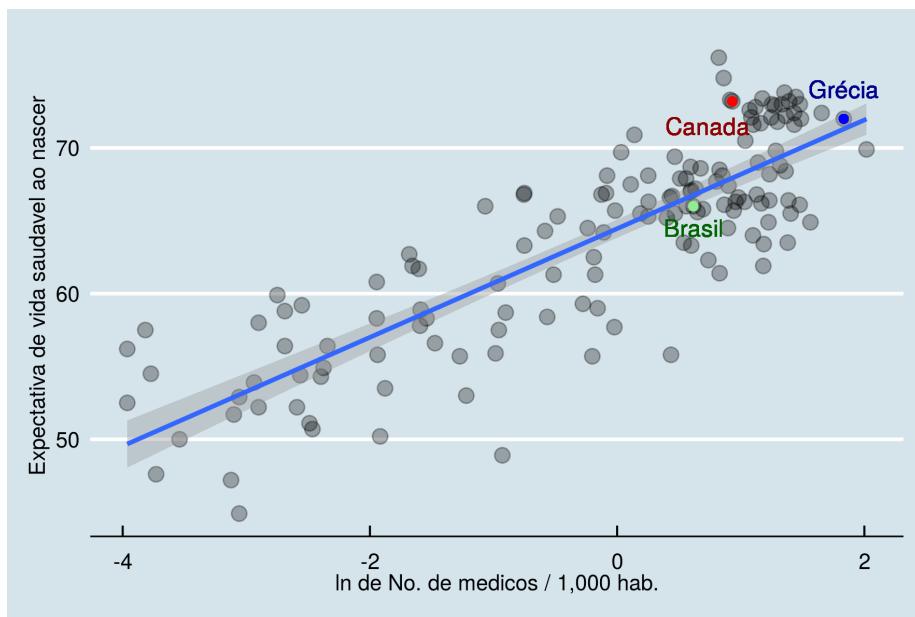
A equação que descreve essa reta nos informa o valor esperado para expectativa de vida dado o número de médicos.

```
>uni_df$log_docs <- log(uni_df$n_docs)
>ggplot(uni_df,aes(x=log_docs,y=hale))+
  geom_point(alpha=0.3,size=3) + geom_smooth(method="lm")+
  geom_point(y=66.0,x=0.61626614,color="light green")+
  geom_text(y=64.5,x=0.61626614,label="Brasil",color="dark green")+
```

```

geom_point(y=73.2,x=0.93177030,color="red")+
geom_text(y=71.5,x=0.73177030,label="Canada",color="dark red")+
geom_point(y=72.0,x=1.833381,color="blue")+
geom_text(y=74.0,x=1.833381,label="Grécia",color="dark blue")+
xlab("ln de No. de medicos / 1,000 hab.")+
ylab("Expectativa de vida saudavel ao nascer")+
theme_economist()

```



Vieses devem ser enderaçados antes de conclusões, mas o modelo é suficientemente interpretável para tomar decisões.

Uma boa política pode comparar o valor de investimento por setores com outros países em condições semelhantes e resultados diferentes.

Assumindo que realmente há uma relação linear, vemos que o Brasil está bastante próximo do esperado para o número de médicos²⁰. Caso a estratégia seja contratar mais pessoas, podemos nos espelhar em programas de países com mais médicos por habitante resultados correspondentes (e.g. Grécia).

Se a estratégia for economizar com a folha de pagamentos e priorizar investimento em estrutura, podemos usar países com expectativa de vida alta para o número de profissionais esperado (e.g. Canada).

²⁰É praticamente consenso entre especialistas que o Brasil possui problema de distribuição de profissionais, com déficit de médicos em áreas mais pobres e pouco populosas.

Predições com modelos lineares

Como adivinhar uma medida com base na outra? Considerando a relação linear descoberta anteriormente, podemos criar uma função que receba como input o valor de uma variável (número de médicos) e retorne como output o valor esperado para a expectativa de vida.

Descobrir a equação que descreve esta função consiste em encontrar a reta que melhor se ajusta à nuvem de pontos, como na figura anterior.

Para isso, calculamos a inclinação (β_1) e o ajuste vertical (β_0) que minimizam a soma das distâncias entre a reta e as observações. O termo ϵ corresponde aos erros, com distribuição normal de média 0 e desvio padrão σ .

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

Ajustamos o modelo usando a função `lm(linear model)` do R:

```
# log_docs : x' = log(x)
> lm(hale ~ log_docs, data=uni_df)

Call:
lm(formula = hale ~ log_docs, data = uni_df)

Coefficients:
(Intercept)    log_docs
       64.46        3.73
```

Temos $\beta_0 \sim 64.46$ e $\beta_1 \sim 3.73$.

Nossa estimativa para a expectativa de vida saudável “começa” em 64.46 anos e aumenta com o número de médicos no país. Especificamente, aumenta em 3.73 para cada unidade de nossa variável transformada ($\log(x)$).

Em nosso dataset, o Brasil possui 1.852 médicos/1,000 hab. Nossa predição então é:

$\hat{y}_{Brasil} = \log(1.852 * 3.73 + 64.46 \sim 66.8$, o que está bastante próximo do número real(66).

Estimadores

Existe mais de uma maneira de estimar esses parâmetros.

Uma de particular interesse, que também servirá em outros contextos, é a de Maximum likelihood (máxima verossimilhança).

Primeiro, determinamos uma função que descreve a probabilidade da observação na variável alvo (y_i) ocorrer dadas medidas das variáveis preditoras (x_i) e um conjunto de parâmetros (β_k).

Podemos adotar como função de verossimilhança (*likelihood function*) para os valores y_i uma distribuição de probabilidades gaussiana cuja média é dada pela

reta $\mu_{yi} = \beta_0 + \beta_1 * x_i$. Assim, a probabilidade de cada valor y_i é dada por uma gaussiana, de acordo com o desvio para o valor previsto pela reta.

$$L \sim N(\mu_{yi}, \sigma^2)$$

Assumindo que as observações são independentes, a probabilidade do conjunto de observações é dada pelo produto delas.

$$L = \prod_{i=1}^n P(y_i | x_i; \beta_0, \beta_1, \sigma^2)$$

Substituindo os valores de μ para a gaussiana pelas previsões da reta:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}$$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y_i - (\beta_0 + \beta_1 x_i)^2}{2\sigma^2}}$$

Essa é nossa função de verossimilhança e expressa a probabilidade de observarmos as medidas y_i dadas as medidas x_i e considerando um conjunto de parâmetros (β_0, β_1) .

O objetivo então é encontrar parâmetros que maximizem essa função. Por conveniência, aplicamos uma transformação logarítmica nesta função (*log likelihood function*). Isso transforma nosso produtório em um somatório e passamos o contradomínio do intervalo $[0; 1]$ para $[-\infty, 0)$.

$$\begin{aligned} \text{log likelihood}(\beta_0, \beta_1, \sigma^2) &= \log \prod_{i=1}^n P(y_i | x_i; \beta_0, \beta_1, \sigma^2) \\ &= \sum_{i=1}^n \log P(y_i | x_i; \beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned}$$

Os parâmetros que maximizam a função de verossimilhança (max. likelihood, ML) são os mesmos que maximizam o logaritmo da função de verossimilhança (log-likelihood).

Introduzimos o racional do estimador ML pois ele será útil futuramente. Em verdade, é fácil entender as fórmulas fechadas para nossos parâmetros, pois apenas expressam as relações lineares exploradas²¹:

$\hat{\beta}_1$ expressa a magnitude da correlação entre X e Y . É natural que seu valor seja a covariância normalizada pela variância do preditor.

$$\hat{\beta}_1 = \frac{cov(XY)}{\sigma_x^2}$$

$\hat{\beta}_0$ é nosso intercepto, então é a diferença entre médias preditas e previsões considerando o valor médio em X.

$$\hat{\beta}_0 = \mu_y - \hat{\beta}_1 \mu_x$$

Por fim, a variância dos erros $\hat{\sigma}^2$ é dada pelo quadrado dos desvios das previsões em relação às medidas.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

As soluções acima fornecem as melhores estimativas que podemos obter minimizando a distância da reta aos pontos.

Devemos então nos preocupar em saber se o modelo linear encontrado é bom na predição dados.

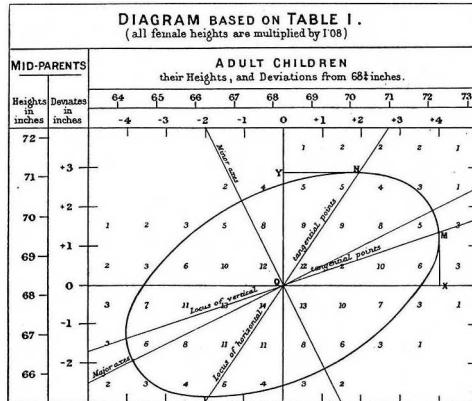


Figure 13: O primeiro gráfico de regressão linear. Ilustração de Francis Galton (1875) relação entre altura de pais e filhos.

²¹Detalhes das deduções dos estimadores OLS and Max. Likelihood: <https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/05/lecture-05.pdf> ; <https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/06/lecture-06.pdf>

Avaliando performance Existem diferentes parâmetros para avaliar a performance de um modelo. Em geral, eles buscam quantificar o quanto os resultados do modelo se distanciam de resultados ideais.

Para regressão linear, o R^2 (coeficiente de determinação) é um coeficiente bastante usado. Expressa a proporção entre **(1)** variância explicada pelo modelo e **(2)** variação total. Chamamos de resíduo(ou erro) a diferença entre valores preditos e valores reais.

(1) Para capturar a magnitude dos erros do modelo, somamos o quadrado de todos os resíduos (*sum of squared residuals, SSR*) em relação aos valores preditos. Sejam y_i as observações e \hat{y}_i as previsões:

$$SSR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

(2) A variabilidade total é quantificada pela soma do quadrado dos desvios em relação à média (*total sum of squares, TSS*), um termo que vimos no cálculo da variância (segundo momento):

$$TSS = \sum_{i=1}^n (y_i - \mu_y)^2$$

Então a fração $\frac{SSR}{TSS}$ é a proporção desejada. Definimos R^2 como:

$$R^2 = 1 - \frac{SSR}{TSS}$$

Uma visualização intuitiva de SSR e TSS:

```
>source("aux/multiplot.R")
>doc_lmfit <- lm(hale ~ log_docs, data=uni_df)
>uni_df$preds[complete.cases(uni_df)] <- predict(doc_lmfit)
>uni_df$hale_mean <- mean(uni_df$hale, na.rm = T)
>ssr_res <- ggplot(uni_df, aes(x=log_docs, y=hale)) +
  geom_point(alpha=0.5, size=3) +
  geom_segment(aes(xend = log_docs, yend = preds)) +
  geom_smooth(method="lm") +
  xlab("") +
  ylab("Expectativa de vida saudável ao nascer") +
  ggplot2::ggtitle("SSR") + theme_economist()

>tss_res <- ggplot(uni_df, aes(x=log_docs, y=hale)) +
  geom_point(alpha=0.5, size=3) +
  geom_segment(aes(xend = log_docs, yend = hale_mean)) +
  geom_abline(slope = 0, intercept = 63.28165) +
  xlab("ln de No. de médicos / 1,000 hab.") +
```

```

ylab("Expectativa de vida saudavel ao nascer")+
ggplot2::ggtitle("TSS")+theme_economist()

>multiplot(ssr_res,tss_res)

```

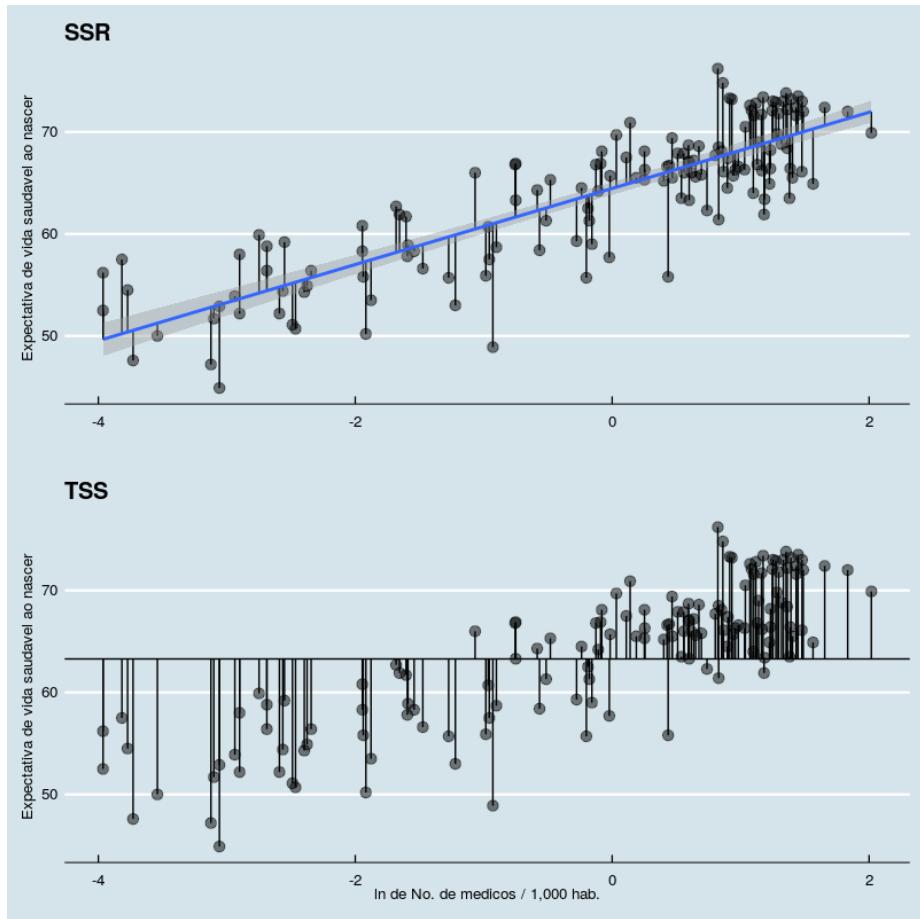


Figure 14: O quadrado da distância entre um ponto e a reta corresponde a um resíduo. Obtemos SSR e TSS somando todos os resíduos nas figuras superior e inferior, respectivamente.

Valores de R^2 próximos a 1 indicam soma de resíduos (SSR) similar a 0. Usar a reta como guia acumula erros quase nulos. Valores de R^2 próximos a 0 indicam $\frac{SSR}{TSS} \sim 1$ e as previsões obtidas pelo modelo são tão boas quanto chutar a média para todos os casos.

```

>lm(hale ~ log_docs, data=uni_df) %>% summary
Call:
lm(formula = hale ~ log_docs, data = uni_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.0964 -2.3988  0.3233  2.8229  8.6708 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 64.4613    0.3162 203.84 <2e-16 ***
log_docs     3.7303    0.2009   18.57 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 3.779 on 143 degrees of freedom
(119 observations deleted due to missingness)
Multiple R-squared:  0.7069,    Adjusted R-squared:  0.7049 
F-statistic: 344.9 on 1 and 143 DF,  p-value: < 2.2e-16

```

Para obter os valores preditos, usamos o método *predict*:

```

>head(predict(doc_lmfit))

 2      3      4      7      8      9 
59.90747 57.23226 65.39962 66.11533 69.54483 68.30608

```

É possível também obter previsões para novos valores especificando o argumento *newdata*. Para um país com 1.5 médicos/1,000 habitantes:

```

>predict(doc_lmfit,newdata = data.frame(log_docs=log(1.5)))
 1 
65.97381

```

Premissas Existem alguns procedimentos auxiliares para checar possíveis falhas e pontos no modelo que precisam de atenção. Por exemplo, os resíduos podem ser assimétricos. Isso indica que o desempenho muda em diferentes intervalos (heteroscedacidade). Diferentes violações necessitam de atitudes diferentes, como tratar outliers ou mudar tipo do modelo. Uma lista completa de premissas, junto aos códigos em R para testá-las, está disponível no material auxiliar (*lm-assumptions.R*)

Correlações e testes não paramétricos

Verificamos minuciosamente análises envolvendo a distribuição normal, a distribuição t e relações lineares. Entretanto, muitas vezes as medidas não seguem uma distribuição definida. Assim, realizar inferências usando os **parâmetros** descritos ($\mu, \sigma, t\dots$) nos levaria a conclusões erradas.

Para lidar com distribuições arbitrárias, vamos abrir mão deles e conhecer ferramentas *não-paramétricas*: o coeficiente de correlação de ranks ρ de Spearman e o teste U de Mann Whitney.

Ranks e o ρ de Spearman

Relações lineares mantêm proporções constantes e aprendemos como quantificá-las. Por outro lado, duas variáveis podem ter relações de outros tipos, não lineares. Em especial, se as medidas apresentam valores muito extremos (*outliers*) um cálculo como o anterior sofre bastante com vieses.

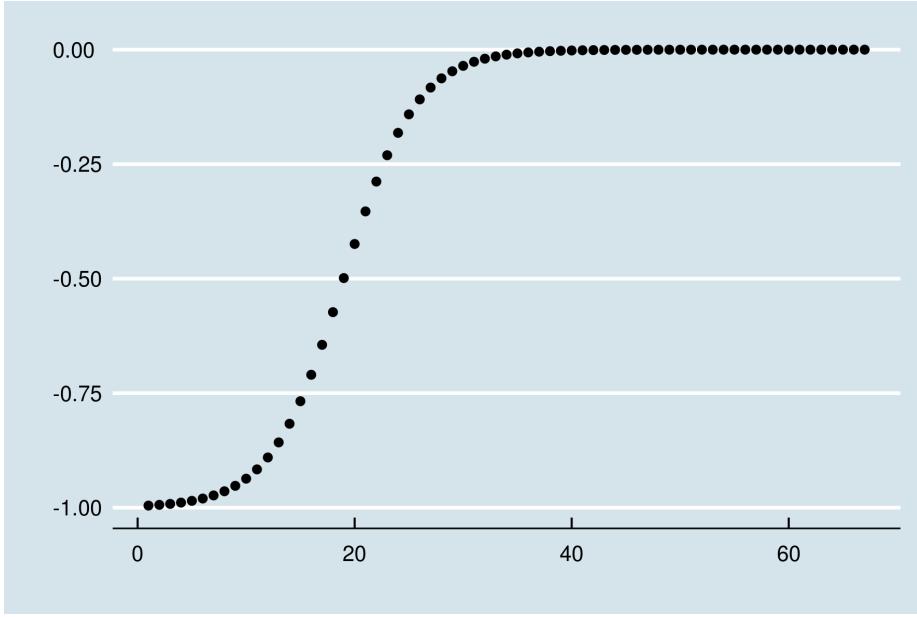
Uma simples solução para esse problema é ranquear os valores. Assim, os itens do conjunto são tratados pela sua posição em relação a outros itens, de forma independente dos valores associados. Exemplo:

$$S = (1, 3, 89, 89, 39, 209) \rightarrow S_{ranked} = (1, 2, 4, 4, 3, 5)$$

O ρ de Spearman é que o coeficiente produto-momento de Pearson aplicado aos ranks. Assim, medimos o grau em que duas variáveis aumentam (ou diminuem) em magnitude observando apenas a ordem das observações. Isto é: **maior que, igual ou menor que**. Especificamente, investigamos se há uma relação de *monotonicidade* entre elas.

Para a relação (sigmoide), entre x e y abaixo:

```
>set.seed(2600)
>sig_data <- data.frame(y_vals = -(1 / (1 + exp(seq(-10,10,by =0.3 )*100 ))),
                           x_vals = 1:67)
>ggplot(sig_data,aes(x=x_vals,y=y_vals))+
  geom_point() + theme_economist() + xlab("") + ylab("")
```



O coeficiente de Pearson é $\rho \sim 0.850^{22}$:

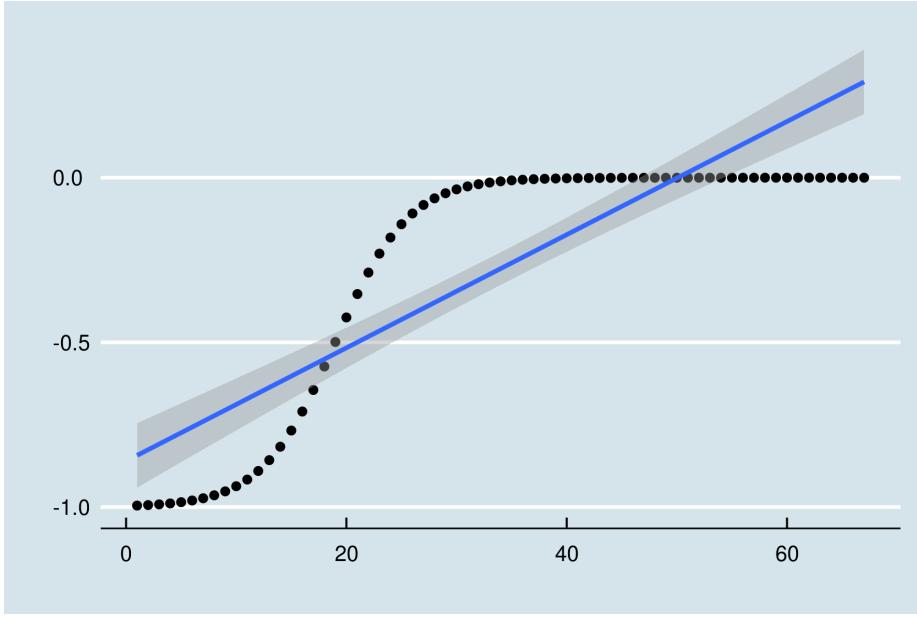
```
>cor.test(sig_data$y_vals,
+          sig_data$x_vals)

Pearson's product-moment
correlation

data: sig_data$y_vals and +sig_data$x_vals
t = 12.993, df = 65, p-value <
2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7658181 0.9051711
sample estimates:
cor
0.8497162

>ggplot(sig_data,aes(x=x_vals,y=y_vals))+
  geom_point()+ geom_smooth(method="lm")+
  theme_economist() + xlab("") + ylab("")
```

²²Como observamos no gráfico, a correlação linear não é tão alta. O coeficiente se aproxima de 1 $\rho \sim 0.850$ pois os desvios superiores compensam simetricamente os inferiores. O exemplo reforça a importância de plotar os dados para um melhor entendimento (ver Quarteto de Anscombe).



Como a relação é perfeitamente monotônica, os pares ordenados (x_i, y_i) sempre possuem o mesmo rank. O quinto valor mais alto em x é também o quinto valor mais alto em y. Portanto, o coeficiente de Spearman é 1:

```
>cor.test(sig_data$y_vals,
+          sig_data$x_vals, method = "spearman")

  Spearman's rank correlation rho

data: sig_data$y_vals and sig_data$x_vals
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

O coeficiente ρ de Spearman é preferível quando as medidas parecem diferir muito quanto à família da distribuição de origem. Especialmente, quando a média não parece corresponder bem ao centro das distribuições. Lembre-se que o coeficiente de Pearson é baseado nos desvios em relação à média em ambas as amostras.

Teste U de Mann-Whitney

O teste U de Mann-Whitney faz uso da estatística U para fazer inferências. O racional é idêntico ao do teste t de Student.

Estabelecemos hipótese nula H_0 e hipótese alternativa H_1 .

Então, calculamos a probabilidade de nossas observações acontecerem caso a hipótese nula seja verdadeira.

Desta vez, usaremos a estatística U. Lembremos que a estatística t era calculada com base em parâmetros extraídos da amostra:

$$t = Z/s = (\mu' - \mu)/\frac{\sigma}{\sqrt{n}}$$

A estatística U não depende de parâmetros (e.g. μ , σ), sendo calculada com base em cada observação.

Primeiro, calculamos os ranks de cada medida r_i unindo as observações das amostras A e B, de tamanhos amostrais n_a e n_b em apenas um conjunto ($N_{tot} = n_a + n_b$).

Depois, separamos novamente as amostras e calculamos a soma dos ranks em cada grupo, chamadas R_a e R_b .

A estatística U é dada pela seguinte expressão:

$$U_a = R_a - \frac{n_a(n_a + 1)}{2}$$
$$U_b = R_b - \frac{n_b(n_b + 1)}{2}$$

Usamos o menor valor de U para consultar a probabilidade (valor p) correspondente para a hipótese nula.

O termo $\frac{n(n+1)}{2}$ corresponde à soma mínima dos ranks para a amostra.

Os ranks são uma sequência regular (1, 2, 3, ...), de forma que a soma de todos os valores é idêntica à soma de uma progressão aritmética de N termos.

$$\Sigma_{ranks} = \frac{N(N + 1)}{2}$$

Enquanto R_i corresponde à soma dos ranks calculados com as duas amostras, o termo acima corresponderia à soma mínima dos ranks para uma amostra, caso os ranks ocupassem a sequência inicial $A = (1, 2, 3, 4, \dots, n_a)$ na amostra conjunta. A definição para o teste não é unânime na literatura, de forma que alguns autores e softwares (e.g. R) implementam o cálculo com a subtração acima e outros (e.g. S-PLUS) não o fazem.

Em R, as funções **dwilcox(x,m,n)** e **pwilcox(q,m,n)** retornam a distribuição e a densidade cumulativa para a estatística U correspondente a amostras com

tamanhos m e n. **wilcox.test(x,y,...)** é a implementação base do teste de Mann Whitney. O teste de Mann Whitney é o teste de Wilcoxon de duas amostras.

Exercícios

1. O coeficiente produto-momento de Pearson descreve quais tipos de relação?

- Ele é útil para modelar relações quadráticas entre variáveis?
- Citamos relações não lineares, como $E = mc^2$. Cite um outro exemplo de fenômeno natural de perfil não-linear em que o ρ de Pearson não funciona.

2. Crie uma função que calcula o n-ésimo momento para uma amostra:

- `n_moment <- function(x,n) {sum((x - mean(x))^n)/length(x)}`
- Calcule o valor de skewness. Como citado no capítulo, é o 3º momento normalizado [pelo 2º momento ao expoente 3/2].

$$\frac{\mu_3}{\mu_2^{3/2}}$$

- Calcule o valor de kurtosis. Como citado, é o 4º momento noramlizado [pelo quadrado do 2º momento menos 3].

$$\frac{\mu_4}{\mu_2^2 - 3}$$

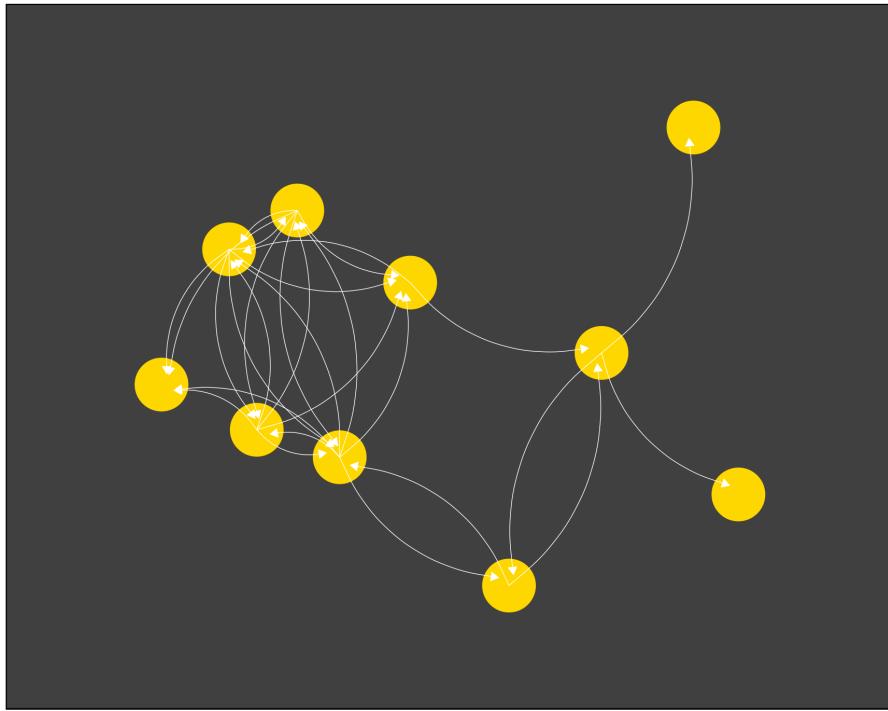
- Os valores podem ser conferidos com as implementações `e1071::skewness` e `e1071::kurtosis`

3. Usando o dataset *iris*, compare as 4 variáveis numéricas (*Sepal/Petal Length/Width*) entre espécies (*Species*) usando teste t de Student e teste de U Mann Whitney. Em algum caso os métodos divergem quanto à rejeição da hipótese nula?

- Obtenha o tamanho de efeito (D de Cohen) para as diferenças.

4. Usando o dataset *iris*:

- Faça um scatterplot entre duas medidas. A função `pairs` pode ajudar.
 - Verifique se há correlação linear significativa entre as variáveis.
 - Se existir, ajuste um modelo de regressão linear.
 - Ajuste um modelo de regressão para cada espécie.
 - Observe os valores de R^2 para cada modelo. Qual a sua impressão sobre as mudanças de performance?



Capítulo 3 : Análise multivariada, grafos e inferência causal

Introdução

Neste capítulo, incorporaremos construtos como base para estudar um conceito do berço da filosofia ocidental: *causalidade*. A filosofia Aristotélica investiga causas materiais, formais, eficientes e finais. Causas exprimem a ideia de isolar relações entre fatores. A maioria das definições envolvem *efeitos* que dependem, mesmo que parcialmente, de *causas* precedentes. Relações de causalidade *exploram* a evolução de sistemas em certas condições.

Até este ponto, aplicamos modelagem matemática para uma ou duas variáveis aleatórias. Procedimentos diferentes foram empregados para correlação, comparação e regressão. Neste capítulo, lidaremos com análise multivariada. Diagramas causais e controle de vieses, mediação, moderação, regressão múltipla, análise de componentes principais e análise fatorial.

Regressão múltipla

Nos modelos lineares simples, calculamos parâmetros para um intercepto β_0 , inclinação da reta β_1 e variância dos erros σ^2_ϵ . No exemplo apresentado, relacionamos o número de médicos (n) com a expectativa de vida saudável $hale$ em um país.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$
$$hale_i = \beta_0 + \beta_1 n_i + \epsilon$$

Na *regressão linear múltipla*, introduzimos mais uma variável preditora. Em nosso exemplo, poderia ser o valor do IDH do país:

$$hale_i = \beta_0 + \beta_1 n_i + \beta_2 IDH'_i + \epsilon$$

Em geral, temos dois objetivos:

- (1) melhorar a performance do modelo ao adicionar informações pertinentes;
- (2) examinar as relações considerando múltiplas variáveis.

O primeiro objetivo é intuitivamente óbvio, entretanto precisamos ter cuidado com redundância de informações. Especificamente, há uma troca quase inevitável entre complexidade e robustez do modelo. Acrescentar variáveis ou usar classes de relações mais flexíveis implica dar liberdade para um sobreajuste aos dados. Isto é, nosso modelo aprenderá idiossincrasias sobre os dados disponíveis (datasets WHO e World Bank) e não sobre a relação entre as abstrações (e.g. expectativa de vida saudável). Veremos nas próximas sessões como mitigar esse problema.

Um outro objetivo para a regressão múltipla é examinar o efeito modificador das variáveis acrescentadas. Em especial, é comum incluir variáveis auxiliares para corrigir estimativas.

Exemplo: queremos estimar um parâmetro β_1 para a relação entre altura e peso. Ajustamos um modelo: $Altura = \beta_0 + \beta_1 * Peso + \epsilon$. Entretanto, sabemos que a altura média de homens é maior que a de mulheres. Ao examinar a relação entre a altura e peso, podemos incluir a variável *sexo* no modelo, $Altura = \beta_0 + \beta_1 * Peso + \beta_2 * Sexo + \epsilon$.

Nossa estimativa de β_1 é modificada de maneira a levar em conta os efeitos do sexo.²³

Veremos uma formalização desse conceito a seguir, com o procedimento para examinar mediação.

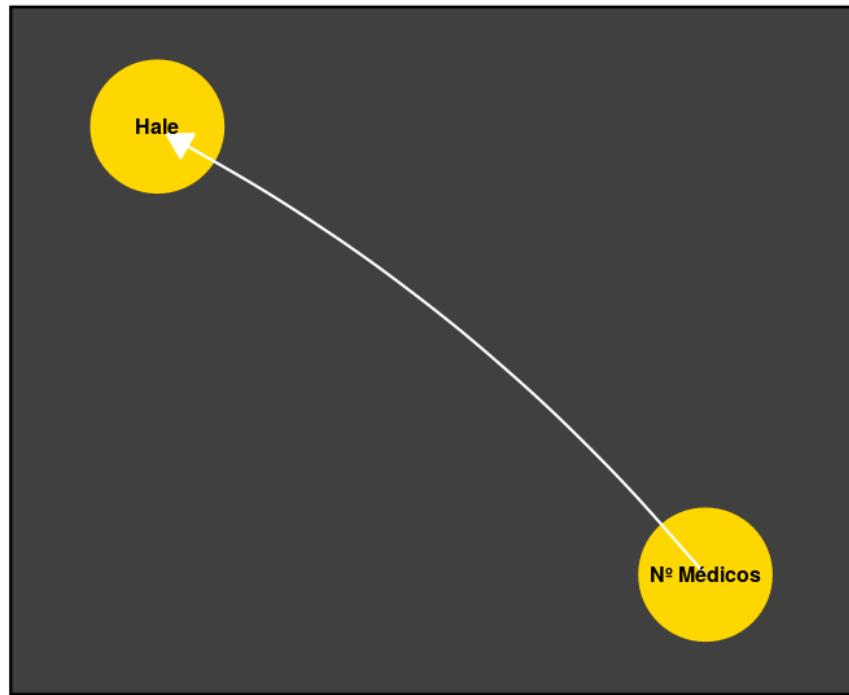
²³Sexo é uma variável dicotômica (macho/fêmea). Costumamos codificá-las de forma binária (0/1; e.g: macho = 1 / fêmea = 0). Assim, um sujeito macho terá a estimativa de altura acrescida em $\beta_2 * 1$, enquanto fêmeas terão este termo zerado $\beta_2 * 0$. Chamamos esse truque de *dummy coding*.

Costumeiramente, traduzimos os procedimentos acima afirmando que a estimativa para “*a relação entre X e Y é controlada para confundidores [A, B e C]*”. A esse ponto, fica óbvio que a simplificação linguística é perigosa. A falta de cautela em traduzir abstrações matemáticas para linguagem natural é responsável pela injusta fama da estatística como ferramenta para enganos.

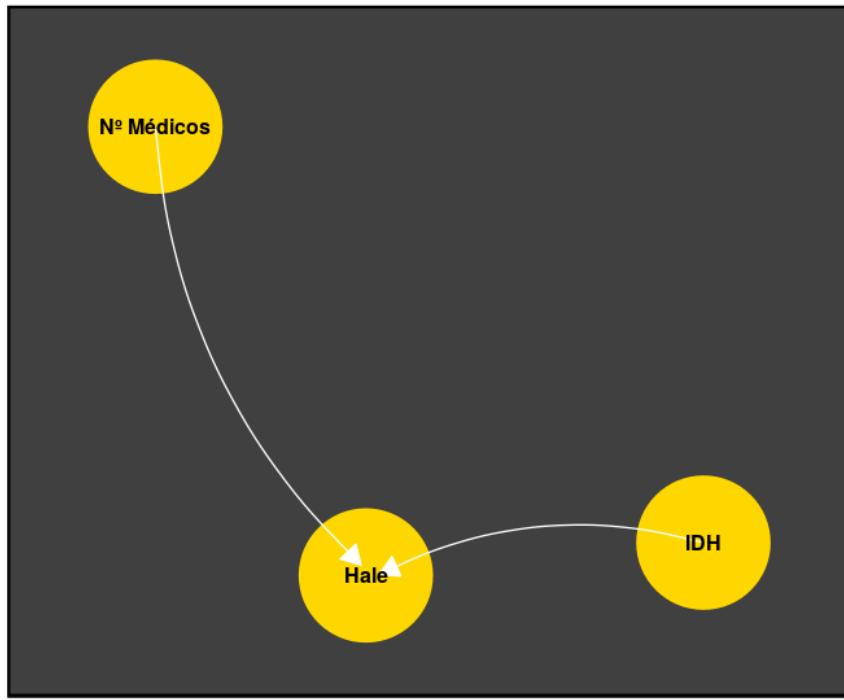
Assim como o valor p é indevidamente interpretado muitas vezes, o “controle para confundidores” nada mais é que o ajuste de estimativas considerando um modelo causal. É recomendado que os confundidores sejam mitigados experimentalmente (e.g. randomização).

Grafos e trajetórias causais

Podemos usar os diagramas a seguir para ilustrar uma regressão linear simples:



Ou múltipla com dois preditores:



É fácil relacionar *nodos com variáveis* e *conexões com relações* descritas pelas equações estimadas. Formalmente, tratamos essas abstrações com o nome de **grafos**. O campo começou a ser tratado por Euler em 1736. Chamamos os pontos de nodos, ou vértices, e as ligações de arestas (*edges*). Cada aresta conecta dois nodos.

O conceito foi usado para resolver o problema das pontes de Königsberg. Dada uma série de pontes conectando partes diferentes da cidade, fazer um percurso que cruzae cada uma apenas uma vez?

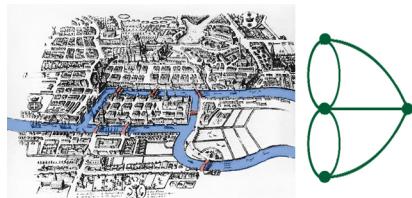


Figure 15: .

Euler mostrou que era impossível. Note que não usamos distâncias. Apenas descrevemos como elementos são conectados. Podemos atrelar diversas estruturas. Os grafos acima, por exemplo, são direcionados e possuem equações vinculadas.

As equações e procedimentos de que lançamos mão anteriormente são soluções equivalentes às representações gráficas. É possível generalizar a ideia, usando diagramas para tratar matematicamente formulações de teorias científicas.

Grafos e trajetórias causais

"The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated.", Sewall Wright, Correlation and Causation, 1921

A pouco conhecida origem deste campo está no trabalho de um geneticista, Sewall Wright. Ele assumiu que a correlação entre variáveis é resultante da influência de muitas trajetórias causais. Então, propôs uma forma de medir a influência de cada trajetória sobre uma variável-alvo.

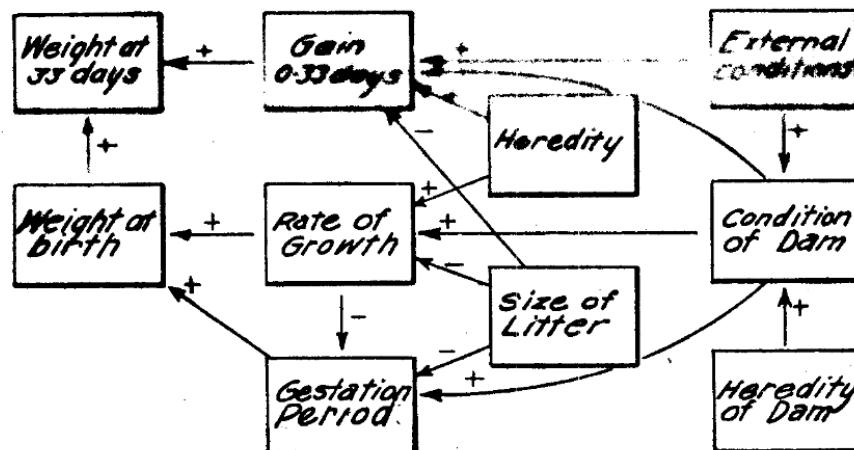


FIG. 1.—Diagram illustrating the interrelations among the factors which determine the weight of guinea pigs at birth and at weaning (33 days).

Figure 16: Diagrama mostrando relação entre fatores influenciando peso de um porquinho-da-índia. Wright, 1921

Usando grafos direcionados (as conexões têm uma origem e um destino), é atrelar as noções de correlação e regressão de forma a ilustrar caminhos causais entre relações lineares. Sewall começou usando apenas grafos acíclicos (sem trajetórias retornando a um mesmo ponto de origem) direcionados, DAGs, em condições restritas.

Décadas depois, o campo foi extrapolado para outros cenários mais gerais. Em específico, o boom de disponibilidade de poder computacional nas décadas de 1960 e 1970 impulsionou o surgimento de estimadores diversos para parâmetros nesses modelos.

É esperado que a quantidade de parâmetros cresça conforme a complexibilidade.

Um trabalho valoroso foi feito por Judea Pearl para unificar as abordagens. Pearl mostrou que muitos *frameworks* são situações especiais de modelos de equação estrutural. Ele escreveu textos comprehensivos alinhando a matemática aplicada a uma base epistemológica.

É especialmente digno de nota o conceito de *contrafactual*. Para estimar um efeito causal, imaginamos quais seriam as condições em um cenário sem ação do agente causal. Pearl conduz um cauteloso estudo lógico-semântico das definições na tentativa de construir um sistema coerente de pesquisa empírica.

Examinando covariáveis com Modelos causais Modelos causais baseados em grafos pressupõem efeitos unidirecionais. Isso preclui a descrição acurada de muitos casos. Por outro lado, o uso parcimonioso é uma ferramenta valiosa para fazer inferências causais.

O DAG a seguir analisa a qualidade de uma cerveja. Ela depende de água, lúpulo (hops) e malte. Queremos entender como a composição dos ingredientes sólidos (lúpulo & malte) interfere na pureza final, avaliada pela ausência de agrotóxicos. Temos dados de algumas fábricas locais. A concentração na água em cada cidade também varia, o que interfere diretamente na pureza final da cerveja. Além disso, a água é usada para regar o solo com lúpulo e malte, interferindo também indiretamente no desfecho.

```
library(dagitty)
library(ggdag)

dagified <- dagify(Quality ~ Water + HopsMalt,
                    HopsMalt ~ Soil, Soil ~ Water,
                    exposure = "HopsMalt",
                    outcome = "Quality")
p1 <- ggdag(dagified) + theme_dag_blank()
p1
```

Traçar um DAG permite examinar formalmente os possíveis caminhos por onde a informação flui e, assim, realizar inferências corretamente. De forma prática, queremos:

- 1 . Testar se o modelo causal proposto é compatível com as observações.
- 2 . Estimar o efeito condicionando-o às covariáveis adequadas.

O grafo implica algumas *independências condicionais*. Isso significa que, se ele estiver correto, algumas variáveis serão independentes.

```

impliedConditionalIndependencies(dagified)
#HopsMalt _||_ Water | Soil
#Quality _||_ Soil | HopsMalt, Water

```

A notação $A \perp\!\!\!\perp B \mid C, D, E, F\dots$ indica que A deve ser independente de B, se condicionarmos a estimativa do efeito às covariáveis C, D, E, F... “Condicionar a” significa incluir a covariável no modelo descritivo. A forma mais simples é através de regressão múltipla:

```

beer_data <- simulateSEM(dagified,b.lower = 0.20,b.upper=0.25)
# HopsMalt _||_ Water | Soil
lm(HopsMalt ~ Water + Soil,beer_data)
# Quality _||_ Soil | HopsMalt, Water
lm(Quality ~ Soil + HopsMalt,beer_data)

```

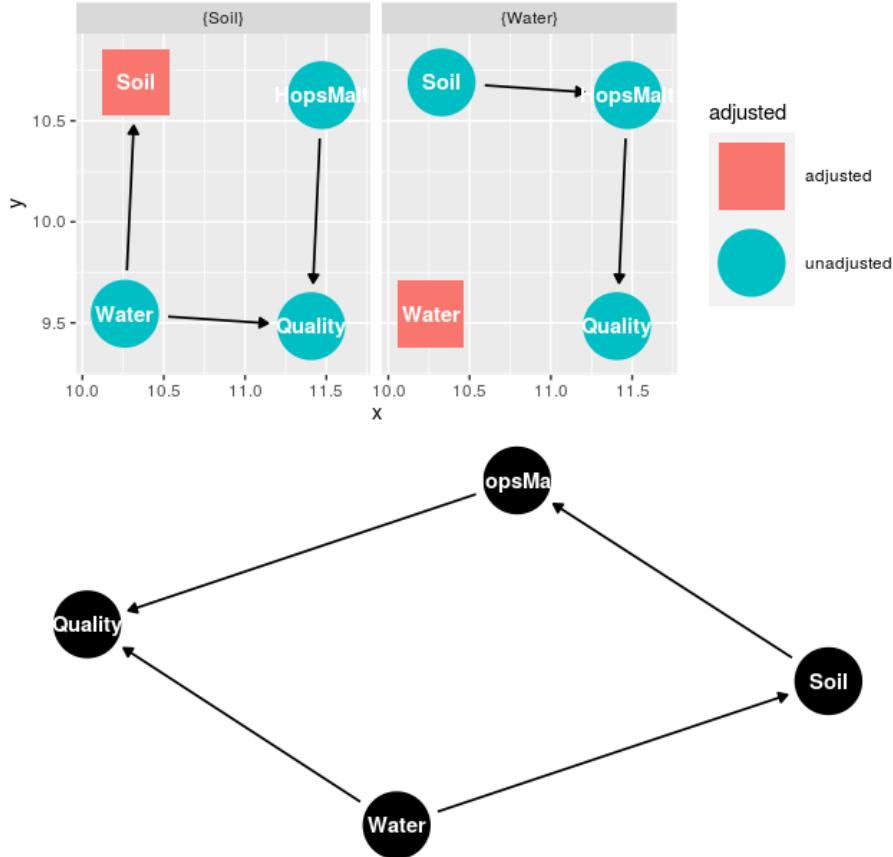
O esperado é que a estimativa do efeito (coeficiente) seja próxima de zero (associação inexistente) uma vez que condicionamos ela às covariáveis indicadas. Verificamos que isso acontece para o exemplo:

```

#(...)
#Coefficients:
#(Intercept)      Water          Soil
#   0.03709       0.02105       0.28069

#(...)
#Coefficients:
#(Intercept)      Soil          HopsMalt
#   0.05652       0.08347       0.15118

```



Uma vez que aceitamos o DAG como adequado, podemos usá-lo como referência para calcular estimativas não-enviesadas (unbiased estimates). Isso significa que estamos ajustando o valor final de acordo com vias pelas quais a informação pode correr nas covariáveis examinadas.

A função `adjustmentSets` quais conjuntos de covariáveis podemos incluir para obter estimativas não enviesadas. A função `ggdag_adjustment_set` informa visualmente quais caminhos estamos fechando ao condicionarmos num grupo de covariáveis. Por vezes (como no exemplo), temos conjuntos alternativos:

```

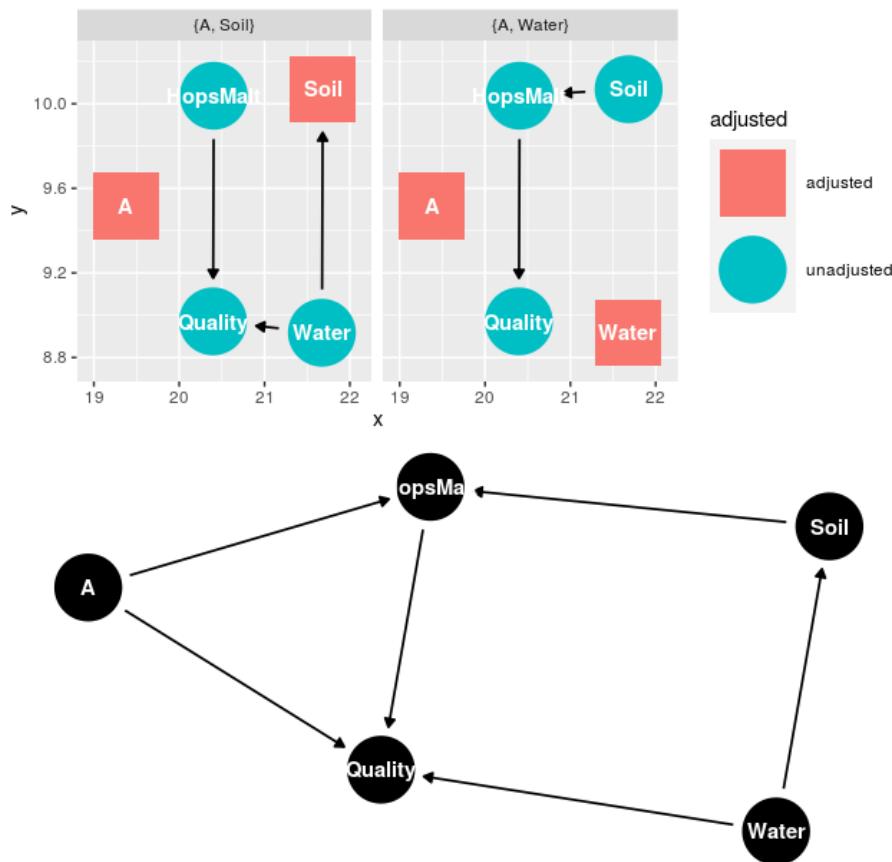
p2 <- ggdag_adjustment_set(dagified, exposure="HopsMalt", outcome="Quality")
multiplot(p1,p2)
adjustmentSets(dagified)
# { Water }
# { Soil }
  
```

Um deles requer condicionar à água e outro requer condicionar ao solo. Os grafos plotados indicam o fluxo de informação em cada conjunto de ajustes. Em geral, reportamos os valores para uma das vias ou ambos os caminhos.

Podemos imaginar um novo fator, que interfere na pureza final e também na dos ingredientes sólidos.

```
dagified2 <- dagify(Quality ~ Water + HopsMalt + A,
                     HopsMalt ~ Soil + A, Soil ~ Water,
                     exposure = "HopsMalt",
                     outcome = "Quality")

p3 <- ggdag(dagified2) + theme_dag_blank()
p4 <- ggdag_adjustment_set(dagified2,exposure="HopsMalt",outcome="Quality")
multiplot(p4,p3)
```



Agora, deveríamos testar as seguintes condições:

```
beer_data2 <- simulateSEM(dagified2,b.lower = 0.20,b.upper=0.25)
#A _/_ Soil
lm(A ~ Soil,beer_data2)
#A _/_ Water
```

```
lm(A ~ Water,beer_data2)
#HopsMalt _/_ Water / Soil
lm(HopsMalt ~ Water + Soil,beer_data2)
#Quality _/_ Soil / A, HopsMalt, Water
lm(Quality ~ Soil + A,beer_data2)
```

As possibilidades de ajuste para estimativa não enviesada são:

```
adjustmentSets(dagified2)
# { A, Water }
lm(Quality ~ HopsMalt + Water + A,beer_data2)
# { A, Soil }
lm(Quality ~ HopsMalt + Soil + A,beer_data2)
```

Mediação e Moderação Mediação

Uma ideia curiosa é de que uma variável pode estar intermediando a ação de outra sobre um desfecho. Um exemplo clássico é a relação entre o hábito de fumar e câncer. Sabemos que existe uma ação nociva pela temperatura do ar inalado, assim como dos componentes químicos absorvidos.

Em modelos de mediação, tentamos quantificar a porção que é explicada por uma variável intermediária. Para tanto, empregamos o seguinte procedimento:

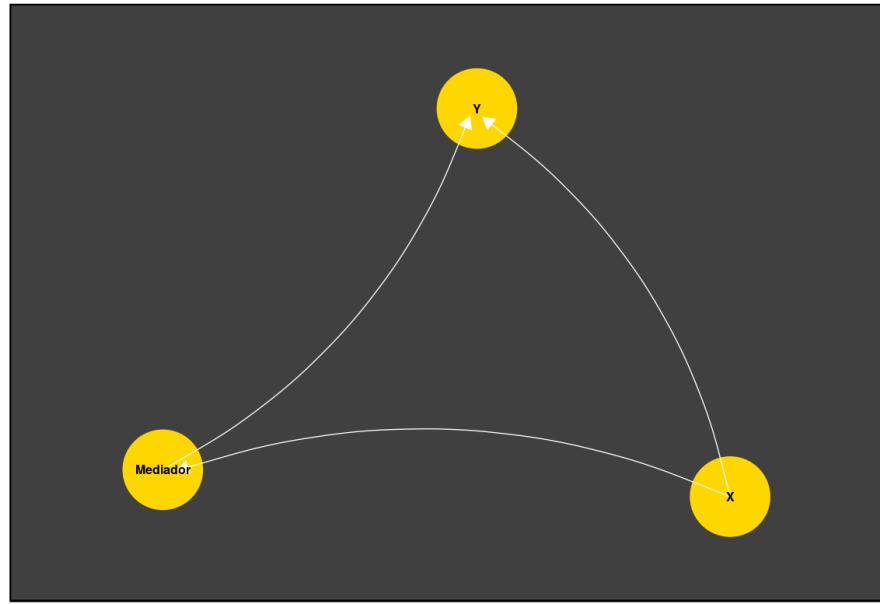
1. Verificar plausibilidade de relações individualmente através modelos de regressão entre variáveis de interesse. Ajustamos 3 modelos:
 - (1) variável independente e variável alvo ($Y \sim X_1\beta_1$),
 - (2) variável mediadora e variável alvo ($Y \sim X_2\beta_2$),
 - (3) variável independente e variável mediadora ($X_2 \sim X_1\beta_3$).

O efeito direto da variável independente sobre a variável alvo é quantificado β_1 .

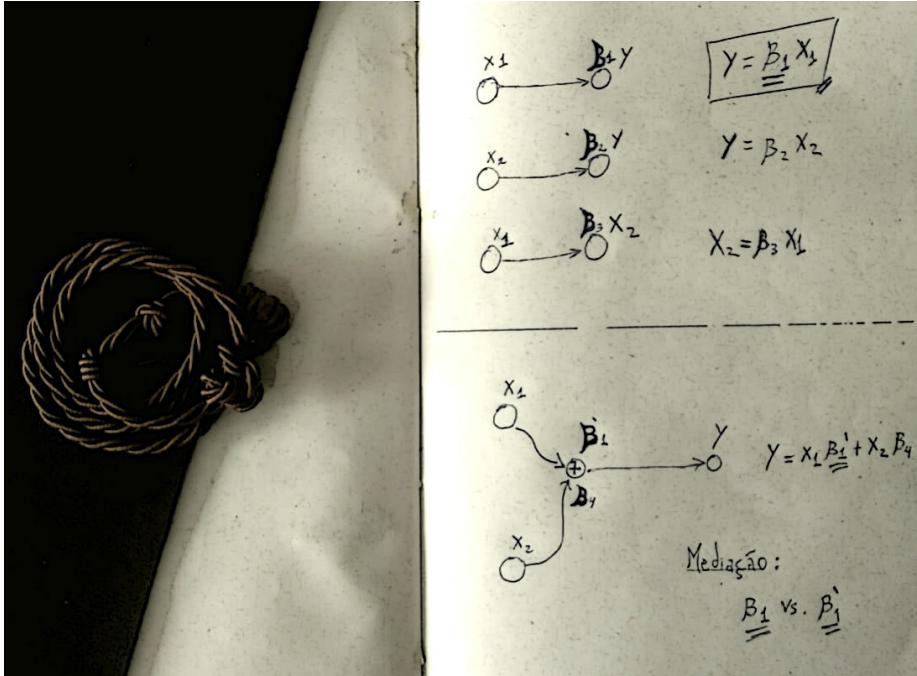
2. Verificar mudanças obtidas pela introdução da variável mediadora.
Ajustamos um quarto modelo (4), com a combinação linear de variável independente e variável mediadora. Observamos então a diferença entre o novo (β'_1) coeficiente de X_1 e o antigo (β_1) $Y = X_1\beta'_1 + X_2\beta_4$.

Caso exista mediação, espera-se que o coeficiente β'_1 seja não significativo ou que possua magnitude bastante reduzida em relação ao coeficiente do efeito direto β_1 .

Seguindo o exemplo sugerido, espera-se que exista uma relação entre hábito de fumar e câncer. Ainda, espera-se que a inclusão de um mediador (e.g. concentração de nicotina) explique parte do efeito, reduzindo o coeficiente de X_1 . O diagrama a seguir expressa a ideia contida no processo desejado.



O diagrama abaixo ilustra passos rigorosamente. As 3 regressões para checar premissas estão na sessão superior e a regressão múltipla no setor inferior. Foram suprimidos termos de erro. Estimativas para a relação entre X_1 e Y são $\hat{\beta}_1$ e $\hat{\beta}_1'$ grifados nas equações. O comportamento desses parâmetros define as conclusões sobre o modelo de mediação.



Não há garantias de que os sistemas reais se comportarão seguindo os parâmetros estimados. Usamos regressão múltipla para estimar o efeito parcial atribuído aos medidores, porém a retirada desses fatores no fenômeno real pode resultar em alterações no sistema não previstas pelo modelo.

A certeza dependeria de uma descrição bastante acurada do fenômeno pelas regressões ($R^2 \sim 1$), o que raramente é verificado fora de fenômenos físicos mais simples.

Portanto, é recomendável que ajustes sejam feitos na fase experimental. Em nosso exemplo, isso implicaria em controlar a concentração de nicotina absorvida *in vivo*. Obviamente, razões éticas e limitação de recursos precluem muitas vezes a manipulação direta do objeto de estudo. Métodos tais como o descrito, ainda que frágeis, permitem estudar interações e relações causais. Entretanto, é necessário atenção aumentada ao fazer conclusões e, especialmente, ao traduzí-las para linguagem natural.

Em R:

```
>fit_yx1 <- lm(y ~ x1, data)
>fit_yx2 <- lm(y ~ x2, data)
# Mediation
>fit_yx1x2 <- lm(y ~ y1 + y2)
```

```

>summary(fit_yx1)
(...)
>summary(fit_yx2)
(...)
>summary(fit_yx1x2)
(...)

```

A diferença numérica entre valores de β_{x_1} é a magnitude do efeito indireto (*Ind. Effect*). Podemos usar uma estimativa do erro padrão para derivar uma estatística t e um valor p associados (teste de Sobel). Usando libs do CRAN: Usando o dataset `bh1996`, com medidas sobre liderança, bem-estar e horas de trabalho.

A pergunta é: clima de liderança media relação entre horas de trabalho e bem-estar?

```

>library(bda)
>library(multilevel) # dataset bh1996
>data(bh1996)

# LEAD : Clima de liderança
# WBEING : Bem-estar
# HRS : Horas de trabalho

>sobel(pred=bh1996$HRS,med=bh1996$LEAD,out=bh1996$WBEING)
$`Mod1: Y~X`
      Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 3.51693620 0.052902697 66.47934 0.000000e+00
pred        -0.06523285 0.004590274 -14.21110 3.078129e-45

$`Mod2: Y~X+M`
      Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 1.86832973 0.06413083 29.13310 1.024201e-176
pred        -0.04311316 0.00421918 -10.21837 2.382257e-24
med         0.48386196 0.01242129 38.95426 4.967825e-302

$`Mod3: M~X`
      Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 3.40718349 0.045154735 75.45573 0.000000e+00
pred        -0.04571488 0.003917997 -11.66792 3.488366e-31

$Indirect.Effect
[1] -0.02211969
$SE
[1] 0.001978985
(...)

>mediation.test(iv = bh1996$HRS,mv = bh1996$LEAD,dv = bh1996$WBEING)

```

```
Sobel      Aroian      Goodman
z.value -1.117729e+01 -1.117391e+01 -1.118067e+01
p.value 5.267356e-29 5.471647e-29 5.070460e-29
# Aroian e Goodman são outros testes para o parâmetro de efeito indireto
```

Moderação e Interações

Modelos incluindo termos de moderação são aqueles que incluem **interação** entre variáveis. Usando o jargão de inferência causal, é o mesmo que modificador de efeito (*effect-modifier*). Como discutimos antes, a relação entre hábito de fumar e câncer pode ser explicada por fatores intermediários, como a concentração de nicotina e presença de variantes genéticas de risco.

Podemos supor que a concentração de nicotina inalada diariamente tenha um efeito independente. Igualmente, uma configuração genética tem efeito causal por si.

$$Risk = Nicotina * \beta_1 + Genes_{(+)}\beta_2$$

Em moderação, adicionamos um termo à nossa combinação linear. É um coeficiente para a multiplicação entre variáveis independentes.

$$Risk = Nicotina * \beta_1 + Genes_{(+)}\beta_2 + Nicotina * Genes_{(+)}\beta_3$$

Será que *fumar e ter genes de risco* é diferente da combinação do efeito de ambos em separado?

Esse é um dos poucos casos em que é mais fácil observar o aspecto algébrico antes. Estamos multiplicando os valores de preditores X_1 e X_2 . Se ambos tiverem mesmo sentido (+ ou -), a interação terá efeito positivo. Caso contrário, negativo. Ainda, vemos que as magnitudes são multiplicadas. O coeficiente β_3 quantifica essa multiplicação em relação ao efeito em y , seja alterando o sentido (β_3 negativo) ou escalando o valor absoluto.

$$y = X_1 * \beta_1 + X_2 * \beta_2 + X_1 X_2 \beta_3$$

A relação de y em relação com cada preditor deixa de ser linear. Como podemos verificar analisando as derivadas parciais. Para $\frac{d}{dx_1}$:

$$\frac{d}{dx_1}(y) = \frac{d}{dx_1}(x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3)$$

O segundo termo não depende de X_1 , então:

$$\frac{d}{dx_1}(y) = \frac{d}{dx_1}(\beta_1 + x_2\beta_3)$$

A inclinação (*slope*), que antes era uma constante (linha reta) β_1 passa a ter um termo somado, que é a multiplicação da constante estimada β_3 pelo valor de x_2 . Então temos inclinação diferente para cada valor de moderador!

Esses detalhes tornam a interpretabilidade dos coeficientes difícil. Normalmente, são usadas heurísticas, como centralizar os dados em torno da média, para simplificar o contexto.

Colinearidade Verificar se há colinearidade (relação linear) entre variáveis preditoras. Se as variáveis preditoras são altamente correlacionadas, é possível que estejamos fornecendo informações redundantes ao modelo, o que é nocivo. Existem alguns indicadores que podem ajudar a tomar essa decisão.

Comumente, observamos o VIF *Variance inflation factor*.

VIF

A intuição aqui é de que se as variáveis são muito relacionadas $X_1 \sim X_2$, os valores de β estimados em $Y = \beta_1 X_1 + \beta_2 X_2 + \dots$ não serão únicos. Por exemplo, poderíamos trocar β_1 por β_2 e a solução permaneceria praticamente inalterada. O VIF estima a colinearidade em relação à combinação de outros preditores usados.

Para calcular o VIF referente a um preditor X' , ajustamos uma nova regressão, em que a variável resposta é X' e as preditoras são as outras variáveis preditoras. O VIF é dado por: $\frac{1}{1-R^2}$, sendo R^2 o coeficiente de determinação da regressão, como calculamos antes.

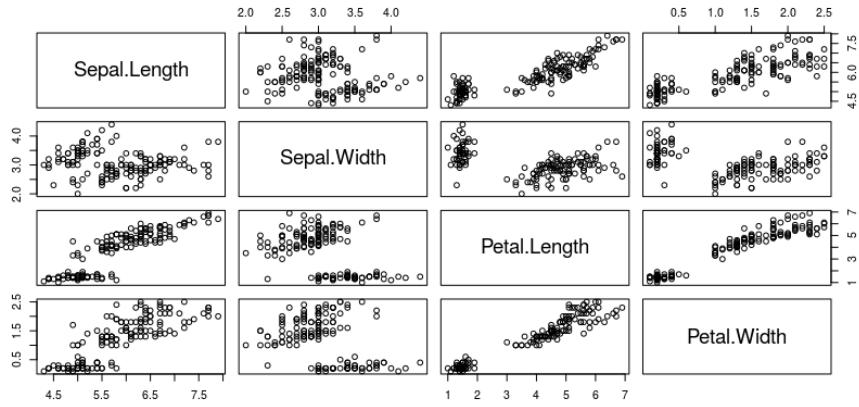
Valores de VIF altos refletem valores de R^2 altos, isto é: a combinação linear de outras variáveis explicaria muito bem a variável preditora em questão. Não há regra canônica, porém $VIF > 10$ ($R^2 = 0.9$) e $VIF > 5$ ($R^2 = 0.8$) são citados como fronteiras indicando colinearidade inaceitável.

A função `vif` do pacote `car` implementa o procedimento. Ajustamos uma regressão linear múltipla para o comprimento das sépalas no dataset `iris` a partir de outras 3 variáveis. Podemos verificar que há colinearidade ($VIF_{pet.leng.} \sim 15.1$, $VIF_{pet.wid.} \sim 14.2$) entre largura e comprimento da pétala. Por outro lado, a colinearidade com o comprimento da sépala é baixa ($VIF_{sep.wid.} \sim 1.3$).

```
> car::vif(lm(Sepal.Length ~ Petal.Length + Petal.Width + Sepal.Width,
  data=iris))
      Petal.Length  Petal.Width  Sepal.Width
  15.097572     14.234335    1.270815
```

Se há colinearidade, é recomendado remover um dos preditores para eliminar a redundância. Como sempre, a inspeção visual ajuda.

```
> pairs(iris[,1:4])
```



Como podemos ver, usar duas variáveis preditoras (regressão múltipla) não colineares aumenta a performance do modelo em relação à regressão simples ($R^2 \sim 0.84$ vs $R^2 = 0.76$).

```
>lm(Sepal.Length ~ Petal.Length,
+     data=iris) %>% summary

(...)

Multiple R-squared:  0.76, Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16

>lm(Sepal.Length ~ Petal.Length + Sepal.Width,
+     data=iris) %>% summary
(...)

Multiple R-squared:  0.8402,   Adjusted R-squared:  0.838
F-statistic: 386.4 on 2 and 147 DF,  p-value: < 2.2e-16
```

Medidas latentes e análise fatorial Considere o problema de medir algo inacessível através de meios secundários.

Por exemplo, o conceito de *qualidade de vida* é facilmente concebível, apesar de não estar atrelado a uma medida tangível, tal qual *altura* ou *tamanho do fêmur*. Uma série de métodos foi desenvolvida para lidar com a tarefa de estimar *variáveis latentes*. Em especial, esses modelos são muito populares entre psicometristas. Podemos aplicar modelos de variáveis latentes para muitos contextos.

Isso é feito quando usamos respostas corretas em um teste formulado por especialistas para quantificar uma habilidade. A *Teoria de Resposta ao Item* é usada em testes como ENEM (Brasil), SAT e GRE (EUA). Relacionamos a estimativa de habilidade (θ) com a probabilidade de acertar (1) ou errar (0).

Traços de personalidade também podem ser estudados dessa maneira. Podemos atribuir um grau de extroversão F de uma pessoa através de sua pontuação em uma bateria de testes X_1, X_2, X_3, \dots relacionados a esse atributo.

Sejam os items:

1. Gosto de estar com outras pessoas (1 a 7)
2. Costumo conversar com desconhecidos (1 a 7)
3. Costumo expressar minhas opiniões (1 a 7)
4. Sou considerado(a) uma pessoa comunicativa (1 a 7)

A pontuação de um indivíduo será uma sequencia de 4 números. Um indivíduo muito extrovertido pode pontuar (7,7,6,7) e um introvertido (2,3,2,1). Podemos pensar que a série de medidas é influenciada por um construto (extroversão).

Análise fatorial parte da premissa de que a **covariância** nas medidas diretas é fruto das **influências latentes compartilhadas** pelos items. Assim, podemos estimar um parâmetro λ para a relação entre cada item e o traço latente F . Para isso, usaremos a matriz de covariâncias.

Os valores de λ quantificam a relação entre items e fatores latentes e servem, por exemplo, para selecionar items mais relacionados aos traços alvo em um instrumento psicométrico.

Como na regressão linear, o modelo descreve cada medida como uma combinação entre score individual para fator latente F multiplicado pelo peso para o item λ_{Item1} e erros.

A medida do item 1 para o i -ésimo sujeito considerando n fatores latentes F_n é:

$$x_{1,i} = \sum_1^n nF_i\lambda_n + \epsilon$$

Assim, falamos em mais de um construto latente. Ao invés de trabalhar com um grande fator latente (extroversão), podemos ligar os quatro items anteriores a dois conceitos menos específicos: “sociabilidade” e “expressividade”.

O valor dos 4 itens para o n -ésimo sujeito, considerando dois fatores latentes, com pesos λ_i, λ'_i é:

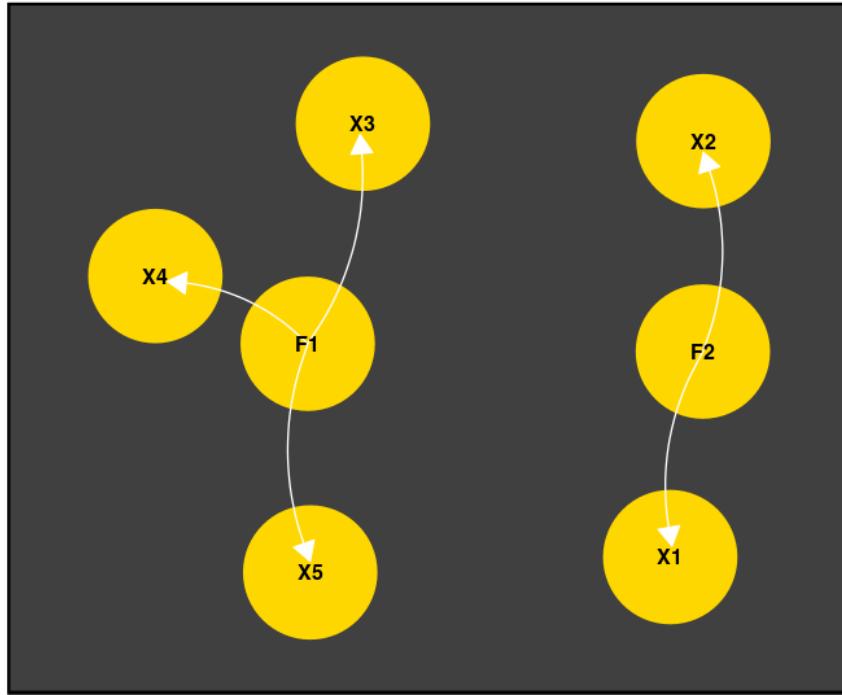
$$x_{1,n} = F_{1,n} \lambda_1 + F_{2,n} \lambda'_1 + \epsilon$$

$$x_{2,n} = F_{1,n} \lambda_2 + F_{2,n} \lambda'_2 + \epsilon$$

$$x_{3,n} = F_{1,n} \lambda_3 + F_{2,n} \lambda'_3 + \epsilon$$

$$x_{4,n} = F_{1,n} \lambda_4 + F_{2,n} \lambda'_4 + \epsilon$$

Podemos perceber que a matriz Λ terá 8 elementos, com 4 pesos para o fator F_1 e 4 pesos para o fator F_2 . Sabendo os dois scores latentes de cada sujeito, seria possível reconstruir as observações com algum grau de perda. Perceba que expressamos qualquer item com apenas dois parâmetros (F_1 e F_2). As informações em nosso dataset poderiam então ser condensadas de $[nx4]$ dimensões para $[nx2]$.



Para estimar os parâmetros acima, supomos que a variância de **cada item** possui uma variância intrínseca e uma variância compartilhada, que é determinada pelos fatores latentes. Usamos uma matriz de covariâncias entre os items para estimar os pesos dos fatores latentes. Além disso, estimamos

parâmetros relacionados à diagonal da matriz (variâncias).
Em nosso exemplo, teríamos uma matriz de dimensão $[4 \times 4]$.

$$CovMat_x = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Como vimos no capítulo 2, cada valor é dado por:

$$Cov(X, X') = \sum_{i=1}^N (x_i - \mu_x)(x'_i - \mu_{x'})$$

A diagonal reflete a covariância de uma variável com ela mesma, a variância:

$$\begin{aligned} Cov(X, X) &= \sum_{i=1}^N (x_i - \mu_x)(x_i - \mu_x) \\ &= \sum_{i=1}^N (x_i - \mu_x)^2 \\ &= Var(X) \end{aligned}$$

Por exemplo, a matriz de covariâncias para o *iris*:

```
> cov(iris[, 1:4])
   Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935 -0.0424340  1.2743154  0.5162707
Sepal.Width     -0.0424340  0.1899794 -0.3296564 -0.1216394
Petal.Length    1.2743154 -0.3296564  3.1162779  1.2956094
Petal.Width     0.5162707 -0.1216394  1.2956094  0.5810063
> var(iris[, 1])
[1] 0.6856935
```

Usando notação matricial, seja X uma matriz com $m = 4$ colunas de $n = 150$ observações, a matriz de covariância $Cov_{4 \times 4}$ é:

$$Cov(X') = X'^T X' \frac{1}{n} = X'^T X' \frac{1}{150}$$

X' é a matriz cujos valores foram centralizados pela média $x' = x - \mu$. Assim o produto de X pela transposta retorna em cada elemento x_{ij} o valor $\sum_i^n (x_i - \mu_i)(x_j - \mu_j)$. Fácil implementar manualmente:

```

> iris2$Sepal.Length <- iris$Sepal.Length - mean(iris$Sepal.Length)
> iris2$Sepal.Width <- iris$Sepal.Width - mean(iris$Sepal.Width)
> iris2$Petal.Length <- iris$Petal.Length - mean(iris$Petal.Length)
> iris2$Petal.Width <- iris$Petal.Width - mean(iris$Petal.Width)
> (t(as.matrix(iris2[,1:4])) %*% as.matrix(iris2[,1:4]))*1/150
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  0.68112222 -0.04215111   1.2658200  0.5128289
Sepal.Width   -0.04215111  0.18871289   -0.3274587 -0.1208284
Petal.Length   1.26582000 -0.32745867   3.0955027  1.2869720
Petal.Width    0.51282889 -0.12082844   1.2869720  0.5771329

```

Com base nos princípios delineados, a solução desejada por nós é tal que:

1. A covariância entre medidas é explicada por combinações de variáveis latentes compartilhadas.
2. Os dados serão explicados por uma matriz de rank mais baixa. Em nosso caso: $\Lambda_{[n \times m]}$, $m < 4$.
3. Para cada observação, teremos um valor de score latente F_i para cada fator. O valor final de um item é dado pela contribuição individual de cada fator mais uma variância individual. Como vimos:

$$x_{1,i} = \sum_1 nF_i\lambda_n + \epsilon$$

4. Cada fator possui uma variância intrínseca, a qual estimaremos somando uma matriz diagonal ψ à nossa matriz de pesos.

Estimamos os parâmetros para maximizar as probabilidades (*Max. Likelihood*) dos valores observados em X dadas as equações.

$$L(X^T X \frac{1}{n} | \Lambda, \psi)$$

Determinamos a função de custo conhecendo Λ e ψ :

$$C \sim \Lambda \Lambda^T + \psi$$

Em que ψ é uma matriz diagonal com mesmo rank que Λ . Como vimos anteriormente, a diagonal contém as variâncias, então os parâmetros em ψ regulam a porção de variância dos items governadas por fatores λ . Dizemos que a diagonal em $\Lambda \Lambda^T$ contém as **communalities** (variância intrínseca).

O processo de otimização para minimizar erros é mais complexo que o da regressão linear. Os estimadores possíveis aqui são muitos, nenhum deles com solução analítica simples ou garantia de convergência.

Semelhanças entre técnicas de redução de dimensões: EFA, PCA probabilístico, PCA, Autoencoder. Podemos levar em consideração a solução anterior sem uma matriz diagonal ψ atrelada:

$$Cov \sim \Lambda\Lambda^T$$

Essa formulação é equivalente à análise de componente principal (Principal Component Analysis, PCA). Aqui, nossos pesos estimarão também a variância intrínseca. É um método computacionalmente barato para reduzir dimensões preservando informações. Matematicamente, a diferença entre PCA e EFA está no fato de o segundo estimar separadamente parâmetros para covariância compartilhada e variância individual. Uma técnica ‘intermediária’ pouco conhecida é o PCA probabilístico (PPCA), em que levamos em conta uma matriz diagonal mais simples.

$$Cov \sim \Lambda\Lambda^T + \sigma^2 I$$

Isto é: uma matriz identidade com ruído introduzido através de apenas um parâmetro (σ^2).

Uma curiosidade é que a diagonal acaba influindo menos com o aumento do rank das matrizes. Então, o resultado das técnicas acima converge em situações com alta dimensionalidade ($n \rightarrow \infty$). Uma discussão mais completa pode ser conferida em outro lugar (ver referências).

Em sumário:

$$\text{PCA} : Cov \sim \Lambda\Lambda^T$$

$$\text{PPCA} : Cov \sim \Lambda\Lambda^T + \sigma^2 I$$

$$\text{EFA} : Cov \sim \Lambda\Lambda^T + \psi$$

(Aqui, usamos \sim não para denominar semelhança, mas sim que maximizaremos o likelihood de Cov com uma expressão em função dos termos à direita)

Ainda, redes neurais do tipo *autoencoder* possuem formulação semelhante. Especificamente, um autoencoder com uma camada interna e certas restrições na função de ativação é idêntico ao PCA. Entretanto, podemos usar **mais** dimensões que o input, além de múltiplas camadas e funções não-lineares. Dessa maneira, incrementamos o poder do modelo gerativo, assim como ficamos mais vulneráveis a sobreajuste.

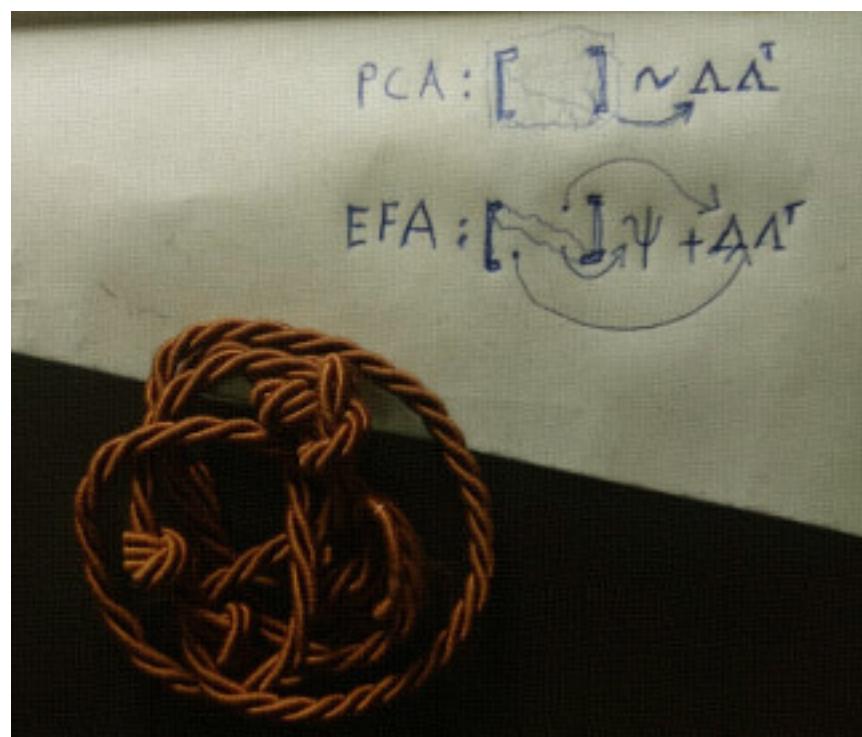


Figure 17: .

Voltaremos ao assunto quando o foco for modelos de ambiente, compressão de informação, modelos gerativos e redução de dimensões.

Número de fatores

Não tocamos em um ponto crucial: qual o número ótimo de fatores? É melhor um modelo que leve em conta *extroversão* ou um que use *sociabilidade* e *expressividade*?

Podemos explicar a covariância usando um número arbitrário de fatores latentes. A tendência é observarmos melhora nos indicadores de performance sob a pena de saturação (e.g. sobreajuste, interpretabilidade difícil). Existem procedimentos estabelecidos para balancear o poder explicativo com simplicidade do modelo.

Em geral, busca-se um número mínimo de fatores que maximize o poder de explicação. Considerando graus de liberdade (df) e erros do modelo (estatística X^2), dois índices populares são o RMSEA e o CFI. Assim como no cálculo de R^2 , o racional é dimensionar erros, porém aqui penalizamos a quantidade de parâmetros.

Outra métrica bastante utilizada é observar a influência de cada fator sobre a matriz de covariância.

Ao multiplicarmos um vetor por uma matriz, mudamos sua magnitude e sua direção.

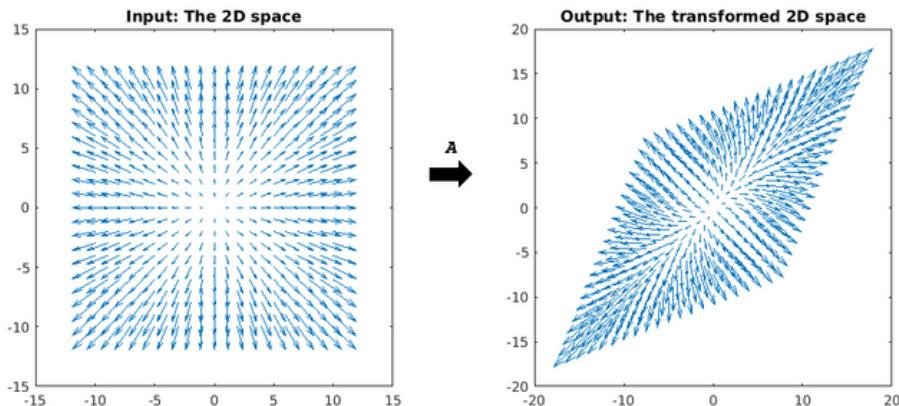


Figure 18: Efeito de multiplicação entre vetores e uma matriz A

Os vetores alinhados com a matriz (e.g. aqueles na diagonal da transformação acima) apenas mudam de tamanho após a transformação.

São os autovetores da matriz.

Uma das formas de extração de fatores é através dos eixos principais. Neste método, decomponemos a matriz original em vetores ortogonais multiplicados

por escalares (autodecomposição, *eigen/spectral decompositon*): autovalores e autovetores (eixos).

Em geral, os primeiros eixos têm maior autovalores. Existem diversas heurísticas recomendando métodos para escolher números de fatores pelo tamanho dos autovalores. Uma delas é considerar apenas autovalores maiores que 1. Outra é considerar o ponto da curva em que há um aparente ponto de descontinuação (“joelho”).

É razoável pensar que autovetores associados a autovalores altos capturam muita informação sobre variância (individual e compartilhada) dos items.

Análise factorial confirmatória

Os processos descritos acima são exploratórios por natureza. Buscamos o melhor ajuste para fatores latentes sem antes determinar uma estrutura. É um bom procedimento para redução de dimensões e compressão de informação, porém, se desejamos interpretabilidade e validade científica, há alguns pontos sensíveis.

Pensando na elaboração de uma escala para medir um traço de personalidade, retomamos o argumento de Popper (capítulo 2) contra o indutivismo. É desejável que tenhamos um modelo prévio e hipóteses testáveis de antemão. Do contrário, é fácil encontrar um modelo oferecendo bom ajuste em quase qualquer caso.

Na análise factorial confirmatória, fazemos uma restrição direta ao modelo. Os parâmetros são pré-determinados com base em diagrama (grafo) expresso por quem conduz a análise. Assim, podemos especificar uma relações. No diagrama acima, o primeiro fator latente possui cargas nas relações com items X_3, X_4, X_5 e o segundo fator com X_1, X_2 .

Nesse caso, os estimadores serão um pouco mais complexos.

Equações estruturais Equações estruturais são o *framework* abrangendo quaisquer dos modelos anteriores, incluindo topologias de grafos e relações arbitrárias (e.g. não paramétrica/probabilísticas).

Assim, podemos desenhar um diagrama de relações entre entidades, declarar relações entre medidas e testar adequação do modelo. Como vimos, Judea Pearl costurou esses métodos quantitativos com uma base filosófica coerente, fazendo uso dos conceitos de contrafactual e testagem de hipóteses.

Esses modelos são úteis em muitos campos para descrever estatisticamente relações de elementos múltiplos num sistemas complexo. Como sempre, devemos ter cuidado com a flexibilidade do modelo. Em especial, alguns procedimentos recomendados são de difícil conciliação com uma base hipotético-dedutiva (e.g. mudança ad-hoc do modelo após observação de índices de modificação).

Aplicações Os grande cinco (Big Five) traços de personalidade são construtos consistentemente encontrados na busca por fatores latentes. Eles são: agradabilidade, neuroticismo, abertura a experiencias, conscienciosidade, extroversão.

Usaremos dados do <https://openpsychometrics.org/>. O dataset BIG5 tem dados demográficos (idade, gênero, país) e 50 medições em items do International Personality Item Pool. O tamanho amostral é de 19,719. Faremos análise factorial exploratória e confirmatória através dos pacotes **psych**, **sem** e **lavaan**.

```
>system("wget http://openpsychometrics.org/_rawdata/BIG5.zip")
(...)

Resolving openpsychometrics.org (openpsychometrics.org)... 69.164.197.103
Connecting to openpsychometrics.org (openpsychometrics.org)|69.164.197.103|:80... connected
(...)

Saving to: 'BIG5.zip'
(...)

2019-02-04 09:09:39 (624 KB/s) - 'BIG5.zip' saved [523351/523351]
> system("unzip BIG5.zip")
Archive: BIG5.zip
inflating: BIG5/codebook.txt
inflating: BIG5/data.csv

>library(psych)
>library(lavaan)
>library(sem)
>bigf_data <- read.csv("BIG5/data.csv",sep = "\t")
>names(bigf_data)
[1] "race"      "age"       "engnat"    "gender"    "hand"
[6] "source"    "country"   "E1"        "E2"        "E3"
[11] "E4"        "E5"        "E6"        "E7"        "E8"
[16] "E9"        "E10"       "N1"        "N2"        "N3"
[21] "N4"        "N5"        "N6"        "N7"        "N8"
[26] "N9"        "N10"       "A1"        "A2"        "A3"
[31] "A4"        "A5"        "A6"        "A7"        "A8"
[36] "A9"        "A10"       "C1"        "C2"        "C3"
[41] "C4"        "C5"        "C6"        "C7"        "C8"
[46] "C9"        "C10"       "O1"        "O2"        "O3"
[51] "O4"        "O5"        "O6"        "O7"        "O8"
[56] "O9"        "O10"
```

Vamos verificar o que acontece se ajustarmos um modelo com 5 fatores latentes:

```
>library(lavaan)
>library(psych)
>efa_big <- fa(bigf_data[,8:57],nfactors = 5)
>efa_big
(..)
```

```
RMSEA index = 0.055 and the 90 % confidence intervals are 0.054 0.055
```

Observamos um valor baixo de RMSEA, o que indica baixa magnitude de erros por grau de liberdade. É interessante notar que não termos indicamos quais items avaliam quais fatores (e.g. Items O1 e O2 estão atrelados à abertura a experiência). Se as premissas estiverem corretas, para cada item, a solução encontrada deve indicar alta carga em um fator e baixa em outros.

É o que se verificar. Selecionando as estimativas para três items de três grupos. O fator com maior carga está marcado com um asterisco.

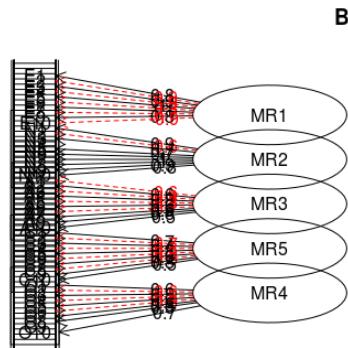
```
(...)
Factor Analysis using method = minres
Call: fa(r = bigf_data[, 8:57], nfactors = 5)
Standardized loadings (pattern matrix) based upon correlation matrix
      MR1    MR2    MR3    MR5    MR4    h2    u2  com
E1   0.69*   0.04 -0.03 -0.01  0.00  0.46  0.54 1.0
E2  -0.70*  -0.08 -0.04  0.04  0.00  0.48  0.52 1.0
E3   0.63*  -0.17  0.16  0.09 -0.06  0.57  0.43 1.3
(...)
N1  -0.06   0.69*   0.10  0.05 -0.05  0.49  0.51 1.1
N2   0.07  -0.50*  -0.01 -0.09  0.05  0.26  0.74 1.1
N3  -0.12   0.61*   0.20  0.10  0.01  0.43  0.57 1.3
(...)
A1   0.05   0.09 -0.44*   0.02 -0.07  0.20  0.80 1.2
A2   0.28  -0.04   0.50*  -0.05  0.06  0.41  0.59 1.6
A3   0.17   0.27 -0.41*  -0.15  0.10  0.27  0.73 2.6
(...)
```

Extraindo a solução:

```
>efa_bigst <- structure.sem(efa_big)
>efa_bigst
      Path      Parameter Value
[1,] "MR1->E1"    "F1E1"     NA
[2,] "MR1->E2"    "F1E2"     NA
[3,] "MR1->E3"    "F1E3"     NA
[4,] "MR1->E4"    "F1E4"     NA
```

Temos os fatores (MR) e os nodos aos quais eles estão conectados, descartando aqueles de menor magnitude/significância. Podemos ajustar um modelo confirmatório a partir dessas especificações:

```
>big_sem <- sem(efa_bigst,S = cov(bigf_data[,8:57]),N = 19719)
>summary(big_sem)
(...)
>sem.diagram(big_sem,main = "Big Five",e.size=0.05)
```



O pacote lavaan permite especificar uma família maior de modelos e é bastante popular para SEM em R. A sintaxe é:

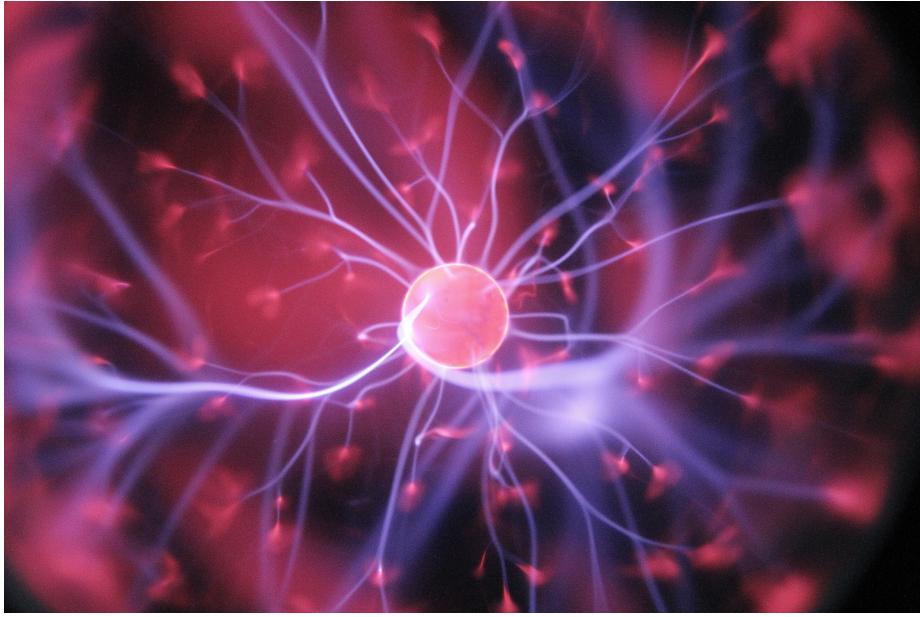
```
>model <- c('
  F1 =~ X1 + X2 + X3
  F2 =~ Y1 + Y2 + Y3')
>lavaan (model,data,...)
```

Referências Wright, S. (1921). “Correlation and causation”. J. Agricultural Research. 20: 557–585.

<https://stats.stackexchange.com/questions/123063/is-there-any-good-reason-to-use-pca-instead-of-efa-also-can-pca-be-a-substitut> <https://steemit.com/steemstem/@dexterdev/linear-transformations-a-20-sbd-coding-contest-announcement>

Exercícios

1. Examine o VIF da regressão múltipla usada no processo de mediação com banco de dados *bh1996*.
 - Há colinearidade entre mediador e preditor principal?
2. Examine a mudança de performance (e.g. R^2) após inclusão do mediador no modelo.
 - Se a variável mediadora M explicar as mesmas vias causais que a variável preditora X_1 , é esperado que essa mudança seja grande? Discuta.
3. Usando dados à sua escolha:
 - Ajuste uma regressão linear simples
 - Adicione outro preditor (regressão linear múltipla)
 - Verifique se há colinearidade
 - Cheque outras premissas observando o material auxiliar **/aux** (e.g. independência dos erros com Durbin-Watson)
 - Teste uma relação de mediação usando 3 variáveis
4. Usando os dados *iris*:
 - Escolha duas medidas corelacionadas e verifique se a espécie *modera* a relação entre elas. Lembre-se: você deve adicionar um termo de interação `var1*var` na fórmula da regressão.
 - Execute **(1)** análise de componentes principais (PCA) e **(2)** análise fatorial exploratória (EFA) para as variáveis numéricas.
 - Extraia **(1)** a projeção de cada observação nos dois primeiros componentes, PC_1 , PC_2 , e **(2)** o score gerado a partir de cada fator. A função `princomp` retorna um objeto acessível `$scores`.
 - Verifique a correlação entre ambos.



Capítulo 4 : Neurônios

Em março de 2016, o software AlphaGo venceu um mestre de Go. Inventado há mais de 2,500 anos, o jogo motivou avanços em matemática. Existem $2,08 \times 10^{170}$ maneiras válidas de dispor as peças no tabuleiro. O polímata chinês Shen Kuo (1031–1095) chegou a um resultado próximo 10^{172} séculos atrás. Vale lembrar que o número de átomos no universo observável é de módicos 10^{80} .

No capítulo anterior, aprendemos formulações básicas de modelo preditivo com regressão. Aqui, conhceremos a primeira máquina inteligente da história implementando um *perceptron*. Ele é capaz de lidar com mais dimensões (e.g. processamento de imagens). Estimadores com solução fechada não existem como na regressão linear, então usamos informações locais para ‘caminhar’ (*gradient descent*) em direção a um mínimo.

Estenderemos nossa caixa de ferramentas para abranger relações mais complexas, não lineares. Encadeando neurônios simples, podemos aprender sinais complexos sem apelar para funções complexas, intratáveis ou demasiadamente flexíveis.

O perceptron de Rosenblatt

Frank Rosenblatt (1928 - 1971) nasceu e morreu em 11 de julho, mas esse não é o fato mais curioso da biografia deste psicólogo. Foi o responsável pelo desenvolvimento do primeiro neurônio artificial. Em suas palavras, o primeiro objeto não biológico a recriar uma organização do ambiente externo com significado.

It can tell the difference between a cat and a dog, although it wouldn't be able to tell whether the dog was to the left or right of the cat. Right now it is of no practical use, Dr. Rosenblatt conceded, but he said that one day it might be useful to send one into outer space to take in impressions for us. - New Yorker, December, 1958²⁴

O aparato reproduzia o entendimento da época sobre o funcionamento de um neurônio. O corpo recebe sinais de dendritos e, após processamentos ocultos, produz um output na forma de sinal elétrico pelo axônio. A primeira matematização viria do modelo de McCulloch & Pitts (“A Logical Calculus of the Ideas Immanent in Nervous Activity”, 1943).

Em 1949, Donald Hebb descreveu em seu clássico *The Organization of Behavior* um mecanismo plausível para a aprendizagem. Comumente expressa na máxima “Cells that fire together wire together” (células que disparam juntas, conectam-se entre si).

Com o objetivo de criar uma máquina que pudesse processar inputs diretamente do ambiente físico (e.g. luz e som), Rosenblatt concebeu uma extensão elegante do modelo em 1957 (“The Perceptron[*do latim, percipio, compreender*] – a perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory”). Composto de três partes: o sistema S (sensório); o sistema A (associação) e o sistema R (resposta).

O neurônio “lógico” criado de McCulloch & Pitts foi modificado de maneira a processar inputs através de pesos antes da saída. A aprendizagem se dá pela modificação desses pesos.

Inicialmente, o perceptron foi simulado em um IBM 704 (também berço das linguagens FORTRAN e LISP). Em seguida, implementado como um dispositivo físico, batizado de Mark I Perceptron.²⁵ Um estudo mais profundo foi publicado por ele em 1962 (*Principles of neurodynamics*).

²⁴Ele consegue diferenciar um gato de um cachorro, ainda que não seja capaz de dizer se o cachorro estava à esquerda ou à direita do gato. No momento, não tem uso prático, Dr. Rosenblatt admitiu, porém disse que um dia pode ser útil para enviar um [aparato] ao espaço para capturar impressões para nós.

²⁵Mark I é um título comumente utilizado para a primeira versão de uma máquina.

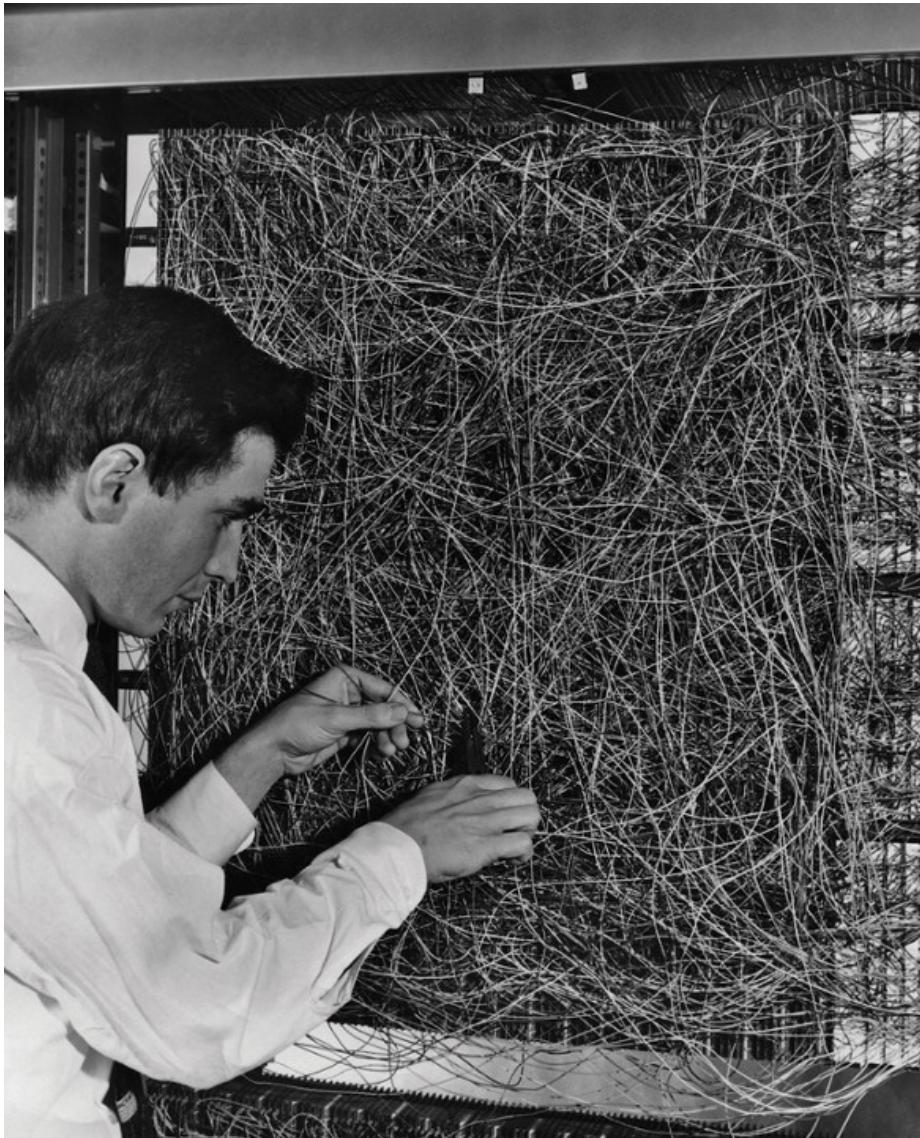


Figure 19: Frank Rosenblatt e Mark I.

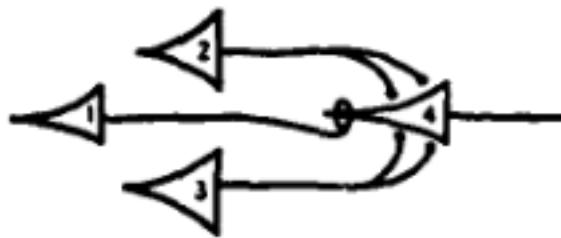


Figure 20: Diagrama de células lógicas em McCulloch & Pitts

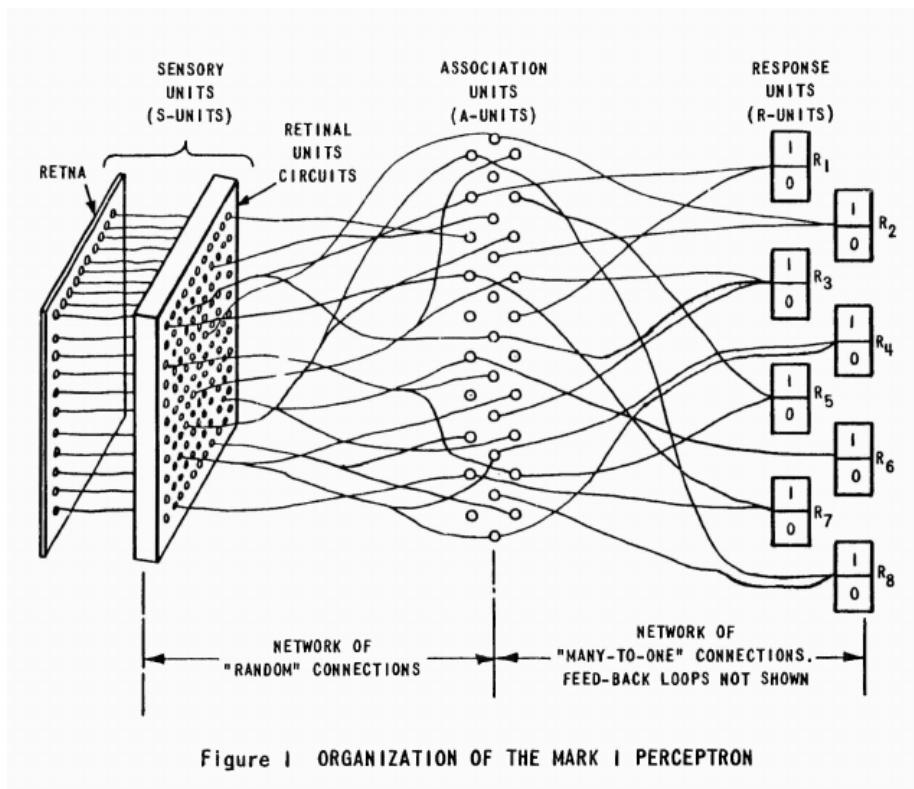


Figure 21: Organização do Mark I, retirado de seu manual de uso original

Rosenblatt protagonizava calorosos debates sobre inteligência artificial na comunidade científica junto a Marvin Minsky, um amigo da adolescência. Em 1969, Minsky e um matemático (Seymour Papert) publicaram um livro centrado no Perceptron (Perceptrons: An Introduction to Computational Geometry). Nele, provaram que o neurônio artificial era incapaz de resolver problemas não-lineares do tipo XOR. Para um problema eXclusive OR (OU eXclusivo) o neurônio deve disparar diante do estímulo A ou do estímulo B, porém não diante de ambos.

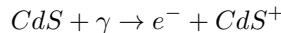
O impacto foi devastador sobre o otimismo vigente e se passou um período de 10 anos de baixíssima produção, conhecido como ‘idade das trevas’ do conexionismo. A retomada dos neurônios artificiais aconteceu somente na década de 80. Infelizmente, Rosenblatt morreu prematuramente em 1972 num acidente de barco, não presenciando o renascimento dos perceptrons.

Sabendo das origens do modelo, é curioso que a maioria dos cursos introduzam perceptrons do ponto de vista puramente matemático, apontando a semelhança com neurônios como mera curiosidade. Pelo contrário, a inspiração em neurônios biológicos e posterior sucesso nas tarefas designadas fala em favor de um fantástico caso de sucesso via engenharia reversa.

Criando neurônios

Mark I foi criado para reconhecimento visual, podendo ser considerado avô da visão computacional.

Possuía um campo de entrada fotossensível de 20x20 (400) células de Sulfeto de Cádmio, as unidades S. Ao reagir com a luz, CdS emite um elétron:



Caso a célula seja ativada, envia o sinal eletrônico a uma unidade intermediária A. A unidade intermediária, por sua vez, transmite um sinal eletrônico à saída. **A intensidade do sinal é regulada por sucessos prévios** de maneira a ajustar o sinal para a classificação correta. O aparato físico mimetiza o modelo matemático do **classificador**.

Um sinal luminoso excita cada campo de maneira diferente, ativando células de acordo com a quantidade de luz captada. Matematicamente, representamos cada neurônio sensível à luz como uma célula na matriz de entrada.

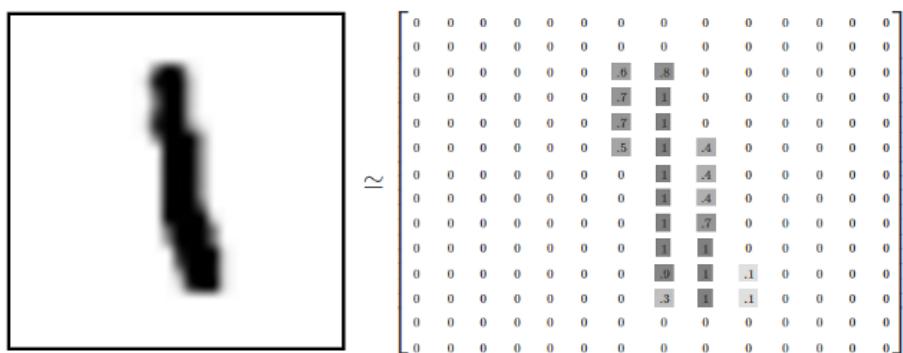


Figure 22: Exemplo de “1” em letra cursiva e sua representação numa matriz 2x2. <http://colah.github.io/posts/2014-10-Visualizing-MNIST/>

O dígito acima ('1') está numa imagem com 14×14 pixels (196 valores entre: 1, preto; e 0, branco). Esses pixels podem ser esticados e vistos como uma matriz X de dimensão $[196 \times 1]$ com valores entre 0 e 1 em cada elemento.

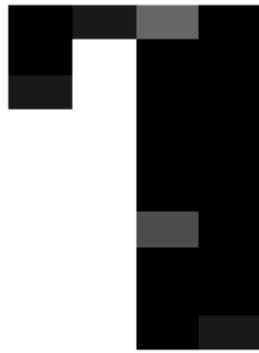
Vamos simular uma imagem semelhante:

```
>library(magrittr)
>set.seed(2600)
>my.image.data <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,0,1,.9,.6,1,0,0,0,0,0,0,0,0,
+                     0,0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,
+                     0,0,0,0,0.9,0,1,1,0,0,0,0,0,0,0,
+                     0,0,0,0,0,0.1,1,0,0,0,0,0,0,0,
```

```

0,0,0,0,0,0,1,1,0,0,0,0,0,0,
0,0,0,0,0,0,1,1,0,0,0,0,0,0,
0,0,0,0,0,0,.7,1,0,0,0,0,0,0,
0,0,0,0,0,0,1,1,0,0,0,0,0,0,
0,0,0,0,0,0,1,1,0,0,0,0,0,0,
0,0,0,0,0,0,1,.9,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0) %>%
matrix(.,14,14,byrow=T)
> image(t(my.image.data[14:1,]), axes = FALSE, col = grey(seq(1, 0, length = 256)))

```



Eis a nossa imagem [14x14]. O computador lê os valores entre 0 (branco) e 1 (branco), dispondo para nós o sinal visual correspondente numa paleta de cores. Aqui usamos 256 tons cinza.

Em regressão linear múltipla, calculamos um peso β_i para cada variável. O racional é parecido: ponderamos cada pixel por seus respectivos pesos w_i . Em analogia, cada imagem é uma observação de 196 variáveis.

Classificação

Na tarefa de regressão linear, o output deveria ser um número real $Y \sim \beta * X$ com $X, Y \in \mathbb{R}$, como o número médio de profissionais ou a expectativa de vida. Usaremos o perceptron para outra tarefa, a classificação, em que as possibilidades de saída são **categorias**. Isto é, o output é *discretizado*, geralmente num conjunto binário (e.g. $\{-1, 1\}$ ou $\{0, 1\}$) que sinaliza pertencimento à classe. Em nossa notação, o neurônio deve disparar (output $y = 1$) caso reconheça um objeto ou permanecer em repouso ($y = -1$) caso não seja.

Algebricamente, é uma multiplicação da matrizes entre imagem x_j , de dimensão $[196 \times 1]$ por uma matriz $W_{[196 \times 1]}$ que traz i pesos (**weights**) estimados para cada pixel para cada classe. Essa formulação é idêntica àquela feita em regressão linear. Para uma saída discreta, forçamos o resultado para +1 ou -1 com uma função de ativação (ϕ). O output linear $W^T X$ é transformado:

$$y = \phi(W^T X)$$

Assim, o produto $W^T X$ deve ter valor proporcional à probabilidade de ativação: se o input pertence à classe o resultado deve ser alto.

Usaremos a função *Heaviside step*:

$$\phi(x) = \begin{cases} +1 & \text{se } x \geq 0 \\ -1 & \text{se } x < 0 \end{cases}$$

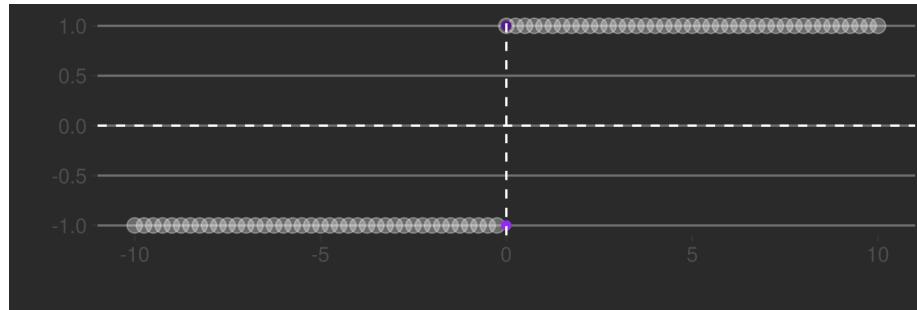


Figure 23: Heaviside step function

Em R:

```
# Heaviside
>phi_heavi <- function(x){ifelse(x >=0,1,-1)}
# Iniciando pesos com base em distribuição normal
>my_weights <- rnorm(196)
>w <- matrix(my_weights,196,1)
# Multiplicacao usando o operador %*%
>as.vector(my.image.data) %*% w
# Score
[,1]
[1,] -0.3794718
# Funcao de ativacao
>as.vector(my.image.data) %*% w %>% phi_heavi
[,1]
[1,] 1
```

Para o exemplo acima, nosso neurônio com pesos aleatórios foi ativado para o estímulo contendo o ‘7’. Inicialmente, estabelecemos pesos aleatórios a partir de uma distribuição normal (`my_weights <- rnorm(...)`). O processo de treinar o classificador é observar as respostas muitos exemplos de imagens x_i , alterando os valores de W para que os scores maiores sejam os das classes corretas. Assim, neurônio só dispara $y = 1$ quando diante do estímulo adequado.

O processo de treino é bastante simples:

Seja x_{i_j} o i -ésimo pixel da observação j . E w_0 o peso correspondente inicial, o peso atualizado, w' é:

$$w' = w_0 + \Delta w$$

Em que Δw indica o magnitude e o sentido da modificação no peso.

Aceitemos, por enquanto, a fórmula:

$$\Delta w_i = \eta(score_j - output_j)x_i$$

Em que x_{i_j} é o valor do i -ésimo pixel, w_i é o i -ésimo peso e η uma constante chamada *tasa de aprendizagem* (learning rate), que determina o tamanho dos incrementos feitos pelo algoritmo. Mostraremos a derivação dessa equação a seguir.

Auto MaRK I Usando as abstrações acima, codificamos nosso perceptron em R, o Auto MaRK I.

Argumentos: Exemplos (x , vetor de números reais) e estados esperados (y , disparar = 1 vs. não disparar = -1) devem ter mesmo tamanho.

Eta: Número especificando constante de aprendizagem.

Auto MaRK I inicializa um peso aleatório para cada entrada e, também numa ordem aleatória, percorre os exemplos atualizando os pesos.

```
>mark_i <- function(x, y, eta) {
  # inicializa pesos randomicos de distribuicao normal
  w <- rnorm(dim(x)[2]) # numero de pesos = numero de colunas em x
  ypreds <- rep(0,dim(x)[1]) # inicializa predicoes em 0
  # Processa as observacoes em x de forma aleatoria
  for (i in sample(1:length(y),replace=F)) {
    # predicao
    ypred <- sum(w * as.numeric(x[i, ])) %>% phi_heavi
    # update em w
    delta_w <- eta * (y[i] - ypred) * as.numeric(x[i, ])
    #nota: x[i,] sera multiplicado como matriz (dot product)
    w <- w + delta_w
    ypreds[i] <- ypred # salva predicao atual
  }
  print(paste("Weights: ",w))
  return(ypreds)
}
```

Vamos testá-lo para o problema proposto, separando flores *setosa* de *versicolor*. Preparação de dados:

```
>train_df <- iris[1:100, c(1, 2, 5)]
>train_df[, 4] <- -1
>train_df[train_df[, 3] == "setosa", 4] <- 1
>names(train_df) <- c("s.len", "s.wid", "species","target")
>head(train_df)
  s.len s.wid species target
1  5.1   3.5 setosa     1
2  4.9   3.0 setosa     1
3  4.7   3.2 setosa     1
4  4.6   3.1 setosa     1
5  5.0   3.6 setosa     1
6  5.4   3.9 setosa     1
> train_df[60:65,]
  s.len s.wid   species target
60   5.2   2.7 versicolor    -1
61   5.0   2.0 versicolor    -1
62   5.9   3.0 versicolor    -1
```

```
63  6.0  2.2 versicolor -1  
64  6.1  2.9 versicolor -1  
65  5.6  2.9 versicolor -1  
>x_features <- train_df[, c(1, 2)]  
>y_target <- train_df[, 4]
```

E então, podemos ativá-lo:

Usando $\eta = 0.002$, obtivemos 72% de acurácia (classificações corretas, diagonal na matriz de confusão). Podemos modificar a taxa de aprendizagem. Com $\eta = 0.05$, ficamos com 51%. Com $\eta = 0.1$, temos 60%. Uma acurácia considerável em relação ao esperado com adivinhação. Contudo, estas soluções não são estáveis e passagens repetidas geram previsões muito diferentes.

```
> y_preds <- mark_i(x_features, y_target, 0.05)
[1] "Weights: -1.26323926081935" "Weights: 1.85983709987067"
> table(y_preds,train_df$target)
y_preds -1  1
-1 35 16
 1 15 34

> y_preds <- mark_i(x_features, y_target, 0.1)
[1] "Weights: -1.83248546552824" "Weights: 3.19075461158561"
> table(y_preds,train_df$target)
y_preds -1  1
-1 31 21
 1 19 29

> y_preds <- mark_i(x_features, y_target, 0.01)
[1] "Weights: -0.250410476080629"
[2] "Weights: 0.447470183281492"
> table(y_preds,train_df$target)
y_preds -1  1
-1 25 27
```

O que há de “errado” com nosso estimador?

Durante a exposição, a seguinte regra nos ajudou, mas não foi explicada.

$$\Delta w_i = \eta(score_j - output_j)x_i$$

Antes, verificamos (Cap. 2) uma solução fechada para o problema de regressão, em que a melhor estimativa para a inclinação da reta, β , poderia ser calculada diretamente.

O perceptron atualiza seus pesos de maneira recursiva, aprendendo um pouco (Δw_i) com cada exemplo. Um novo estímulo determina quanto (magnitude em Δw) e em que direção (+ ou -) um peso deve mudar para diminuir erros.



Gradient Descent para o Perceptron

Ao otimizar estimativas, nos concentramos em encontrar máximos ou mínimos para espaços definidos. Em geral, estes são superfícies descrevendo o tamanho dos erros em função dos pesos adotados pelo modelo. O nosso objetivo é encontrar o local mais *baixo*. Para superfícies muito irregulares, aceitamos um ponto suficientemente *baixo*.

Em regressão linear, o espaço é conhecido, é possível ir ao ponto mais baixo diretamente. Para outros modelos, isso não é tão simples.

Δw_i pode ser obtido usando o conceito de *Gradient Descent*.

O processo é como descer uma ladeira *de olhos vendados*. Só podemos saber a inclinação local (diferença entre pé esquerdo e pé direito). Podemos descer dando passos sempre na direção do pé mais baixo.

O que precisamos então é da inclinação da superfície relacionada aos erros em função dos pesos.

Levando em conta cada j -ésima observação, primeiro definimos uma função de perda L expressando a soma dos erros nos n exemplos.

$$L = \sum_j^n E(score_j, output_j)$$

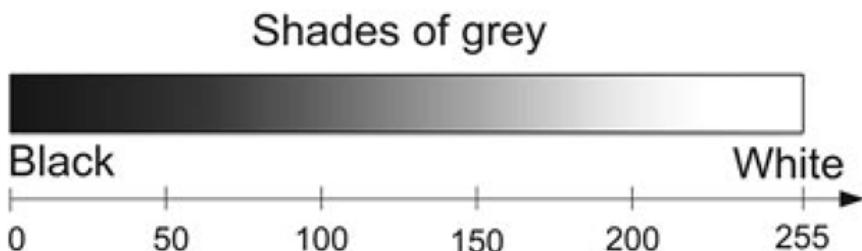
Usaremos para nossa função de erro a distância euclidiana entre score desejado e output. O score desejado é a resposta ótima e o output é um produto entre pesos e entrada:

$$E = d_{eucl.}(score_j, output_j) = (score_j - output_j)^2$$

Essa função descreve a superfície em função dos erros usando uma relação quadrática: errar para cima tem o mesmo peso que errar para baixo e erros extremos são magnificados (x^2) polinomialmente.

O processo envolve implementar uma função de erro entre resultados da rede e um espaço virtual de scores ótimos. O sucesso do treinamento depende de uma correspondência entre a função de distância escolhida e a distância real no espaço em que os dados foram gerados. Não sabemos se isso reflete a realidade. No exemplo, cada pixel reflete um sinal de 0 a 255.

A figura abaixo mostra a correspondência entre valores da medida e escala visual.

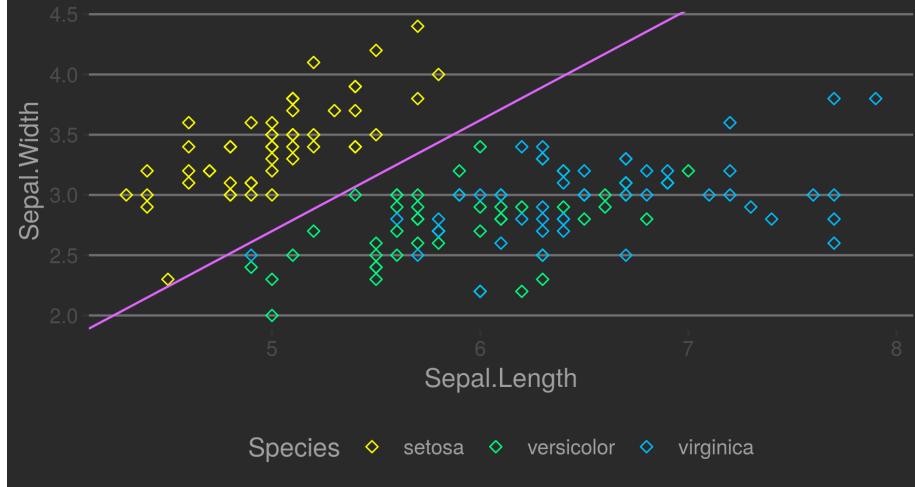


A intuição para sensibilidade à luz pode ser percebida num intervalo contínuo entre incidência total de luz (valores extremos de branco, medida: 255) e ausência total (valores extremos de preto, medida: 0). Supondo que podemos atribuir um rótulo a cada tom de cinza e que esse conjunto é ordenável pela *clareza*, dizemos que há isomorfismo de ordem entre os conjuntos. Isso implica que a distância eulidiana deve funcionar razoavelmente em nossas medidas como em números reais \mathbb{R} .

Resta saber se a projeção das observações é linearmente separável. É intuitivo para seres humanos saber quais problemas serão separáveis: basta imaginar a tarefa de diferenciar tipos de imagens com uma régua numa tela em preto e branco.

Se os dados são linearmente separáveis, o algoritmo converge com um número suficiente de exemplos. Usando o *iris*, funcionaria para separar flores *setosa* de outra classe, mas não teríamos bons resultados separando *virginica* de *versicolor*.

```
>ggplot(iris,aes(x=Sepal.Length,y=Sepal.Width,color=Species))+  
  geom_point(shape=5)+ geom_abline(slope = 0.92,intercept = -1.9,color="mediumorchid1") +  
  scale_colour_manual(values = c("yellow", "springgreen", "deepskyblue")) +  
  theme_hc(style = "darkunica")
```



Para descobrir o valor mínimo de L , vamos encontrar polos através de derivadas parciais. Ou, seu equivalente para funções de múltiplas variáveis (espaços multidimensionais), o gradiente (∇).

Para cada observação x_j , a derivada parcial da função de perda em relação a um peso w_i expressa a taxa de variação no erro global em função daquele peso.

$$\frac{d}{dw_i} L(w_i) = \frac{d}{dw_i} \frac{1}{n} \sum_j nE(score_j, output_j)$$

Sabemos então se devemos ajustar o peso para cima ou para baixo, assim com a magnitude do passo. Algebricamente, modificaremos w seguindo o inverso do gradiente. A taxa de aprendizagem é um hiperparâmetro que regula artificialmente o tamanho desse passo:

$$\begin{aligned}\Delta w_i &= -\eta \frac{dL}{dw_i} \\ &= -\eta \frac{d}{dw_i} \frac{1}{n} \sum_j^n E(score_j, output_j)\end{aligned}$$

Lembrando que o erro é dado pela distância euclidiana:

$$= -\eta \frac{d}{dw_i} \frac{1}{n} \sum_j^n (score_j - output_j)^2$$

Fazemos $f(x) = (score_j - output_j)$ e $g(x) = x^2$, de maneira que

$$L = \frac{1}{n} \sum_j^n E(score_j, output_j) = (g \circ f)$$

$$= \frac{1}{n} \sum_j^n (score_j - output_j)^2$$

Podemos resolver $\frac{d}{dw_i} L$ aplicando a regra de cadeia

$$(g \circ f)' = (g' \circ f)f'$$

e a ‘regra do tombo’ para derivadas de polinômios ($\frac{d}{dx}(x^n) = nx^{n-1}$).

Então,

$$f' = \frac{d}{dw_i} (score_j - output_j)$$

O output é dado pelo produto escalar entre pesos w_j e entradas x_j :

$$f' = \frac{d}{dw_i} (score_j - w_j \cdot x_j)$$

O score desejado não depende dos pesos, portanto a primeira derivativa é 0.

$$\begin{aligned} f' &= 0 - \frac{d}{dw_i} w_j \cdot x_j \\ &= -\frac{d}{dw_i} \sum_{i,j}^n w_{i,j} * x_{i,j} \\ &= -\frac{d}{dw_i} (w_0 * x_0 + \dots + w_i * x_i + w_n * x_n) \end{aligned}$$

Os termos não dependentes de w_i também são zerados e ficamos com o primeiro termo da soma:

$$f' = -\frac{d}{dw_i} w_i x_i$$

A função a ser derivada agora descreve uma relação linear (polinômio de grau 1) em w_i e temos:

$$f' = (-x_{i,j})$$

Sabendo f' , buscamos o outro termo em $(g \circ f)'$:

$$(g \circ f) = (score_j - output_j)^{2-1}$$

$$(g' \circ f) = 2(score_j - output_j)^{2-1}$$

$$= 2(score_j - output_j)$$

Por fim, a derivada parcial da função de perda para o i-ésimo peso w_i é:

$$\frac{dL}{dw_i} = \sum_j^n \frac{d}{dw_i} (score_j - output_j)^2$$

$$= \sum_{i,j}^n 2(score_j - w_j \cdot x_j)(-x_{i,j})$$

Para simplificar a expressão e estabelecer o tamanho dos incrementos sobre os pesos, escalamos por uma constante, dada por $-\frac{1}{2}\eta_0$:

$$-\frac{1}{2} * \eta_0 \frac{dL}{dw_i} = -\frac{1}{2} \eta_0 * 2(score_j - output_j)(-x_j)$$

$$\Delta w_i = \eta_0 \sum_j^n (score_j - w \cdot x)(x_j)$$

E η_0 é um [hiper]parâmetro que simplifica a equação e define o tamanho dos incrementos usados.

Como implementamos antes no Auto MaRK I.

```
(...)
ypred <- sum(w * as.numeric(x[i, ])) %>% phi_heavi
delta_w <- eta * (y[i] - ypred) * as.numeric(x[i, ]) #<-----
w <- w + delta_w
(...)
```

Chamamos η de hiperparâmetro. A escolha de valores para hiperparâmetros é um dos desafios em aprendizagem estatística. Repetindo a aprendizagem com *iris*, vamos testar:

```

> y_preds <- mark_i(x_features, y_target, 0.01)
[1] "Weights: -0.0153861618736636" "Weights: 0.0812191914731158"
> table(y_preds,train_df$target)
y_preds -1 1
-1 25 27
 1 25 23
> y_preds <- mark_i(x_features, y_target, 0.01)
[1] "Weights: -0.685141728446126" "Weights: 1.03174770234754"
> table(y_preds,train_df$target)
y_preds -1 1
-1 47 10
 1 3 40
> y_preds <- mark_i(x_features, y_target, 0.01)
[1] "Weights: -0.193515893657872" "Weights: 0.180589056542887"
> table(y_preds,train_df$target)
y_preds -1 1
-1 19 37
 1 31 13
> y_preds <- mark_i(x_features, y_target, 0.01)
[1] "Weights: -0.0672147799277951" "Weights: 0.115145797950982"
> table(y_preds,train_df$target)
y_preds -1 1
-1 45 12
 1 5 38

```

Usando $\eta = 0.01$, temos 48%. Entretanto, rodar repetidas vezes retorna classificações muito boas (Acc. > 0.8) ou muito ruins. O que se passa?

Em geral, passos grandes impossibilitam ajustes finos e podem não convergir, assim como é impossível para um animal grande explorar um vale estreito.

Taxas pequenas levam mais tempo (n de observações) para atingir um mínimo. Se o espaço for irregular, também existem mais chances de se atingir um mínimo secundário ao invés do fundo do espaço. Assim como um animal pequeno percorre o caminho mais lentamente e pode ter a ilusão de que atingiu mínimos rapidamente.

Uma maneira trivial é testar muitos valores possíveis e observar o desempenho, entretanto isso não é exequível para grandes volumes de dados e/ou muitos parâmetros. Existem diversos processos heurísticos e algoritmos para encontrar valores ótimos. Podemos também ajustar parâmetros ao longo do processo de aprendizagem ou testar pontos diferentes de partida.

Uma forma popular para otimizar o treinamento é partitionar o dataset em pedaços e apresentar os particionamentos (epochs) repetidas vezes ao classificador ou acumular os erros de epochs ao invés de exemplos individuais. Assim, calculamos erros agregados e evitamos mínimos locais. Para evitar muitas alterações e andar em círculos, avançamos por mais tempo em apenas uma direção

antes de recalcular a rota. Epochs podem ser recombinados e/ou reapresentados, aumentando artificialmente o n para calcular gradientes.

Deep learning



Intuições

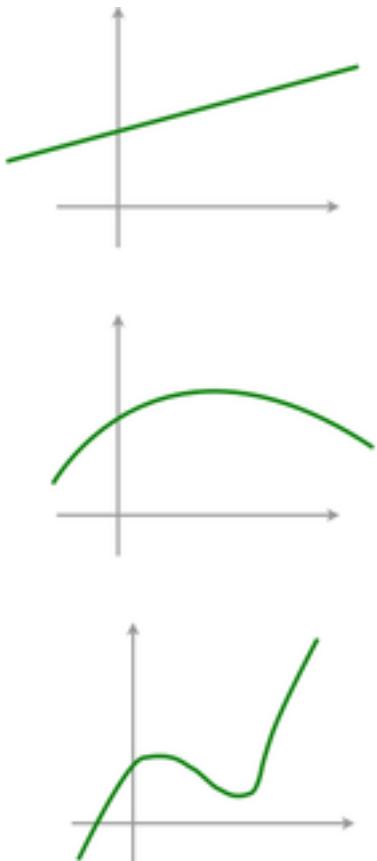
Com o aprendizado através de exemplos, otimizamos nosso classificador (mudando pesos W) para minimizar a perda gradualmente. Uma das condições para o *perceptron* acima funcionar foi a separabilidade linear das classes no espaço examinado.

Alguns problemas são mais difíceis, sendo separados por curvas. Outros são ainda mais difíceis, exigindo muitas transformações e funções específicas. Uma alternativa é usar polinômios de ordem maior. Ao invés de $Y \sim \beta_0 + \beta_1 X$, podemos introduzir termos com expoentes maiores em X :

$$Y \sim \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \dots$$

A inclusão flexibiliza a função, que pode se adequar melhor aos dados.

Na regressão linear, ajustamos o ângulo e a altura de uma barra fixa para reduzir a distância até os pontos. Com termos quadráticos, é possível dobrar essa barra em relação ao centro, mas as pontas devem ir numa mesma direção. Com termos cúbicos, isso é flexibilizado.



A introdução de termos polinomiais de ordem maior torna consideravelmente mais difícil a otimização das estimativas.

Um neurônio *linearmente sensível* a input e dotado de uma barreira (*threshold*) para disparos é capaz de resolver problemas de classificação mais simples. Para problemas mais difíceis, ao invés de implementar células de processamento radicalmente diferentes e/ou mais complexas, a natureza usa um artifício engenhoso. Neurônios comuns são encadeados: cálculos simples e comunicação local das unidades possibilita a aprendizagem.

Os dados são apresentados aos perceptrons na linha de frente. O output das primeiras células é usado como input para neurônios da próxima camada. Assim, conseguimos implementar transformações adequadas (rotações, torções, escalonamentos, dobras) em sequência, de maneira que abstrações complexas possam ser capturadas.

Going Deep

As versões reais da maioria dos conceitos criados por seres humanos não são idênticas umas às outras. Em outras palavras, não existe um conjunto rígido de regras para classificarmos a maior parte das entidades ao nosso redor. Muitas entidades são diferentes, porém similares o suficiente para pertencer a uma mesma categoria.



Todos são naturalmente reconhecidos como felinos, mas apresentam variações de tamanho, cor e proporção em todo o corpo. Esse é um problema interessante e antigo, mais conhecido na ideia de entes platônicos, os quais capturam a essência de um conceito.

Alguns filósofos contemporâneos tomam as abstrações humanas como instâncias de um conceito mais genérico: mapas biológicos contidos em redes neurais. Uma brilhante exposição é feita por Paul Churchland em *Plato's Camera*.

Esses mapas estão associados de forma hierarquizada. Numerosos padrões em níveis inferiores e menos deles em camadas superiores.

No caso da visão, neurônios superficiais captam pontos luminosos. O padrão de ativação sensorial captado na retina é enviado ao córtex visual primário é o primeiro mapa, que é torcido e filtrado caminho cima.

Em níveis superiores, sinais individuais de cones sensíveis faixas de energia compõem a paleta de cores que percebemos.

Neurônios intermediários possuem configurações que identificam características simples: olhos e subcomponentes da face. Por fim, temos camadas mais profundas, ligadas a abstrações complexas e funções superiores (e.g. linguagem).

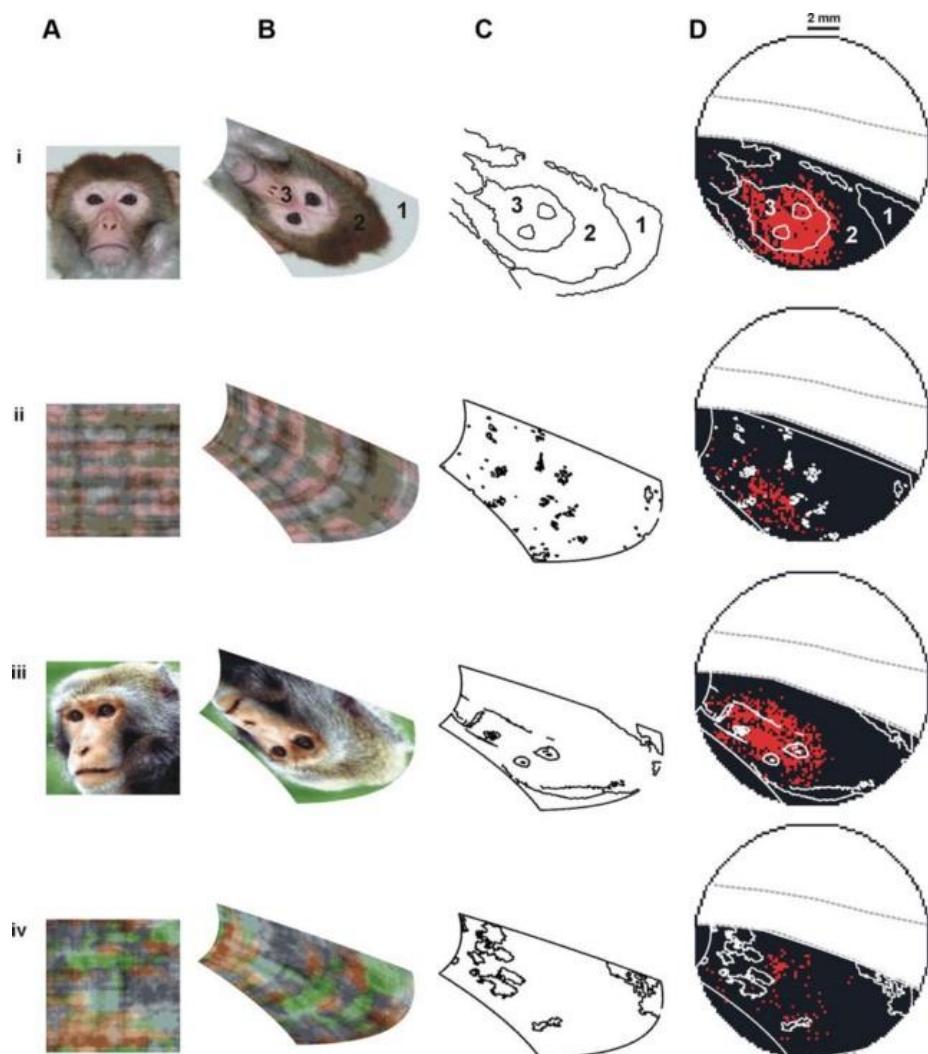


Figure 24: Resposta a estímulos visuais em V1 de *Macaca fascicularis* <http://www.jneurosci.org/content/32/40/13971>

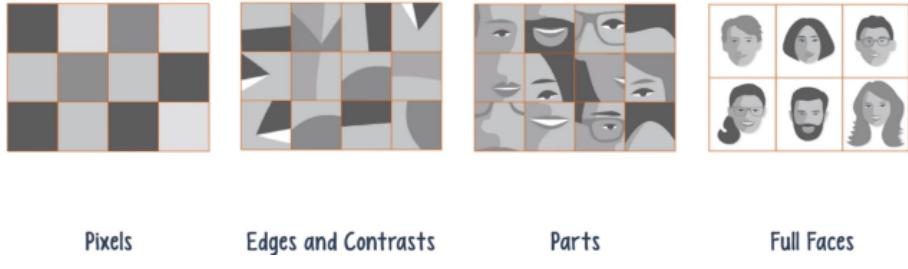
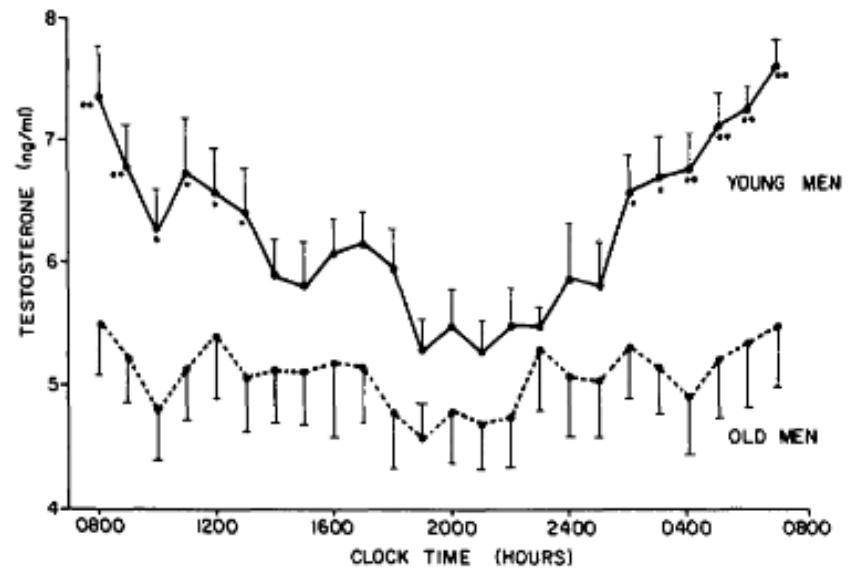
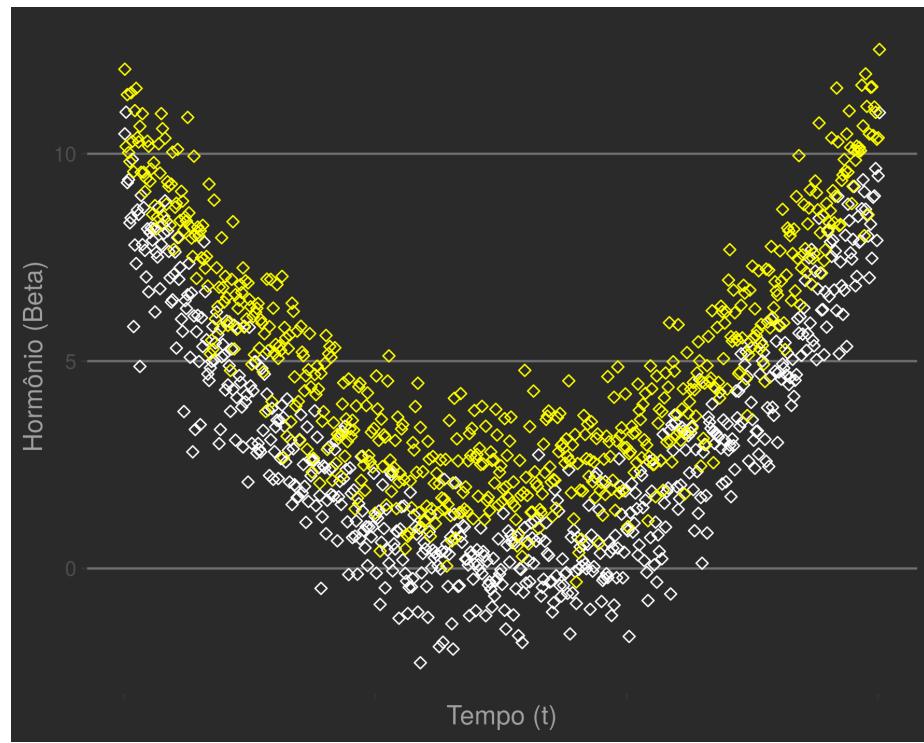


Figure 25: Retirado de: <https://www.youtube.com/watch?v=SeyIg6ArS4Y>

Deduzindo superfícies

Um classificador deve capturar essa estrutura abstrata a partir de modelos matemáticos tratáveis. Para examinarmos esse aspecto, usemos um exemplo. O gráfico abaixo representa milhares de amostras com: (1) a curva diária natural de testosterona (branco) e a curvas para medidas sob uso de esteroides anabolizantes (amarelo).

```
>normal <- (purrr::map(seq(-3,3,0.01), .f =function(x) x^2) %>%
+  as.numeric)+ rnorm(601)
>over <- (purrr::map(seq(-3,3,0.01), .f =function(x) x^2+2) %>%
+  as.numeric)+ rnorm(601)
>horm_df <- data.frame(norm = normal, ov = over,time=1:601)
>ggplot(data=horm_df,aes(y=norm,x=time))+ 
+   geom_point(color="white",shape=5)+ 
+   geom_point(data=horm_df,aes(y=over,x=time),color="yellow",shape=5)+ 
+   ylab("Hormônio (Beta)")+xlab("Tempo (t)")+ 
+   scale_x_continuous(labels=NULL)+ 
+   theme_hc(style="darkunica")
```



Como hipotéticos membros de uma comitê atlético, nosso objetivo aqui é, dada uma amostra, saber se o atleta está sob efeito de esteroides.
Quando experimentamos, normalmente haverá ruídos (erros) na medida e rece-

beremos medições imprecisas da curva. Variações na dieta daquele dia, micções, sudorese, stress e outros fatores. Sabemos que a testosterona flutua diariamente seguindo uma curva.

Para cada medida, temos o tempo (t , eixo horizontal) e o nível hormonal (β , eixo vertical).

Um modelo bastante popular para classificações é o de regressão logística. Nele, estimamos probabilidade para um evento com base nas probabilidades de uma função sigmoide. Temos uma probabilidade (valor entre 0 e 1) definida por:

$$P(h, \beta) = \frac{1}{1 + e^{-(i+t*h+\beta*y+\epsilon)}}$$

ϵ representa o erro e i é uma constante.

A equação parece estranha, mas aparece quando buscamos calcular probabilidades a partir de uma combinação linear dos nossos parâmetros:

$$P(x) \sim i + t * x + \beta_i * y + \epsilon$$

Isso permite que o processo de estimação seja quase idêntico ao da regressão linear, que é facilmente tratável.

Em uma linha de R:

```
>class_df <- class_df <- data.frame(measures=c(horm_df$norm,horm_df$ov),
  time=c(horm_df$time,horm_df$time),
  target=c(rep(0,601),rep(1,601)))
>logist.fit <- glm(target ~ measures + time, family=binomial,data=class_df)
```

Outra consequência é de que uma relação linear torna magnitude e o sentido dessas relações interpretáveis. Por exemplo, um parâmetro positivo (e.g. $\beta = 0.241$) indica que aumentos em X aumentam a probabilidade de ativação e parâmetros negativos (e.g. $\beta = -0.9517$) têm efeito contrário.

Muitas avaliações de risco em saúde ou avaliação de crédito em finanças estimam probabilidades com base nos parâmetros de uma regressão logística.

Usamos um limite de decisão (*decision boundary*) dependente de relações lineares. Tecnicamente, um hiperplano. Um hiperplano divide o espaço em duas partes. É a generalização de plano (curvatura zero) para quaisquer dimensões. O hiperplano é um espaço de $n - 1$ em um espaço n dimensões. A reta é um hiperplano em duas dimensões (nossa caso), o plano tradicional é um hiperplano em 3 dimensões. Para dimensões mais altas, a visualização é menos simples.

Para nosso exemplo não-linear, seria difícil capturar as diferenças entre atletas dopados usando apenas esta equação.

Acima, um neurônio sigmoide, que equivale à regressão logística. É como o plano anterior, mas visto de cima, dividimos ele em duas regiões para classificação. Por

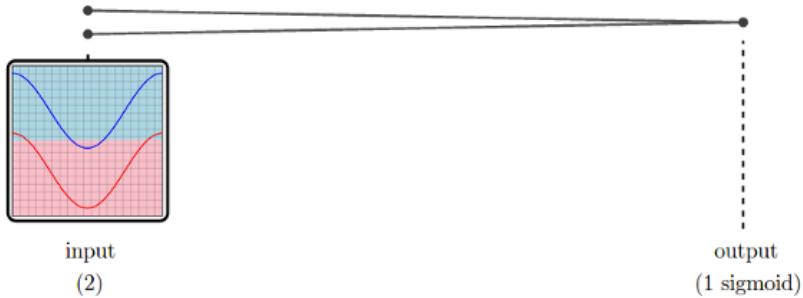


Figure 26: <http://colah.github.io/posts/2015-01-Visualizing-Representations/>

que? O classificador linear otimiza suas respostas levando em conta apenas o valor absoluto da medida hormonal. Isto é, valores acima de um limite serão considerados doping, não considerando horário.

O coeficiente para o tempo estimado foi tende a ser próximo de 0. Ao tentar dividir os grupos com uma régua, o melhor é tentar uma reta paralela ao eixo x . Podemos verificar isso diretamente através dos parâmetros estimados em nosso modelo de regressão.

Mudar isso tornaria a reta divisória inclinada em relação ao eixo x , piorando a classificação para valores baixos ou altos.

```
> summary(logist.fit)
Call:
glm(formula = target ~ measures + time, family = binomial, data = class_df)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.93641 -1.02791 -0.07236  1.12396  1.63490 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -9.439e-01  1.504e-01 -6.276 3.48e-10 ***
measures    2.411e-01  2.186e-02 11.027 < 2e-16 ***
time        -2.597e-05  3.621e-04 -0.072    0.943  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1666.3  on 1201  degrees of freedom
Residual deviance: 1526.0  on 1199  degrees of freedom
AIC: 1532
```

```

Number of Fisher Scoring iterations: 4
> prob <- predict(logist.fit,type=c("response"))
> class_df$prob <- prob
> curve <- roc(target ~ prob, data = class_df)
> curve
Call:
roc.formula(formula = target ~ prob, data = class_df)

Data: prob in 601 controls (target 0) < 601 cases (target 1).
Area under the curve: 0.6964

```

Quem poderá nos ajudar?

Voltamos às redes neurais para resolver o problema. Quando processamos o sinal em etapas, cada camada modifica os dados para as camadas posteriores, transformando e filtrando/dando forma.

As camadas intermediárias permitem a transformação gradual do sinal, e o sistema acerta usando apenas dois classificadores simples (sigmoids). No exemplo acima, temos uma camada de 2 neurônios entre o input e o output.

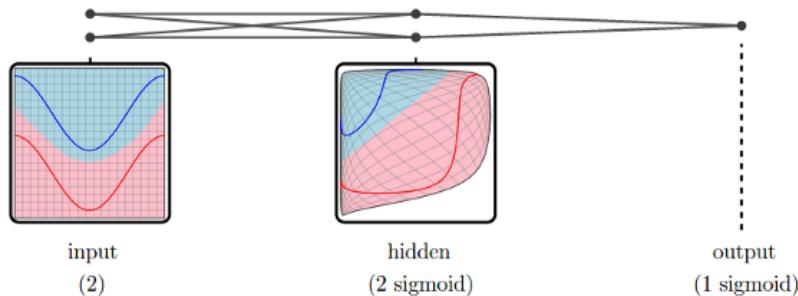


Figure 27: Visualização do processamento de sinal, tornando-o linearmente separável. Fonte: <http://colah.github.io/posts/2015-01-Visualizing-Representations/>

Agora, a primeira camada (hidden) modifica a entrada com duas unidades sigmoids e a segunda camada pode classificar corretamente usando apenas uma reta, algo que era impossível antes.

Em tese, esse modelo pode capturar melhor as características que geraram os dados (flutuação hormonal ao longo do dia).

Neurônios

Notem que o diagrama acima lembra uma rede neural. Esse tipo de classificador foi inspirado na organização microscópica de neurônios reais e acredita-se que seu funcionamento seja de alguma forma análogo. A arquitetura de redes convolucionais (convolutional neural networks), estado da arte em reconhecimento de imagens, foi inspirada no córtex visual de mamíferos²⁶. Outros modelos bio inspirados (Spiking neural networks, LTSMs...) apresentam desempenhos inéditos para tarefas complexas e pouco estruturadas, como reconhecimento de voz e tradução de textos. A teoria mais aceita é de que o maquinário neural dos animais foi desenhado por processos evolutivos, como a seleção natural. Assim, apresenta coloridas formas de complexidade a depender da tarefa desempenhada.

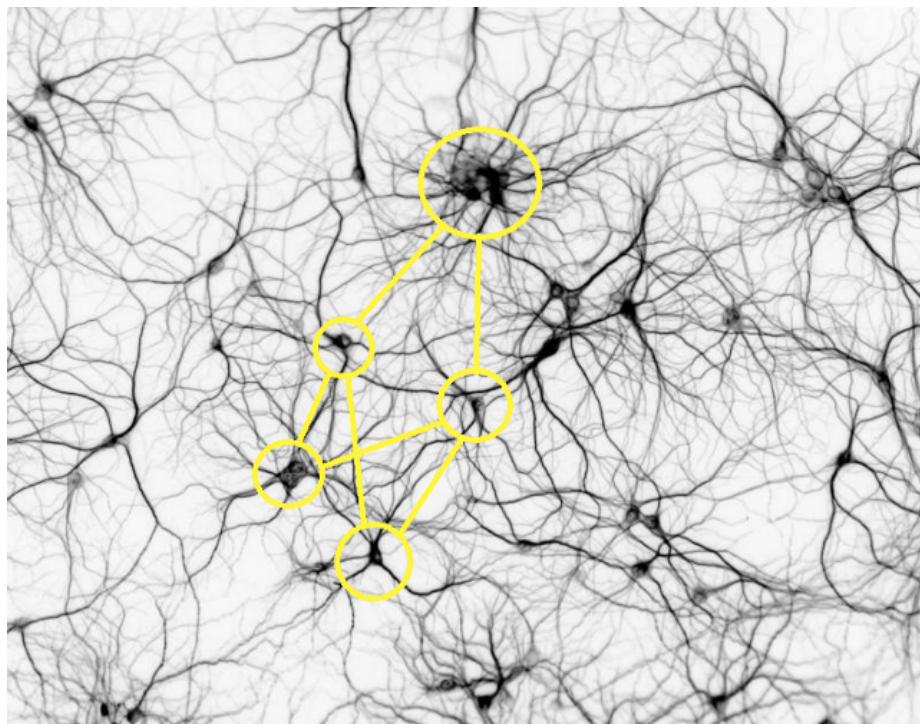


Figure 28: Modificado de <http://www.rzagabe.com/2014/11/03/an-introduction-to-artificial-neural-networks.html>

Como podemos ver, as redes biológicas são complexas, com até dezenas de bilhões de unidades de processamento paralelas conectadas. Zona destacada possui grafo isomorfo ao descrito no texto.

Nos modelos profundos (deep) de reconhecimento de rosto, neurônios de camadas

²⁶(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557912/>)

superficiais capturam bordas, ângulos e vértices, camadas intermediárias detectam presença de olhos, boca, nariz. Por fim, camadas ao final da arquitetura decidem se é um rosto ou não e a quem ele pertence.

Eficiência e aplicações Podemos demonstrar formalmente que uma rede neural com apenas uma camada interna é capaz de aproximar qualquer função. A prova não é lá essas coisas, já que, no fundo o que fazemos é criar uma tabela de consulta (lookup table) para os valores de entrada e saída usando os neurônios. Na prática, é difícil obter boas performances. Tão difícil que as redes neurais também passaram décadas esquecidas. Se você rodar o modelo abaixo, baseado no nosso exemplo, verá que a acurácia é próxima da regressão logística. É necessário algum conhecimento e tempo para afinar os detalhes.

Normalmente, depende da qualidade e da quantidade dos dados. O boom veio com a descoberta de topologias de rede especificamente boas para certas tarefas (e.g. LSTM para linguagem natural, *Conv Nets* para visão computacional). Em outras palavras, modelar uma rede neural para problemas inéditos pode ser algo desafiador.

O código a seguir mostra como implementar uma rede com topologia similar usando a lib **caret**. Conseguimos acurácia de 81% usando 5 neurônios.

```
# Neural Net para o exemplo
>library(caret)
> class_df$time_sc <- scale(class_df$time)
> nn_horm <- caret::train(x = class_df[,c(1,5)], y=factor(class_df$target),method="mlp")
Multi-Layer Perceptron

1202 samples
  2 predictors
  2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1202, 1202, 1202, 1202, 1202, 1202, ...
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.6488305  0.2948640
  3     0.8181583  0.6355261
  5     0.8198874  0.6393824

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was size = 5.
```

As redes neurais passaram algum tempo esquecidas, até que algumas reviravoltas

²⁷ permitiram o treinamento eficaz dessas redes. Algoritmos para melhorar o treinamento, assim como arquiteturas econômicas ou especialmente boas em determinadas tarefas. Além disso, o uso de processadores gráficos (GPU), desenhados para as operações de álgebra linear que discutimos (com matrizes) permitiu treinar em um volume maior de dados.

Backpropagation

Backpropagation é um processo chave em para permitir o treinamento de classificadores em deep learning. É o conceito de propagar gradientes da função de perda ao longo da rede de maneira a atualizar cada nodo. Historicamente, surgiu nos estudos sobre teoria do controle.

Como vimos, podemos encarar a rede neural como uma sequência de funções plugadas. Algebraicamente, se o primeiro nodo é $q(x, y)$, o neurônio f que recebe sua saída como input tem valor $f(q(x, y))$ ou $f \circ q$.

Exemplo

Neurônio de input: $q(x, y) = 3x + 2y$

Segundo neurônio: $f(z) = z^2$

Output final: $f(q(x, y)) = q^2 = (3x + 2y)^2$

À primeira vista funções complexas vão possuir gradientes difíceis de calcular. Além disso, temos que calcular valores para cada neurônio em camadas diferentes. *Backpropagation* usa a *regra de cadeia* para calcular as derivadas por camada. Encadeando sequências de funções elementares com derivada conhecida, podemos atingir mapeamentos complexos e ainda assim calcular o gradiente sem muito esforço.

Podemos obter o gradiente da função de perda no nodo de hierarquia mais alta (f), com respeito a uma das variáveis de entrada (x) na hierarquia mais baixa. A operação é computacionalmente barata, bastando multiplicar as derivadas parciais dos erros em cada parte.

$$\frac{df}{dx} = \frac{df}{dq} \frac{dq}{dx}$$

É possível calcular de forma recursiva, portanto local e paralela, ao longo das camadas. Fazendo o mesmo acima para df/dy , teremos os valores de df/dx e df/dy que é precisamente nosso gradiente em f .

```
# Valor duplo (x,y) para inputs
>x <- 1
>y <- 3
q <- 3*x + 2*y # primeira camada
f <- q^2 # segunda camada
```

²⁷(<http://people.idsia.ch/~juergen/who-invented-backpropagation.html>)

```

# Backprop - Mudanças em hierarquia superior
# dadas por entradas de camadas inferiores
dfdq <- 2*q # derivada de x^2 ; variação de f em função de q
dqdx <- 3 # Derivada de 3x ; variação de q em função de x
dqdy <- 2 # Derivada de 2x ; variação de q em função de y
# Obter gradiente de f(x,y) multiplicando as parciais
dfdx = dfdq*dqdx
dfdy = dfdq*dqdy
grad = c(dfdx,dfdy)
> grad
[1] 24 16

```

Usando essa lógica, calculamos os gradientes para a função de erro e treinamos o modelo.

Podemos então implementar nossa rede neural, Mark II.

Mark II Nossa rede terá um perceptron de entrada com dimensão igual à do input. Entretanto, acrescentamos um peso a mais, que corresponderá a um intercepto.

Note que

$$y = w_0 + w_1x_1 + w_2x_2$$

é o mesmo que

$$y = 1 * w_0 + w_1x_1 + w_2x_2$$

Assim, adicionamos um peso w_0 e também forçamos uma dimensão a mais no input, que sempre terá valor 1. Chamamos esse artifício de adicionar um intercepto (*bias*) de *bias trick*. Ajuda a estabelecer um valor basal para o output, facilitando a convergência.

```

library(magrittr)
library(ggplot2)
set.seed(2600)

mark_ii <- function(x, y, eta, reps=1) {

  # inicializa pesos randomicos de distribuicao normal
  w1 <- rnorm(n = (dim(x)[2]+1)) %>% as.matrix # numero de pesos = numero de colunas em x +

```

Em seguida, neurônios da camada intermediária, dois, cada um com dois pesos.

```

w21 <- rnorm(2) %>% as.matrix
w22 <- rnorm(2) %>% as.matrix

```

Zeramos as previsões e iniciamos os loops de treinamento. Para a rede neural, precisamos de muitos exemplos de exposição, então embutimos em Mark II

um parâmetro (`reps`) responsável por repetir o processo de treinamento com o dataset.

A rigor o melhor seria particionar o dataset em fragmentos menores para cada epoch, mas vamos manter as coisas simples.

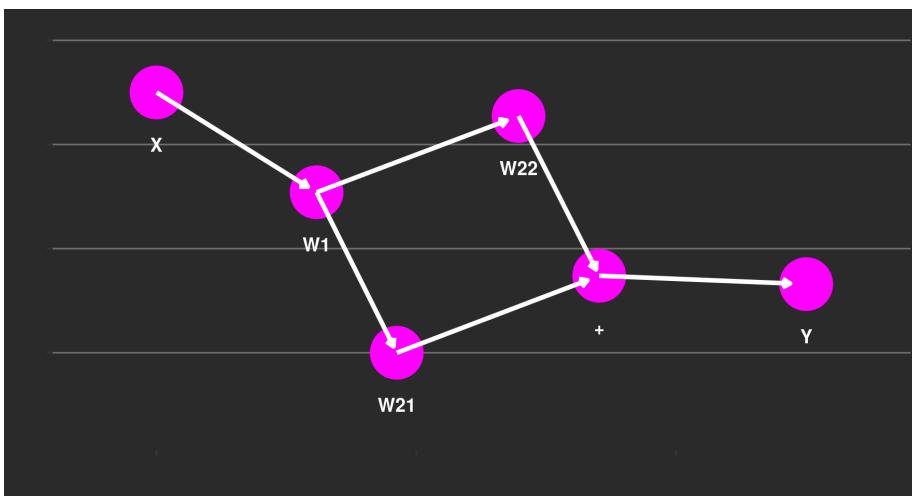
```
ypreds <- rep(0,dim(x)[1]) # inicializa predicoes em 0
yerrors <- rep(0,dim(x)[1]) # inicializa predicoes em 0
for (j in 1:reps){
  print(paste("This is training epoch:",j))
  print(paste("Current weights:",w1,w21,w21))
```

Predições: a primeira camada soma o produto de seus pesos pela entrada e pela unidade (*bias trick*). Os neurônios da segunda camada somam o produto de seus pesos pelo output. O output final é a soma dos outputs na camada intermediária.

```
# Processa as observacoes em x de forma aleatoria
for (i in sample(1:length(y),replace=F)) {
  # predicao
  ypred1 <- sum(w1 %*% c(as.numeric(x[i, ]),1))

  ypred21 <- sum(w21 %*% as.numeric(ypred1))
  ypred22 <- sum(w22 %*% as.numeric(ypred1))

  out <- sum(ypred21,ypred22)
```



Agora, as regras de atualização dos pesos seguindo derivações com regra de cadeia. Para os neurônios intermediários, temos: $\frac{d}{dw_{21}}$ e $\frac{d}{dw_{21}}$ de $(target - (pred22 + pred21))^2$.

$$\frac{d}{dw_{21}}(target - (pred22 + pred21))^2$$

Aplicando a regra de cadeia e sabendo que a predição do segundo neurônio W_{22} não depende dos pesos em W_{21} :

$$\begin{aligned} &= 2(target - (pred22 + pred21)) * \frac{d}{dw_{21}}(-1)(pred22 + pred21) \\ &= 2(target - (pred22 + pred21)) * \frac{d}{dw_{21}}(-1)(w_{21} * ypred1) \end{aligned}$$

Que é a derivada para os pesos do perceptron:

$$= 2(target - (pred22 + pred21)) * (ypred1)(-1)$$

Entretanto, calcular os pesos de w_1 em função da saída requer um pouco mais:

$$\begin{aligned} &\frac{d}{dw_1}(target - (pred22 + pred21))^2 \\ &= 2(target - (pred22 + pred21)) * \frac{d}{dw_1}(-1)(pred22 + pred21) \\ \\ &= 2(target - (pred22 + pred21)) * \frac{d}{dw_1}(-1)(\sum w_{22} \sum w_1 x + \sum w_{21} \sum w_1 x) \end{aligned}$$

Usando a derivada de somas e verificando que os termos não relacionados ao w_1 avaliado somem:

$$2(target - (pred22 + pred21)) * (-1)(\sum w_{22}x + \sum w_{21}x)$$

```
# update em w . Eta ja ajustado para 1/2*eta
delta_w22 <- eta * (-1) * (y[i] - (ypred21 + ypred22)) * ypred1
delta_w21 <- eta * (-1) * (y[i] - (ypred21 + ypred22)) * ypred1
delta_w1 <- eta * (y[i] - (ypred21 + ypred22)) * -1 *
  (sum(w21 %*% c(as.numeric(x[i,]),1)) + sum(w22 %*% c(as.numeric(x[i,]),1)))

w1 <- w1 - delta_w1
w21 <- w21 - delta_w21
w22 <- w22 - delta_w22
ypreds[i] <- out # salva predicao21 atual
yerrors[i] <- ypreds[i] - y[i]
}
print(paste("Mean squared error:", mean((yerrors)^2)))
}
return(ypreds)
}
```

Então, podemos testá-lo em um dataset.

```
>train_df <- iris[, c(1, 2, 3)]
>names(train_df) <- c("s.len", "s.wid", "p.len")
>head(train_df)
>train_df[60:65,]

>x_features <- train_df[, c(1, 2)]
>y_target <- train_df[, 3]

# Convergência boa
>mark_ii_preds <- mark_ii(x = x_features, y = y_target,
                           eta=0.000001, reps = 40)
[1] "This is training epoch: 1"
[1] "Current weights: -0.45050790019773 -0.0197893400687895 -0.0197893400687895"
[2] "Current weights: 0.150011803623929 2.13458518518008 2.13458518518008"
[3] "Current weights: 1.48235899015804 -0.0197893400687895 -0.0197893400687895"
[1] "Mean squared error: 1133.22204821886"
(...)
[1] "This is training epoch: 2"
[1] "Current weights: -0.67126807499406 -0.0609395311239563 -0.0609395311239563"
[2] "Current weights: -0.0707483711724013 2.09343499412492 2.09343499412492"
[3] "Current weights: 1.26159881536171 -0.0609395311239563 -0.0609395311239563"
[1] "Mean squared error: 176.747586724131"
(...)
[1] "This is training epoch: 4"
[1] "Current weights: -0.791488817323548 -0.0700721883119202 -0.0700721883119202"
[2] "Current weights: -0.19096911350189 2.08430233693696 2.08430233693696"
[3] "Current weights: 1.14137807303222 -0.0700721883119202 -0.0700721883119202"
[1] "Mean squared error: 7.32496712284895"
[1] "This is training epoch: 5"
[1] "Current weights: -0.805708526415977 -0.0705118739404967 -0.0705118739404967"
[2] "Current weights: -0.205188822594319 2.08386265130838 2.08386265130838"
[3] "Current weights: 1.12715836393979 -0.0705118739404967 -0.0705118739404967"
[1] "Mean squared error: 3.31246116798174"
(...)
[1] "Mean squared error: 2.50706426321967"
(...)
[1] "Mean squared error: 2.50638029884829"
(...)
[1] "Mean squared error: 2.50640582517322"
```

Podemos observar o modelo convergindo à medida em que os pesos se estabilizam e nossa medida de erro cai. Usando o η acima, a rede costuma convergir com correlação $\rho \sim 0.60$ entre previsões e dados originais.

```

>acc_data <- data.frame(y_preds=mark_ii_preds,
                         y_targs=y_target)

>acc_data$errors <- y_target - mark_ii_preds
>cor.test(acc_data$y_preds, acc_data$y_targs)

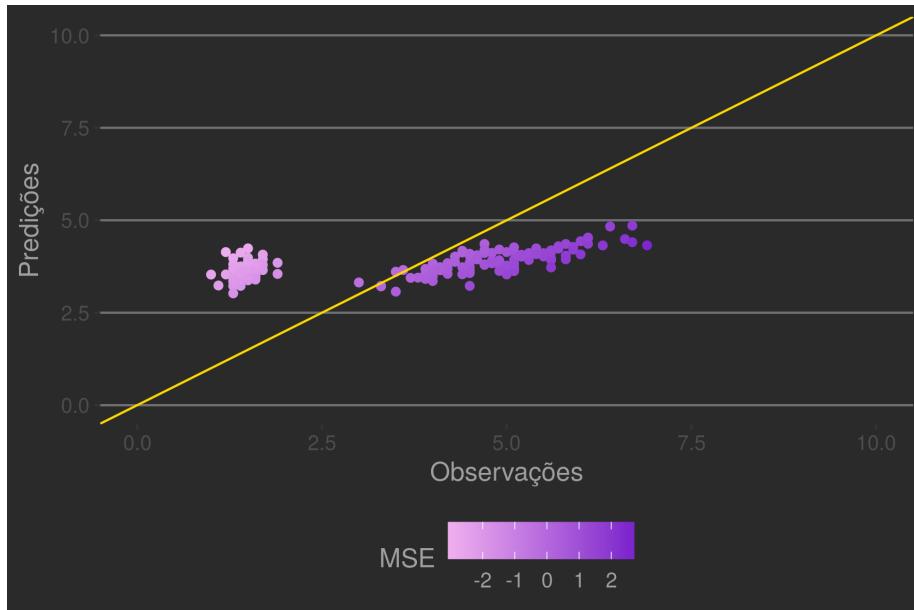
Pearson's product-moment correlation

data: acc_data$y_preds and acc_data$y_targs
t = 8.9717, df = 148, p-value = 1.203e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.4788098 0.6883163
sample estimates:
cor
0.5935271

>ggplot(acc_data,aes(y=y_preds,x=y_targs,color=errors))+  

  geom_point() + xlim(0,10) + ylim(0,10) +
  geom_abline(slope = 1,intercept = 0)

```



De maneira prática, não precisamos calcular os gradientes ou a topologia da rede (número de neurônios, camadas e como estão conectados). Bibliotecas voltadas à aprendizagem de máquina automatizam partes do processo, oferecendo rápida usabilidade para muitos classificadores eficientes. Usando a lib **caret**:

```

> library(caret)
# https://topepo.github.io/caret/train-models-by-tag.html
> train(x=x_features,y = y_target,method = "mlpWeightDecay")
  Multi-Layer Perceptron

150 samples
  2 predictors

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 150, 150, 150, 150, 150, 150, ...
Resampling results across tuning parameters:

  size  decay    RMSE      Rsquared     MAE
  1     0e+00   1.830946  0.3132915  1.5795672
  1     1e-04   1.831956  0.4041400  1.5641681
  1     1e-01   2.203828  0.5889224  1.9507964
  3     0e+00   1.035326  0.6731265  0.8242900
  3     1e-04   1.129702  0.6322950  0.8921468
  3     1e-01   2.230236  0.6531256  1.9114274
  5     0e+00   1.094755  0.6558700  0.8567348
  5     1e-04   1.121093  0.6523228  0.9007250
  5     1e-01   2.143342  0.6639255  1.7652741

RMSE was used to select the optimal model using the
smallest value.
The final values used for the model were size = 3 and decay = 0.

```

Temos $R^2 \sim 0.673$ com 3 unidades escondidas. Outras arquiteturas (e.g. defina `method = "brnn"`) incluem nodos com funções de ativação diferente, assim como variações para o funcionamento de outros pontos.

Referências Para uma história completa sobre redes neurais: J. Schmidhuber. Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61, p 85–117, 2015. (Based on 2014 TR with 88 pages and 888 references, with PDF & LATEX source & complete public BIBTEX file).

<http://web.csulb.edu/~cwallis/artificialn/History.htm> https://sebastianraschka.com/Articles/2015_singlelayer_neurons.html <https://rpubs.com/FaiHas/197581>

Exercícios

1. Uma câmera é posicionada no teto e precisamos criar um algoritmo que determine se a bolinha está do lado esquerdo ou direito. Um perceptron como o apresentado seria capaz de aprender corretamente como indicar posse de bola?
2. Em neurônios biológicos, modelamos a ativação em função da voltagem na membrana neuronal. Quais modelos de função de ativação existem? Consulte softwares livres avançados para simulação de redes em Neural Ensemble (<https://neuralensemble.org/projects/>)
3. Reformule o algoritmo de aprendizagem (loop `for`) para que a taxa de aprendizagem η seja reduzida para $\frac{\eta}{10}$ nos últimos exemplos.
4. Implemente Mark I adaptado para aprender com epochs e teste com gradiente η pequeno.
5. Explore outras arquiteturas de rede neural usando *caret*.

Capítulo 5 : Contexto e Inferência Bayesiana

Probabilidades

“*O provável é aquilo que acontece na maioria das vezes*”, Aristóteles, Retórica.

Uma abordagem probabilística da matemática aplicada que tem se popularizado é a de *Inferência Bayesiana*. Os procedimentos apresentados antes são usualmente chamados de *frequencistas*. Muitas vezes, a informação obtida é quase idêntica, mas a perspectiva muda de forma considerável.

Por princípio, usamos caminhos diferentes.

Frequencistas e Bayesianos

Abordagens frequentistas situam probabilidades como aproximações para cenários com um número infinito de eventos. Os exemplos visitados nos primeiros capítulos muitas vezes faziam essa analogia.

Retomando um exemplo trivial: se jogarmos uma moeda honesta infinitas vezes, a proporção de *caras* tende a que valor? Para muitos sorteios, a proporção tende a 0,5.

Simulação:

```
> set.seed(2600)
> coin_t <- function(x) {
  sample(size=x,x=c(0,1), prob = c(0.5,0.5), replace = T) %>%
  (function(y) sum(y)/length(y))}
> coin_t(3)
[1] 0.6666667
> coin_t(10)
[1] 0.4
> coin_t(30)
[1] 0.5666667
> coin_t(100)
[1] 0.51
> coin_t(1000)
[1] 0.498
> coin_t(100000)
[1] 0.50098
> coin_t(10000000)
[1] 0.4999367
```

É comum a ideia de populações ou procedimentos hipotéticos infinitos.

O método hipotético-dedutivo relaciona teorias a observações através de hipóteses falseáveis. A concepção mais aceita, compilada recentemente por K. Popper, trata diretamente de probabilidades como entes importantes para as ciências naturais.

Mais que isso, ilustra o conceito de calcular a plausibilidade de resultados experimentais na vigência de uma hipótese em estudo.

Calculamos uma probabilidade associada à ocorrência de uma observação. No teste t para duas amostras (capítulo 1), definimos a hipótese nula em função das médias dos bicos(μ) e outros parâmetros (σ, df). $H_0 : \mu_{amostra_1} = \mu_{amostra_2}$. O procedimento de imaginar os eventos observados como instâncias de uma família de eventos semelhantes se adequada perfeitamente a preceitos Popperianos. Continua sendo o feijão com arroz da ciência normal para testar previsões de um determinado paradigma. O refinamento gradual de uma teoria envolve o acúmulo de conhecimento e testagem de *hipóteses auxiliares* resultantes de premissas basilares (*hard core* na terminologia de Imre Lakatos).

Prismas bayesianos instrumentalizam probabilidades como entes primitivos, noções mais básicas relacionadas a *plausibilidade*, *grau de crença*, *expectativa* para uma determinada situação. O ponto chave é de que deixamos de guiar os procedimentos objetivando uma probabilidade para os eventos. As probabilidades em si passam a ser entidades centrais. Especificamente, como nossas crenças sobre algo mudam após observações.

No caso dos pássaros:

Inferência Frequencista: Supondo que a diferença média entre tamanho dos bicos seja 0, qual a probabilidade para minhas observações?

Sendo H_0 definida por $H_0 : \mu_{amostra_1} = \mu_{amostra_2}$, queremos saber:

$$P(H_0) < 0,05?$$

Inferência Bayesiana: Quais as probabilidades associadas aos valores possíveis para a diferença entre $\mu_{amostra_1}$ e $\mu_{amostra_2}$? Considerando um modelo e os dados, qual a distribuição probabilística de $\mu_{diff_{1-2}}$

$$P(\mu_{diff_{1-2}}) = ?$$

Além de construtos intuitivos, uma plataforma bayesiana oferece dois recursos poderosos: sensibilidade a informações prévias sobre um fenômeno (*priors*) e estimadores estocásticos (e.g. *Markov Chain Monte Carlo*). Assim, podemos (1) fazer uso de informações arbitrárias (e.g. intuição de um especialista) e (2) reduzir a dependência de soluções analíticas (fechadas) para equações que descrevem os modelos.

Epistemologia Bayesiana? Antes, associamos cenários a hipóteses e estimamos parâmetros (probabilidades) para testá-las. Agora, os *parâmetros* têm um papel conceitual mais central.

Um parâmetro é um símbolo, uma aproximação para uma ideia (*para*, “perto”, *metron*, “medida”). Nos capítulos iniciais, usamos parâmetros para construtos que se comportam como números (e.g: existem elementos que podem ser ordenados por alguma noção de tamanho e operações, como soma e multiplicação).

Estimamos parâmetros (μ_{diff} e valor p) para testar uma hipótese sobre a diferença média entre tamanho dos bicos nas espécies A e B. No capítulo 2, um parâmetro (β e um valor p) para testar uma hipótese sobre a correlação entre expectativa de vida saudável e número de médicos em um país. Mais do que isso, usamos estatísticas para testar hipóteses e calcular intervalos de confiança.

É muito difícil entender a utilidade dos procedimentos anteriores desconhecendo o norte hipotético-dedutivo guiando-os. O seguinte trecho está em *Data Analysis, A Bayesian Tutorial* (Sivia & Skilling, 2006), de professores da Oxford: “*The masters, such as Fisher, Neyman and Pearson, provided a variety of different principles, which has merely resulted in a plethora of tests and procedures without any clear underlying rationale. This lack of unifying principles is, perhaps, at the heart of the shortcomings of the cook-book approach to statistics that students are often taught even today.*”

Podemos, inclusive, usar probabilidades obtidas via inferência bayesiana para continuar testando hipóteses. Entretanto, é conveniente introduzir ferramentas bayesianas junto ao pensamento de filósofos que ofereceram outras alternativas²⁸.

Muitos métodos científicos: Feyerabend, Carnap e Quine

No primeiro capítulo, entramos em contato com o método hipotético-dedutivo e a falseabilidade como critério de demarcação científica. Apesar de dominante, esse racional possui vulnerabilidades interessantes. Entenderemos melhor argumentos contrários e propostas alternativas através de três filósofos do século XX. Esse é um momento conveniente, uma vez que tiramos os holofotes das hipóteses.

Paul Feyerabend (1924 - 1994)

Conhecido pela personalidade ímpar e por ideias radicais, Paul Feyerabend, em *Contra o Método*(1975), argumenta que boa parte dos avanços significativos aconteceram fora do método científico.

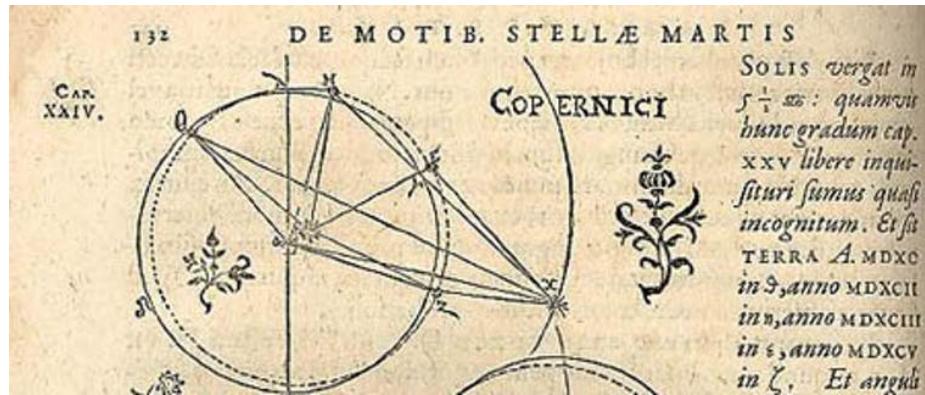
Crenças pessoais e detalhes biográficos são responsáveis por mudanças em nosso conhecimento. Mais que isso, usar falsificabilidade e o método hipotético-dedutivo teriam nos feito rejeitar o heliocentrismo e outras ideias chave para o progresso. Na verdade, o sistema geocêntrico (Terra no centro do sistema) de Ptolomeu era mais acurado (!) que o de Copérnico (Sol ao centro) usando um mesmo número de parâmetros para cálculos das órbitas. O modelo copernicano estava mais próximo da realidade como entendida hoje, porém o estágio intermediário de concepção teórica era ‘pior’ ²⁹.

Além de menos acurado, era mais complexo em alguns aspectos, incluindo mais epiciclos: órbitas auxiliares usadas como artifício para cálculos. A Revolução

²⁸ Existe um programa de pesquisa mais abrangente em filosofia sobre epistemologia Bayesiana, mas este não é nosso foco. Consulte The Open Handbook of Formal Epistemology

²⁹ Stanley E. Babb, “Accuracy of Planetary Theories, Particularly for Mars”, Isis, Sep. 1977, pp. 426

Copernicana somente consolidou a mudança de paradigma com contribuições subsequentes de Tycho Brahe, Kepler, Galileo e Newton, cerca de 1 século depois.



Diante das incongruências entre um método e as inevitáveis imprevisibilidades da empreitada humana em conhecer o Universo, Feyerabend propõe o *anarquismo epistêmico* sob o mote “*Anything goes*” ('Vale tudo'). Isto é, quaisquer recursos são válidos na tentativa de atacar um problema ou conceber um modelo de realidade.

É tentador pensar que, dada a profundidade do trabalho, a defesa de uma postura tão contundente é obviamente uma aplicação dos preceitos defendidos no livro como necessários para disseminar uma idéia. Outros filósofos nos ajudam a conceber uma ciência não pautada num método hipotético-dedutivo de maneira menos radical.

Rudolph Carnap (1891 - 1970)

Carnap, do Círculo de Viena, também contrapôs Popper. Em “Testability and Meaning” (1936-7), argumenta que falsificabilidade não difere de verificacionismo. Envolve a testagem de cada assertiva em si, um problema que outros também endereçaram.

Diante de resultados inesperados em um experimento, o procedimento automático para um cientista envolve checar a integridade das condições desenhadas. Verificar a composição da amostra, os métodos de coleta, mecanismos de perda, critérios de exclusão e inclusão, premissas da análise. Isso não é desonestade intelectual: são fatores menores reais e facilmente abordáveis que podem ter invalidado a teoria de base. O mesmo se dá para técnicas de análise e conceptualização de construtos.

O cuidado com esses pontos é deseável e desnuda o inevitável calcanhar de Aquiles da falsificabilidade.

É impossível refutar uma hipótese/assertiva de maneira isolada. Cada procedimento experimental ou lógico envolve a interdependência entre os símbolos usado.

Willard van Orman Quine (1908 – 2000)

Uma escola filosófica parte do problema acima. A tese de Duhem-Quine postula que é impossível testar qualquer hipótese científica, uma vez que sempre há premissas aceitas como verdade.

Em ‘*Os dois dogmas do empiricismo*’, Quine considera as proposições e as relações lógicas entre elas apenas um sistema, que só pode ser estudado em conjunto.

Os exercícios ilustrados no volume anterior testa a adequação dos dados à família de distribuições t. Também assume que tamanhos dos bicos são mensuráveis usando números e que estes podem ser comparados com valores de outras amostras.

A princípio, essas declarações parecem triviais. Entretanto, considerando os fatores humanos da ciência, a mudança de lentes é significativa. Discutivelmente, abordar um problema dessa maneira é historicamente mais frutífero. As contribuições mais contundentes são advindas de cientistas dedicados a estudar um contexto ou problema como um todo. É raro, talvez inédito, que um grupo testando hipóteses sem um eixo consistente tenha obtido avanços admiráveis.

Estimar livremente os parâmetros de que falamos naturalmente é muito mais intuitivo que adequar uma ideia aos procedimentos hipotético-dedutivos.

Inferência Bayesiana

No capítulo 1, ao fazer um teste t, calculamos a estatística t correspondente às diferenças encontradas e então a probabilidade de obter valores iguais ou mais

extremos.

É possível usar inferência bayesiana para analisar uma situação idêntica. Como aludido antes, não estamos muito interessados no valor p.

A pergunta é “*Quais são os valores prováveis para a diferença entre A e B?*”.

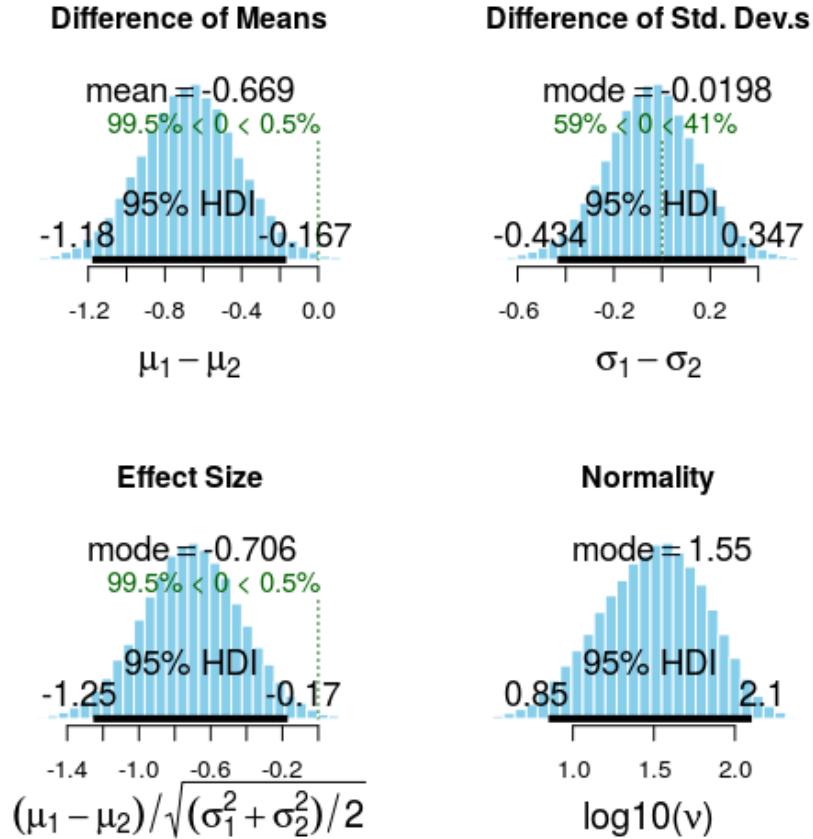
A distribuição probabilística obtida representa nossas crenças na plausibilidade de cada valor.

Usando a library BEST e 30 observações retiradas de amostras de distribuição normal ($\mu_a = 0$; $\mu_b = 0.6$; $\sigma_a = \sigma_b = 1$) normais.

```
> library(ggthemes)
> library(rstan)
> library(reshape2)
> library(BEST)
> library(ggplot2)
> options(mc.cores = parallel::detectCores() - 1)
> set.seed(2600)
> a <- rnorm(n = 30, sd = 1, mean = 0)
> b <- rnorm(n = 30, sd = 1, mean = 0.6)

# BEST
> BESTout <- BESTmcmc(a, b)

### BEST plots
> par(mfrow=c(2,2))
> sapply(c("mean", "sd", "effect", "nu"), function(p) plot(BESTout, which=p))
> layout(1)
```



A distribuição no canto superior esquerdo corresponde às nossas estimativas para possíveis valores da diferença entre A e B. Podemos usar a média como estimativa pontual: ($diff_{\mu_a\mu_b} = -0.669$). O intervalo apontado como 95% HDI (High density interval) contém 95% da distribuição. Seu significado é mais próximo da intuição de uma região provável para os valores que o clássico intervalo de confiança.

Por trás das cortinas

Obviamente, vamos entender a arte envolvida aqui. A flexibilidade e o poder dos modelos bayesianos permite lidar com uma série de problemas dificilmente tratáveis de outra forma. Entretanto, é fácil cair em armadilhas ou esbarrar em dificuldades durante o processo.

É extremamente importante entender os componentes envolvidos para não cometer erros importantes.

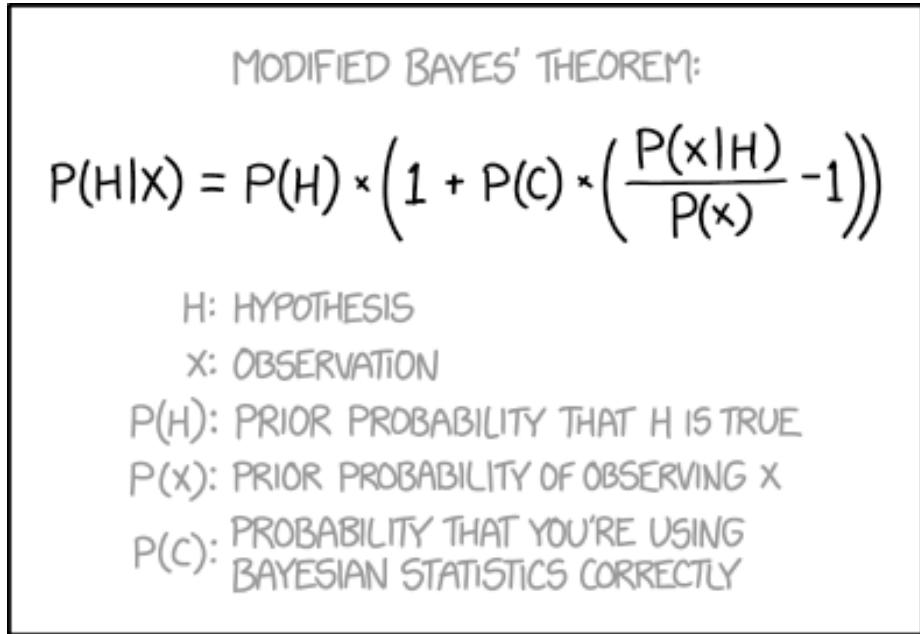


Figure 29: <https://xkcd.com/2059/> Teorema de Bayes modificado, incluindo a probabilidade de você estar usando estatística bayesiana correntemente

O Teorema do Bayes

$$P(B | A) = \frac{(A | B)P(B)}{P(A)}, P(A) \neq 0$$

É a forma célebre do teorema e nos conta sobre probabilidades de eventos subsequentes/concorrentes.

Costuma ser apresentado para tratar problemas simples: *sabendo o resultado de um teste médico positivo, qual a probabilidade de o paciente ter a doença?*. O teorema de Bayes relaciona a probabilidade basal da doença com a probabilidade um teste positivo subsequente. Algumas armadilhas da intuição são quebradas: ainda que o teste tenha boa sensibilidade (probabilidade alta de resultado positivo diante da doença), a probabilidade será baixa se as chances basais também forem.

O teorema foi concebido num esforço maior do reverendo (Thomas Bayes, 1701-1761) para um problema de inferência. Curiosamente, ele é bastante semelhante ao que empreenderemos.

Suponha que atribuímos uma probabilidade $p(0 \leq p \leq 1)$ para o lançamento de uma moeda com resultado *coroa*. Ao observar alguns resultados, podemos calibrar nossa estimativa. Podemos começar supondo uma moeda honesta 0.5. Com uma frequência alta de *coroas*, é racional aumentar a nossa estimativa sobre

o valor de p ($p \sim 1$). Bayes demonstrou como fazer essas atualizações diante de evidência.

Intuições O texto de **An essay towards solving a Problem in the Doctrine of Chances (1973)** apresenta uma série de demonstrações até chegar ao enunciado:

Proposition 4 : *If there be two subsequent events be determined every day, and each day the probability of the 2nd [event] is $\frac{b}{N}$ and the probability of both $\frac{P}{N}$, and I am to receive N if both of the events happen the 1st day on which the 2nd does; I say, according to these conditions, the probability of my obtaining N is $\frac{P}{b}$. (...)*

O estilo é um pouco complicado. Com notação atual:

Considerando dois eventos subsequentes, (1) a probabilidade do segundo acontecer é $\frac{b}{N}$ ($P(A)$), (2) a probabilidade de ambos acontecerem é $\frac{P}{N}$ ($P(A \cap B)$). (3) Sabendo que o segundo aconteceu, a probabilidade de o primeiro também ter acontecido é $\frac{P}{b}$. N é cancelado e (3) é a razão entre (2) e (1):

$$P(B | A) = \frac{P(A \cap B)}{P(A)}, P(A) \neq 0$$

Considerando dois eventos, **A** e **B**, a probabilidade de B acontecer sabendo que A aconteceu ($P(B | A)$) é idêntica à probabilidade de A e B ($P(A \cap B)$) acontecerem, normalizada pela probabilidade de A acontecer individualmente.

Pela definição de probabilidade condicional, $P(A \cap B) = P(A | B)P(B)$, então:

$$P(B | A) = \frac{(A | B)P(B)}{P(A)}, P(A) \neq 0$$

Assim, podemos estimar probabilidades de eventos. Em inferência Bayesiana, empregamos o teorema para estimar os valores prováveis (distribuição probabilística) de um parâmetro (θ) diante de observações (X).

$$P(\theta | X) = \frac{P(X | \theta)P(\theta)}{P(X)}, P(X) \neq 0$$

Posterior Chamamos o primeiro termo, a estimativa do parâmetro após a calibração pelas observações $P(\theta | X)$, de **distribuição posterior**(*posterior distribution* traduz bem para o português). Todos os procedimentos são desenhados para calculá-la e representa a distribuição usada nas inferências finais. Por exemplo, queremos a distribuição posterior dos valores para a diferença entre A e B.

Probabilidade marginal O denominador do termo à direita é a probabilidade independente para a ocorrência dos dados ($P(X)$). É usada para normalizar as quantidades e chamada de probabilidade/verossimilhança marginal, **marginal likelihood**, ou ainda evidência do modelo, **model evidence**.

Likelihood O primeiro termo à direita, $P(X | \theta)$, chamamos de verossimilhança (**likelihood**) e determina a probabilidade de ocorrência das observações $P(X)$ dado um parâmetro θ .

Provavelmente, é o ponto mais sensível na modelagem, pois descreve como se dá a relação entre modelo teórico e observações. Como discutido antes, equações correspondem a leis precisas envolvendo mais de um construto. O mapeamento entre observações $P(X)$ e um parâmetro é dado pela *função de verossimilhança (likelihood function)* escolhida, $f(\theta)$.

Exemplo: o número de células de combate do sistema imune circulante no sangue está associado a uma resposta inflamatória. Quanto mais alto, mais provável é uma infecção para o médico. Mas qual a lei que associa o número de células (entre 0 e 10^5) com a probabilidade de infecção?

Se os desfechos estudados são binários ($y_i \in \{0, 1\}$, e.g. diagnóstico positivo ou negativo), podemos usar uma relação logística (ver Cap. 4) para estimar probabilidades em função de variáveis observadas (X) e parâmetro(s) θ .

$$P(X | \theta) \sim f(X, \theta) : y_i = \frac{1}{1 + e^{-(\theta * x_i + c)}}$$

Outras funções poderia ser escolhidas (e.g. Heaviside step do capítulo anterior). Isto depende do fenômeno, da teoria e das medidas analizadas.

Priors Como estimamos as probabilidades infecção antes de ver os resultados do teste? Antes exame, temos alguma noção de como o parâmetro se comporta. Ela pode ser bem precisa ou trazer muita incerteza. Chamamos a estimativa basal $P(\theta)$ de **prior** e aparece na expressão multiplicando o valor da verossimilhança. Em linguagem das probabilidades, ela é uma distribuição. Nossas crenças prévias podem ser pouco informativas (e.g. não examinamos o paciente; distribuição uniforme sobre os valores possíveis) ou bastante definidas (e.g. o paciente está assintomático; distribuição concentrada nas proximidades de 0).

```
> a <- runif(10000)
> b <- runif(10000, min = 0, max=0.2)
> priors <- data.frame(uniform=a, low=b)
> ggplot(priors)+ 
  geom_density(aes(x=uniform),color="#F0E442")+
  geom_jitter(aes(y=uniform*4.5,x=seq(0,0.2,length.out = 10000)),
  color="#009E73",alpha=0.015)+
  geom_density(aes(x=low),color="#009E73")+
  geom_jitter(aes(y=low*3,x=seq(0,1,length.out = 10000)),
```

```

color="#F0E442",alpha=0.01)+ylab("Densidade")+
xlab("Priors: informativo(verde) ou incerto (amarelo)")+
theme_hc(style="darkunica")+theme(axis.text.y=element_blank())

```

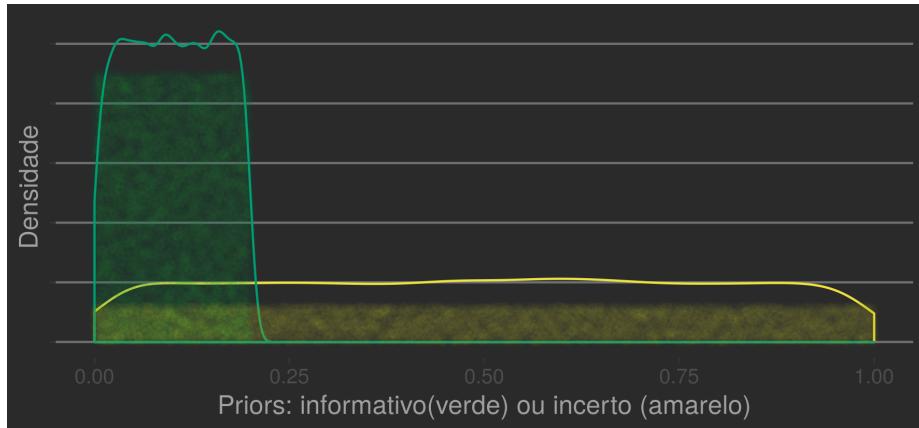


Figure 30: O prior verde supõe maiores probabilidades para valores baixos. O prior amarelo é pouco informativo, atribuindo probabilidades semelhantes em todo o intervalo

Conhecendo nossos construtos, podemos então reescrever os procedimentos:

$$\text{Posterior} = \frac{\text{Prob. de observações dada por } f(X, \theta) * \text{Prior}}{\text{Prob. marginal para observações}}$$

Para obtermos o *posterior*, multiplicamos a probabilidade dada pela *função de verossimilhança* pelas nossas estimativas prévias (*prior*) e normalizamos pela *probabilidade marginal* das observações.

As narrativas posteriores são construídas de acordo com a distribuição do *posterior*.

Mestre Foo e o Recrutador³⁰

Um recrutador técnico, ao descobrir que os caminhos dos hackers Unix lhe eram estranhos, buscou conversar com Mestre Foo para aprender sobre o Caminho. Mestre Foo encontrou o recruta nos escritórios de recursos humanos de uma grande corporação.

O recruta disse, “Eu tenho observado que os hackers Unix desdenham ou ficam nervosos quando pergunto a eles quantos anos de experiência têm em uma linguagem de programação nova. Por que isso acontece?”

³⁰<http://www.catb.org/~esr/writings/unix-koans/recruiter.html>

Mestre Foo levantou e começou a caminhar pelo escritório. O recrutador ficou intrigado, e perguntou “O que está fazendo”?

“Estou aprendendo a andar”, replicou Mestre Foo.

“Eu vi você entrar pela porta andando” o recrutador exclamou, “e você não está tropeçando em seus pés. Obviamente, você sabe como andar.”

“Sim, mas este piso é novo para mim” replicou Mestre Foo.

Ao ouvir isso, o recrutador foi iluminado.

Dear Stan

As implementações dos modelos Bayesianos são feitas em Stan, um pacote em C++ especializado em inferência bayesiana. Os modelos são escritos num dialeto próprio, mas a sintaxe é bastante semelhante à da notação matemática, então a tradução das análises do capítulo é direta.

Especificamos o modelo num arquivo auxiliar de extensão *.stan*, que é manipulado por pacotes em R para visualização e outras utilidades.

Lá e de volta outra vez

Reproduziremos à maneira bayesiana dois exemplos conhecidos: diferença entre médias (análogo ao test t) e correlação.

Aqui, fica claro que o racional é mais direto que o anterior.

Comparando amostras de distribuição normal Lembremos (cap. 1) que, para comparar amostras usando o teste t: (1) assumimos normalidade na origem dos dados; (2) imaginamos a distribuição das médias normalizadas pelo erro padrão em amostras hipotéticas semelhantes, retiradas da mesma população; (3) calculamos o valor p conhendo a distribuição (Student's t).

Agora, podemos obter uma distribuição posterior para a diferença entre amostras. (1) Assumimos a normalidade na origem dos dados (likelihood function); (2) fornecemos nossas estimativas prévias (prior); (3) atualizamos os valores diante dos dados e para obter o posterior.

Adotamos a seguinte parametrização:

Valores observados nas amostras 1 e 2, vetores N dimensões: y_1, y_2
Parâmetros alvo desconhecidos, as médias em cada amostra e um desvio-padrão em comum: μ_1, μ_2, σ

Priors supondo média 0 em ambos os grupos e desvio-padrão de 1: $\mu_1 \sim N(0, 1), \mu_2 \sim N(0, 1), \sigma \sim N(1, 1)$ Função de verossimilhança, indicando que cada observação é de uma população com distribuição normal: $y \sim N(\mu, \sigma)$

Também especificamos para o Stan que gere (1) valores para diferença entre as distribuições posteriores de μ_1 e μ_2 , μ_{diff} e (2) tamanho de efeito com D de Cohen, dividindo o valor pelo desvio-padrão.

O código deve ser salvo num arquivo “.stan”.

```
data {
    int<lower=0> N;
    vector[N] y_1;
    vector[N] y_2;
}
parameters {
    real mu_1;
    real mu_2;
    real sigma;
}
model {
    //priors
    mu_1 ~ normal(0, 1);
    mu_2 ~ normal(0, 1);
    sigma ~ normal(1, 1);

    //likelihood - Verossimilhança
    for (n in 1:N){
        y_1[n] ~ normal(mu_1, sigma);
        y_2[n] ~ normal(mu_2, sigma);
    }
}
generated quantities{
    real mudiff;
    real cohenD;

    mudiff = mu_1 - mu_2;
    cohenD = mudiff/sigma;
}
```

Então, vamos iniciar as análises através da interface em R. Criamos uma lista com componentes homônimos às variáveis do arquivo Stan (y_1: amostral 1, y_2: amostral 2, N: tamanho amostral).

```
> a <- rnorm(n = 100, sd = 1, mean = 0)
> b <- rnorm(n = 100, sd = 1, mean = 0.6)
> sample_data <- list(y_1=a,y_2=b,N=length(a))
> fit <- rstan::stan(file="aux/bayes-t.stan",
  data=sample_data,
  iter=3000, warmup=100, chains = 6)
SAMPLING FOR MODEL 'bayes-t' NOW (CHAIN 1).
```

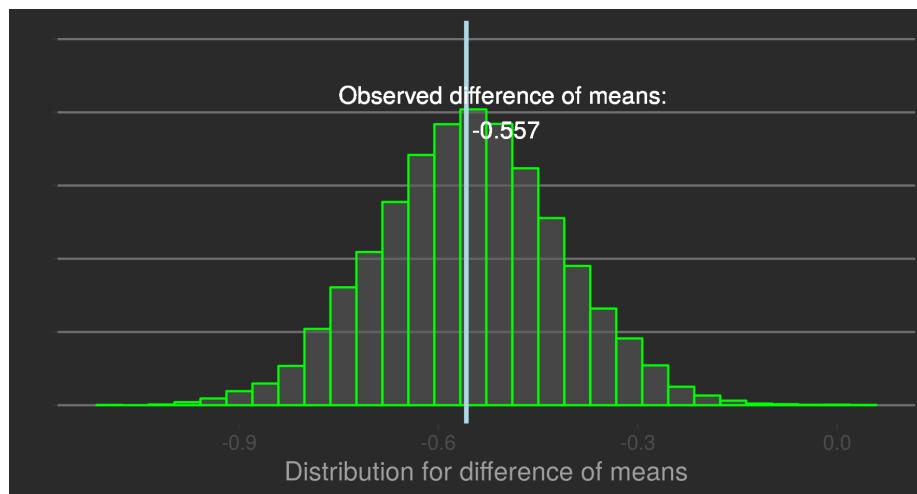
(...)

O comando acima iniciará os cálculos. Vamos plotar as distribuições posteriores de μ_1 , μ_2 e a diferença entre essas (μ_{diff})

```
> obs_diff <- mean(a) - mean(b)
> obs_diff
[1] -0.5579295
> posteriors <- extract(fit, par = c("mu_1", "mu_2", "mudiff"))
> lapply(posteriors, mean)
$mu_1
[1] 0.07303457

$mu_2
[1] 0.6261336

$mudiff
[1] -0.553099
> ggplot(data.frame(muDiff=posteriors$mudiff), aes(x=muDiff))+
  geom_histogram(alpha=0.6, color="green")+
  geom_vline(xintercept=obs_diff,
             color="light blue", size=1) # line for observed difference
  xlab("Distribuição para diferença de médias") + ylab("") + ylim(0, 2500) +
  geom_text(label="Diferença observada:\n -0.557",
            color="white", x=mean(muDiff)+0.05, y=2000) +
  theme_hc(style="darkunica") +
  theme(axis.text.y=element_blank())
```



A distribuição acima contém outras informações. Perdemos a elegante estimativa analítica de Student para testar a hipótese sobre um parâmetro(e.g. $H_0 : \mu_{\text{diff}} =$

0). Por outro lado, temos uma visão global sobre toda a distribuição estimada para μ_{diff} !

Correlação linear Reproduziremos a análise de correlação do capítulo 2, quando falamos em indicadores de saúde. As variáveis importantes são o logaritmo do número de médicos e a expectativa de vida saudável (Health Adjusted Life Expectancy). O banco foi criado com nome `uni_df`, contendo as variáveis `log_docs` e `hale`.

Sistematizando nossa abordagem, vamos escolher **Priors**:

Correlação ρ : Vamos assumir que ela é positiva entre número de médicos e expectativa de vida saudável. Vamos indicar um valor baixo (0,1) para essa correlação.

$$N(0.1, 1)$$

Médias e desvios μ e σ : Não temos muita ideia média para o logaritmo do número de médicos. Uma leve inspeção mostra que os valores têm baixa magnitude. Vamos indicar priors pouco informativos para $\mu_{\text{medicos}}, \sigma_{\text{medicos}}$ na forma de gaussianas de média 0 e desvios altos.

$$\mu_{\text{medicos}} \sim N(0, 2), \sigma_{\text{medicos}} \sim N(0, 10)$$

Uma breve consulta em mecanismos de busca sugere que uma média μ_{hale} de 60 anos seja um chute razoável. Vamos estimar o prior do desvio-padrão σ_{hale} em 5.

$$\mu_{\text{hale}} \sim N(60, 3), \sigma_{\text{hale}} \sim N(5, 2)$$

Likelihood function: Nosso modelo para os dados é de que eles são dados através de uma distribuição normal bivariada, com médias μ_1, μ_2 e desvios σ_1, σ_2 . Como vimos antes, a definição para o coeficiente de Pearson entre as amostras X e X' é

$$\rho_{XX'} = \frac{\text{cov}(X, X')}{\sigma_X \sigma_{X'}}$$

Então,

$$\text{cov}(X, X') = \sigma_X \sigma_{X'} * \rho_{XX'}$$

Podemos então definir a matriz de covariância de nossa distribuição bivariada:

$$\text{Cov. Matrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_{2'} * \rho \\ \sigma_1 \sigma_{2'} * \rho & \sigma_2^2 \end{pmatrix}$$

Nosso código em Stan:

```

data {
    int<lower=1> N;
    vector[2] x[N];
}

parameters {
    vector[2] mu;
    real<lower=0> sigma[2];
    real<lower=-1, upper=1> rho;
}

transformed parameters {
    // Matriz de covariancias
    cov_matrix[2] cov = [[      sigma[1] ^ 2      , sigma[1] * sigma[2] * rho],
                         [sigma[1] * sigma[2] * rho,      sigma[2] ^ 2      ]];
}

model {
    // Priors
    sigma ~ normal(0,1);
    mu ~ normal(0.2, 1);

    // Likelihood - Bivariate normal
    x ~ multi_normal_lpdf(mu, cov);
}

generated quantities {
    // Amostras com pares ordenados
    vector[2] x_rand;
    x_rand = multi_normal_rng(mu, cov);
}

```

E então podemos iniciar as estimativas.

```

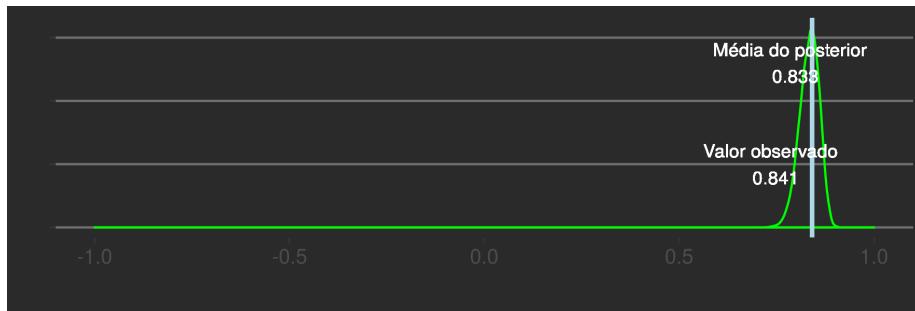
# Stan não aceita missing values
> c_cases <- uni_df[complete.cases(uni_df[,c(3,4)]),]
> vec_2 <- matrix(data = c(c_cases$hale,c_cases$log_docs),ncol = 2,nrow = 145)
> health_data <- list(N=nrow(c_cases),x = vec_2)
> fit <- rstan::stan(file="aux/corr-docs.stan",
                      data=health_data,
                      iter=3000, warmup=120, chains = 6)
SAMPLING FOR MODEL 'corr-docs' NOW (CHAIN 1).
(...)
```

E então, vamos observar nossa estimativa posterior para o valor de ρ :

```

> obs_rho <- cor.test(vec_2[,1],vec_2[,2])$estimate
> posterior <- rstan::extract(fit,par = c("rho"))
> ggplot(data.frame(rho=posterior$rho), aes(x=rho))+ 
  geom_density(alpha=0.6,color="green")+
  geom_vline(xintercept=obs_rho,
             color="light blue",size=1) # line for observed difference
  xlab("")+ylab("")+ xlim(-1,1)+ 
  geom_text(label="Valor observado \n 0.841",
            color="white",x=obs_rho-0.1, y = 5,
            size=3)+ 
  geom_text(label="Média do posterior \n 0.833",
            color="white",x=obs_rho-0.05, y = 13,
            size=3)+ 
  theme_hc(style="darkunica")+
  theme(axis.text.y=element_blank())

```



Notamos que as estimativas posteriores para ρ foram razoavelmente distribuídas ao redor do valor empíricamente calculado na amostra. Podemos ainda observar na distribuição intervalos com alta densidade de probabilidade (HDI, High density intervals) ou ainda outros fins.

```

> quantile(posterior$rho,probs = c(0.025,0.5,0.975))
  2.5%      50%     97.5%
0.7790645 0.8353651 0.8777544
> cor.test(vec_2[,1],vec_2[,2])$conf.int
[1] 0.7854248 0.8828027

```

O HDI muitas vezes é próximo do intervalo de confiança como calculado tradicionalmente, mas isso não é garantido.

Podemos plotar nossa amostra aleatória gerada a partir do posterior e inspecionar visualmente como os valores da amostra estariam dentro da probabilidade estimada.

```

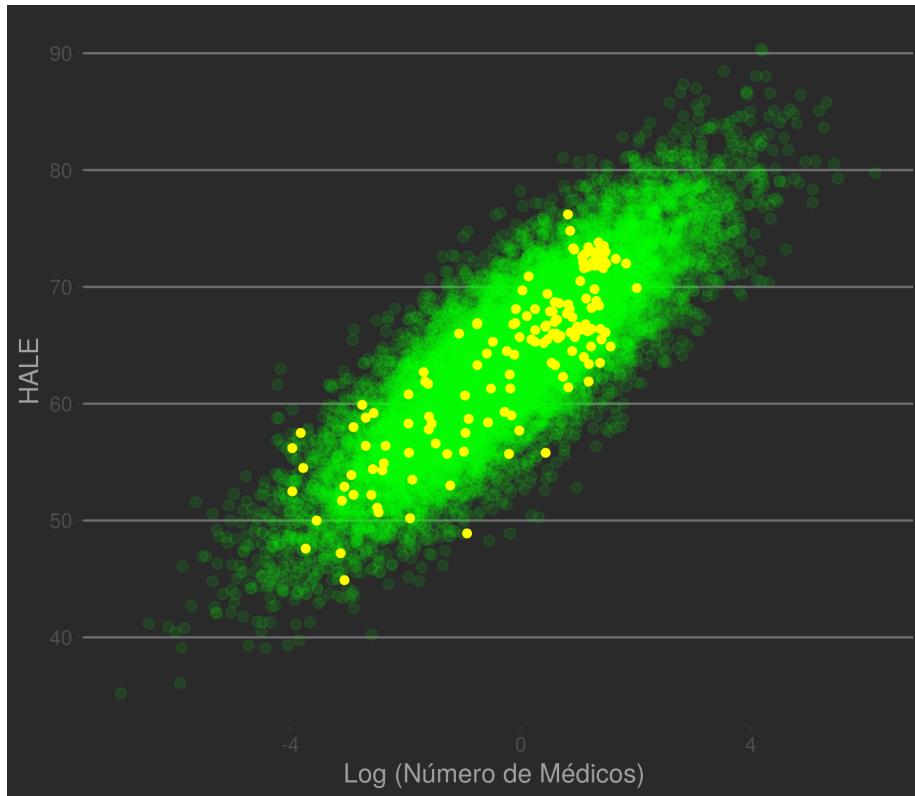
>x.rand = extract(fit, c("x_rand"))[[1]]
>plot(uni_df[,c("log_docs","hale")],
      xlim=c(-5,5), ylim=c(20, 100), pch=16)

```

```

>dataEllipse(x.rand, levels = c(0.75,0.95,0.99),
             fill=T, plot.points = FALSE)
> sample_data <- data.frame(x.rand)
> names(sample_data) <- c("HALE","Logdocs")
> ggplot(sample_data,aes(x=Logdocs,y=HALE))+ 
  geom_point(alpha=0.1,color="green",size=2)+ 
  xlab("Log (Número de Médicos) ") + ylab("HALE")+
  geom_point(data=uni_df,aes(x=log_docs,y=hale),color="yellow")+
  theme_hc(style="darkunica")

```



Você pode experimentar com diferentes priors (famílias e parâmetros) observando como o valor final muda.

Estimadores e métodos Markov Chain Monte Carlo

Nas implementações acima, partimos da equação envolvendo priors, likelihood e probabilidades marginais.

$$P(\theta | X) = \frac{P(X | \theta)P(\theta)}{P(X)}, P(X) \neq 0$$

Usando Stan, informamos priors, a função de verossimilhança, observações e todo o trabalho sujo é realizado sem mais esforços.

A estimativa de $P(\theta | X)$ pode ser feita de diferentes maneiras.

Uma delas envolve partir de uma distribuição $P(\kappa)$ e gradualmente minimizar uma medida da diferença (em geral, a *divergência de Kullback-Leibler*) entre ela e $P(\theta | midX)$. Esses métodos (cálculo variacional, *Variational Bayesian methods*) envolvem soluções analíticas para cada modelo.

Abordaremos um outro método: **Markov Chain Monte Carlo**.

Nem todos que andam sem destino estão perdidos ³¹

Soluções fechadas

Quando falamos em regressão (Cap. 2), estimamos as inclinações de reta β_i . Lançamos mão de uma *função de verossimilhança (likelihood function)*, com o mesmo sentido aqui empregado, definindo a probabilidade das observações dado um modelo teórico.

Obtivemos soluções que maximizassem essa função (*maximum likelihood*). Para o caso da regressão linear, apontamos soluções fechadas

$$\begin{aligned} & \text{Max log likelihood}(\beta_0, \beta_1, \sigma^2) \\ &= \text{Max log} \prod_{i=1}^n P(y_i|x_i; \beta_0, \beta_1, \sigma^2) \end{aligned}$$

Por exemplo, o coeficiente angular (β_1) é

$$\hat{\beta}_1 = \frac{\text{cov}(XY)}{\sigma_x^2}$$

Gradient Descent

No capítulo 4, mostramos outra maneira de estimar parâmetros, analisando uma função de perda. Usando derivadas parciais, calculamos o gradiente, análogo à *inclusão* de uma superfície em 3 dimensões. Isso foi possível pois sabíamos as derivadas em cada nodo (neurônio). A rede consiste no sequenciamento de unidades em camadas, então a regra cadeia funciona perfeitamente (*backpropagation*).

$$(g \circ f)' = (g' \circ f)f'$$

³¹ All that is gold does not glitter,/Not all those who wander are lost; The old that is strong does not wither,/ Deep roots are not reached by the frost./From the ashes, a fire shall be woken,/A light from the shadows shall spring;/Renewed shall be blade that was broken,/The crownless again shall be king. **J.R.R Tolkien. The Fellowship of the ring 1954,**

Markov Chain Monte Carlo

Estimadores Markov Chain Monte Carlo (MCMC) funcionam para tratar problemas sem solução fechada e em que não sabemos os gradientes com exatidão. Outras formas de tratamento existe. Aqui abordamos uma estratégia de MCMC chamada Metropolis-Hastings. Para estimar nosso posterior, $P(\theta | \text{mid}X)$, usamos um algoritmo que permite obter amostras representativas de $P(\theta | \text{mid}X)$. Para isso, a condição é de que exista uma função $f(x)$ proporcional à densidade de $P(\theta | \text{mid}X)$ e que possamos calculá-la.

1 - Começamos com parâmetros em um estado (e.g. $s_0 : \beta_0 = 0.1, \beta_1 = 0.2$) e analisamos a função (e.g. f : log likelihood function) naquele estado ($f(s_0)$) considerando os parâmetros em s_0 . 2 - Em seguida, damos um passo em direção aleatória, modificando dos valores de β_i . Uma opção bastante usada é a de uma gaussiana com centro no estado anterior (*random walk*). Reavaliarmos o estado ($f(s_1)$).

2.1 - Se ele é mais provável, $f(s_1) > f(s_0)$, então s_1 é aceito como novo ponto de partida.

2.2 - Se ele é menos provável, mas próximo o suficiente do estado anterior, $f(s_1) - f(s_0) < \epsilon$, também tomamos s_1 como ponto de partida para o próximo passo aleatório.

2.3 - Se ele é menos provável com uma margem grande, $f(s_1) - f(s_0) > \epsilon$, rejeitamos s_1 e sorteamos um novo estado aleatório.

O processo caminha para estados mais prováveis, com alguma probabilidade de visitar estados menos prováveis. Se a função escolhida é proporcional à densidade do posterior, $f(x) \sim \text{dens}(P(\theta | \text{mid}X))$, as frequências de parâmetros na amostra de estados visitados, s_i , correspondem ao posterior. É uma prática comum descartar as primeiras iterações (*warm up*), pois os valores ser muito representativos de locais com baixa densidade.

Equações Para fins práticos, vamos trabalhar com um parâmetro desconhecido μ e considerar $\sigma^2 = 1$.

A função f proporcional deve ser proporcional à densidade do posterior.

$$\text{Posterior} \propto \frac{\text{Prior} \times \text{Likelihood}}{\text{Prob. Marginal}}$$

Probabilidades marginais É a probabilidade das observações $P(X)$. Elas são constantes no processo, servindo apenas para normalizar estimativas, então:

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

Priors

Nosso prior é normal, com média 0 e desvio-padrão 1, $P(\mu) \sim N(0, 1)$.

Likelihood Se as observações são independentes, precisamos apenas multiplicar a probabilidade cada uma delas.

Assumimos que a distribuição das medidas é normal, com média μ e desvio σ^2 . Para o estado s_i , a probabilidade das observações X considerando o μ_i é:

$$\begin{aligned} P(X|\mu_i) &= \\ \prod_{j=1}^n P(x_j|N(\mu_i, 1)) &= \\ \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j-\mu_i)^2}{2}} \end{aligned}$$

Função proporcional à densidade do posterior Usaremos o log likelihood pelas vantagens descritas antes: produtório se torna um somatório e passamos o contradomínio do intervalo $[0; 1]$ para $[-\infty, 0)$ (ou $(0, +\infty]$ multiplicando por -1).

$$\log(\text{Posterior}) \propto \log(\text{Prior} \times \text{Likelihood})$$

$$\begin{aligned} f : L(s_i) &= \log(P(X|\mu_i, 1) \times N(0, 1)) \\ \log\left(\prod_{j=1}^n P(x_j|N(\mu_i, 1)) \times N(0, 1)\right) &= \\ \log\left(\prod_{j=1}^n P(x_j|N(\mu_i, 1))\right) + \log(N(0, 1)) &= \end{aligned}$$

O segundo termo é uma distribuição normal com média e variância conhecidas. Precisaremos apenas usar valores transformados por logaritmo.
O primeiro termo é³² :

$$\begin{aligned} \sum_{j=1}^n \log(P(x_j|N(\mu_i, 1))) &= \\ = -\frac{n}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{j=1}^n (x_j - \mu_i)^2 \end{aligned}$$

Finalmente, podemos calcular para cada estado um valor para os parâmetros μ_i, σ_i , aceitá-los ou rejeitá-los.

³²Dedução em <https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood>

Implementação Implementaremos MCMC como prova de conceito para ilustrar o mecanismo de convergência. Para uma aplicação real com resultados robustos, alguns esforços a mais seriam necessários. Por exemplo, os passos do nosso programa serão sempre idênticos, a normalização dos valores foi feita artesanalmente para a amostra e usamos apenas uma cadeia para estimar o posterior.

Stan usa uma versão altamente sofisticada de MCMC, em que a evolução do sistema é guiado por uma função (Hamiltoniana) da energia total. É possível observar um gradiente e, assim como em fenômenos físicos, estados com menores níveis de energia têm maior probabilidade de serem ocupados (e.g. distribuição de Boltzmann em mecânica estatística).

Usando o algoritmo descrito acima para a diferença entre médias, geramos as amostras a e b , $n = 400$, de populações com médias $\mu_a = 0, \mu_b = 0.6$, e distribuição normal.

```
>set.seed(2600)

>n_obs <- 400
>a <- rnorm(n=n_obs, sd =1, mean=0)
>b <- rnorm(n=n_obs, sd=1, mean=0.6)
```

Vamos definir nossa função de verossimilhança (usando transformação de $-\log$):

```
>likel <- function(n,x, mu, sigma){
  l_val <- (-n/2)*log(2*pi*sigma^2) - (1/2*sigma^2)*sum((x - mu)^2)
  return(-l_val) # multiplica(-1)
}
```

Definindo a função para fornecer $\log(N(0, 1))$. Obteremos as probabilidades e o logaritmo delas para um n grande, representativo. Esse número será normalizado pelo tamanho de nossa amostra para permitir passos numa escala razoável nos cálculos da cadeia.

```
>log_norm <- function(n, mu, sigma){
  require(magrittr) # para o operador %>%
  # Truque para obter distribuição ~ uniforme em [-Inf, +Inf]
  unif_dist <- 1/runif(n = n, min = -1, max = 1)
  l_val <- dnorm(x=unif_dist, mean = 0, sd = 1, log=T)
  l_val <- car::recode(l_val, "-Inf:-1000=-1000") %>% sum # recod. valores extremos
  return(-l_val)
}
```

E um loop para rodar a simulação MCMC:

```
# MCMC chain
>mc_chain <- function(obs, iter=4000, n_obs=length(obs)){
```

```

# seeds e objetos
sample <- matrix(nrow = iter, ncol = 2)
s1_mu <- rnorm(n=1,mean=0) # media inicial
s_sigma <- 1 # variância = 1
s1_lik <- 2000
for (i in 1:iter){
  # Salva estado
  s0_mu <- s1_mu
  s0_lik <- s1_lik

  # Realiza um passo (random walk)
  s1_mu <- s1_mu + rnorm(n=1,mean = 0, sd=0.5)
  s1_lik <- likel(n=n_obs,x=obs,mu=s1_mu,sigma=s_sigma) +
    # log do prior se baseian numa densidade de n=10000 e é normalizado por 1000
    log_norm(n=10000, mu=0, sigma=1)/1000

  # Rejeita diferenças maiores que 5, assumindo o valor no estado anterior
  if(s1_lik - s0_lik > 5)
    s1_mu <- s0_mu
  sample[i,] <- c(s1_mu,s_sigma) # Salva
}
return(sample[1001:iter,]) # Descarta as primeiras 1000 amostras (warm-up)
}

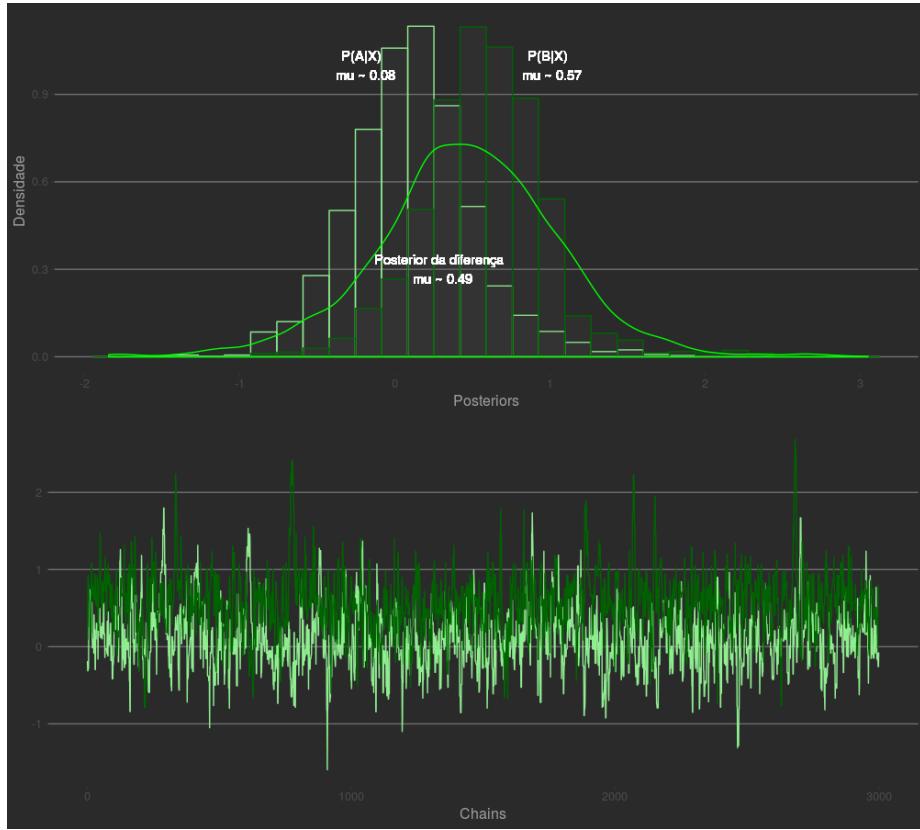
```

Podemos então obter nossas distribuições posteriores para μ_A, μ_B e para a diferença. Também vamos visualizar a evolução dos estados ao longo do tempo.

```

>posterior_a <- mc_chain(obs = a,iter = 4000)
>posterior_b <- mc_chain(obs = b,iter = 4000)
>posteriors_data <- data.frame(post_a=posterior_a, post_b=posterior_b)
>posts_plot <- ggplot(data = posteriors_data, aes(x=posterior_a)) +
  geom_histogram(aes(y=..density..),color = "light green", alpha=0.1) +
  geom_histogram(aes(x=posterior_b, y=..density..), alpha=0.1, color="dark green") +
  geom_density(aes(x=(posterior_b - posterior_a)), color="green") +
  xlab("Posteriors") + ylab("Densidade") +
  geom_text(label="P(A|X) \n mu ~ 0.08",color="white",x=-0.2,y=1)+ 
  geom_text(label="P(B|X) \n mu ~ 0.57",color="white",x=1,y=1)+ 
  geom_text(label="Posterior da diferença \n mu ~ 0.49",color="white",x=0.3,y=0.3)+ 
  theme_hc(style = "darkunica")
>traces_plot <- ggplot(data=posteriors_data,
  aes(y=posterior_a,x=1:nrow(posteriors_data)))+
  geom_line(color="light green")+xlab("Chains")+ylab("")+
  geom_line(aes(y=posterior_b,x=1:nrow(posteriors_data)),
  color="dark green")+
  theme_hc(style="darkunica")
> multiplot(posts_plot,traces_plot,cols = 1)

```



O painel superior da visualização destaca distribuições posteriores de A (verde claro) e B (verde escuro), assim como da diferença. Elas refletem razoavelmente bem as distribuições de origem ($N(0, 1)$, $N(0.6, 1)$) inferidas a partir dos dados. No painel inferior, temos as cadeias para A (média menor, com sinal oscilando num nível menor) e B (média maior, com sinal oscilando acima). Ainda que seja um modelo ilustrativo, o resultado parece bom, com distribuições representativas.

Exercícios

1. Usando Stan, implemente regressão linear para dados à sua escolha. A *likelihood function* para observações pode ser uma gaussiana cuja média é pela equação de regressão. O guia de usuários deve ajudar. https://mc-stan.org/docs/2_18/stan-users-guide/linear-regression.html
 - Implemente regressão linear com mais de um preditor.
 - Compare a média dos posteriores para os coeficientes β com a estimativa pontual clássica usando `glm`.
2. Com a biblioteca **BEST** conduza a comparação de médias do exemplo final, invocando a função `BESTmcmc` e especifique o argumento `numSavedSteps = 3000`.
 - Extraia as distribuições posteriores, `mu1` e `mu2`, do objeto resultante.
 - Obtenha a diferença entre distribuições `mu1 - mu2` e compare visualmente (densidade ou histograma) com o posterior que geramos através do MCMC artesanal.
3. Aperfeiçoe a simulação MCMC modificando a função `mc_chain`.
 - Obtenha a amostra final para o posterior sorteando valores gerados por 4 cadeias independentes.
 - Faça com que o tamanho dos passos diminua linearmente com o número de simulações decorridas.