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 consideravel.

Por princípio, usamos caminhos diferentes.

Frequencistas e Bayesianos

Abordagens frequencistas 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 fenomeno (*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 saudavel 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 alternativas[1].

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.

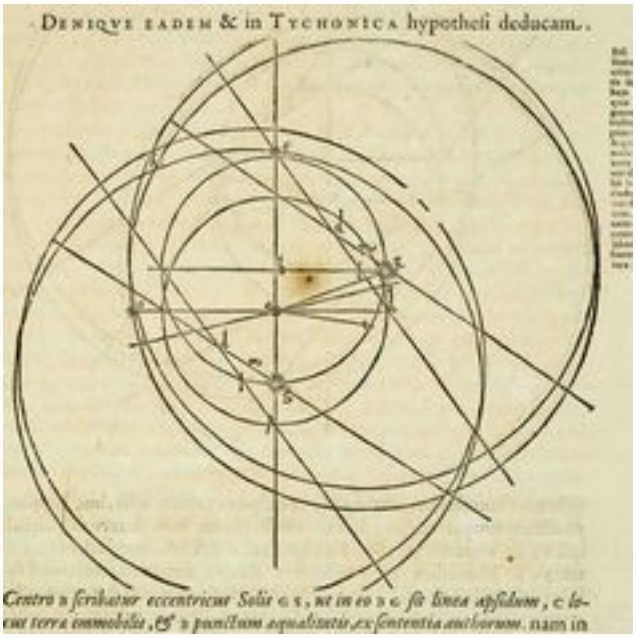
Paul Feyerabend (1924 - 1994)

[1] Existe um programa de pesquisa mais abrangente em filosofia sobre epistemologia Bayesiana, mas este não é nosso foco.

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' [2].

Além de menos acurado, era mais complexo em alguns aspectos, incluindo mais epiciclos: órbitas auxiliares usadas como artifício para cálculos. A Revolução Copernicana somente consolidou a mudança de paradigma com contribuições subsequentes de 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.

[2]Stanley E. Babb, "Accuracy of Planetary Theories, Particularly for Mars", Isis, Sep. 1977, pp. 426

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 desonestidade 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 desejável 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?"*.

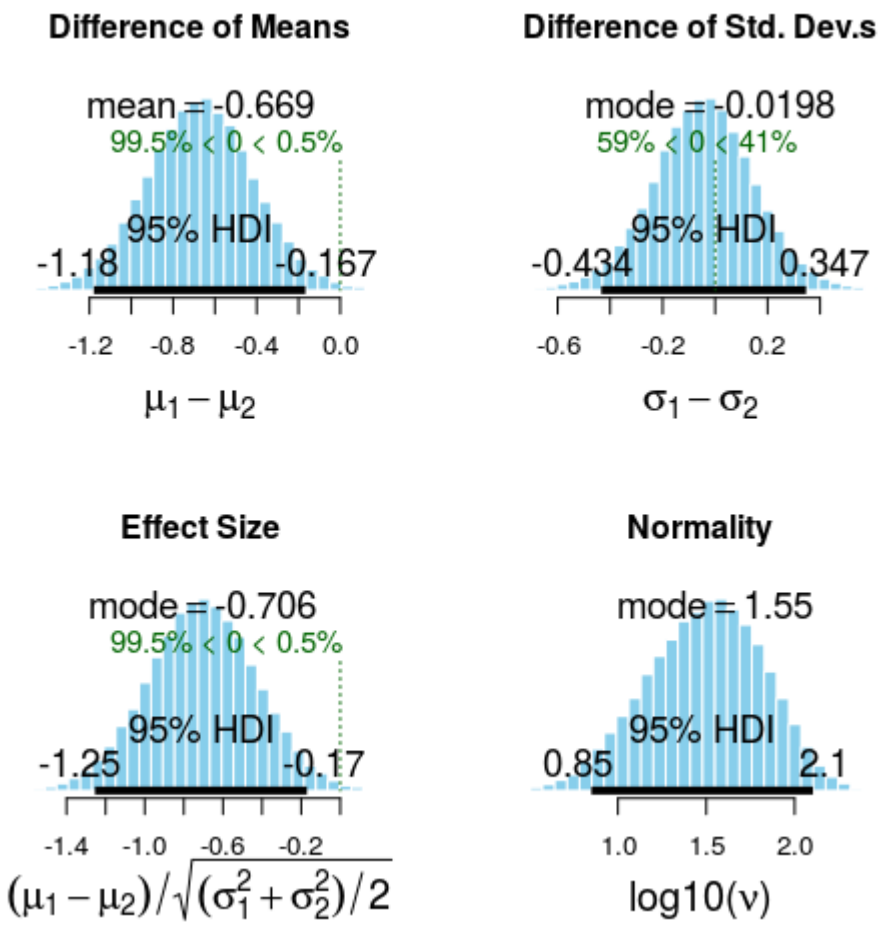
A distribuição probabilística obtida representa nossas crenças na plausibilidade de cada valor.

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.

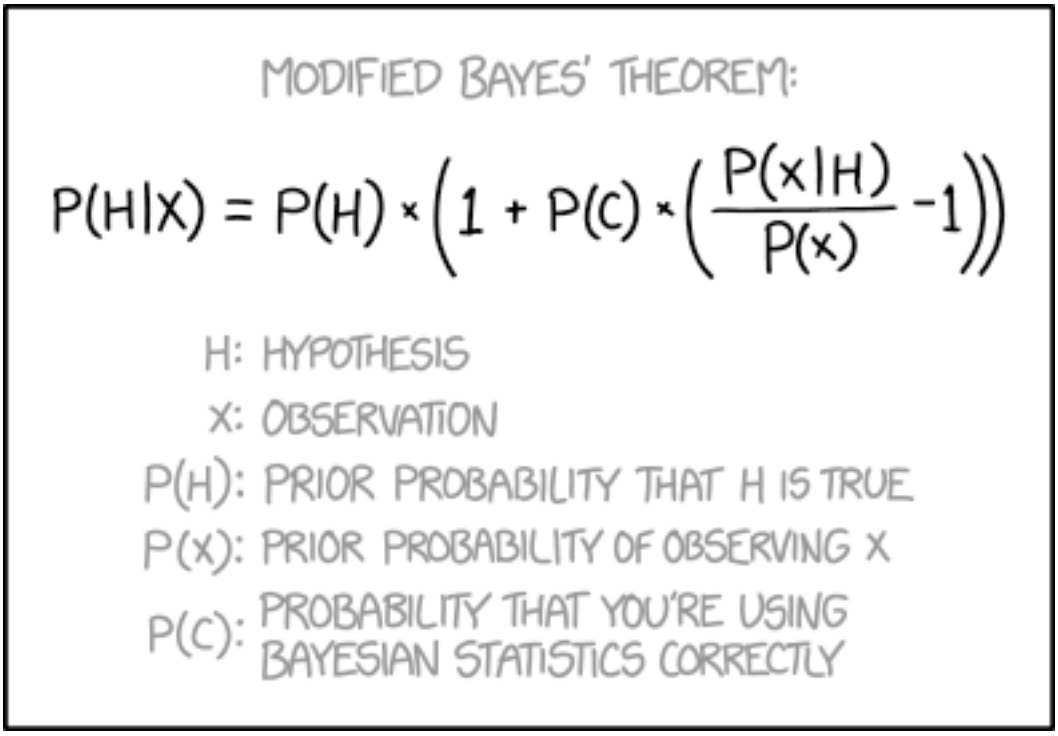


Figure 1: https://xkcd.com/2059/ Teorema de Bayes modificado, incluindo a probabilidade de você estar usando estatística bayesiana correntamente

O Teorema do Bayes

$$P(B \mid A) = \frac{(A \mid B)P(B)}{P(A)}, P(A) \neq 0$$

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 $p(0 \leqslant p \leqslant 1)$ 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 subesequent 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)*:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, P(A) \neq 0$$

Considerando dois eventos, **A** e **B**, a probabilidade de B acontecer sabendo que A aconteceu ($P(B \mid A)$) 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, $P(A \cap B) = P(A \mid B)P(B)$, então:

$$P(B \mid A) = \frac{(A \mid B)P(B)}{P(A)}, P(A) \neq 0$$

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).

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}, P(X) \neq 0$$

Posterior

Chamamos o primeiro termo, a estimativa do parâmetro após a calibração pelas observações $P(\theta \mid X)$, 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, $P(X \mid \theta)$, 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.

$$P(X \mid \theta) \sim f(X, \theta) : y_i = \frac{1}{1 + e^{-(\theta * x_i + c)}}$$

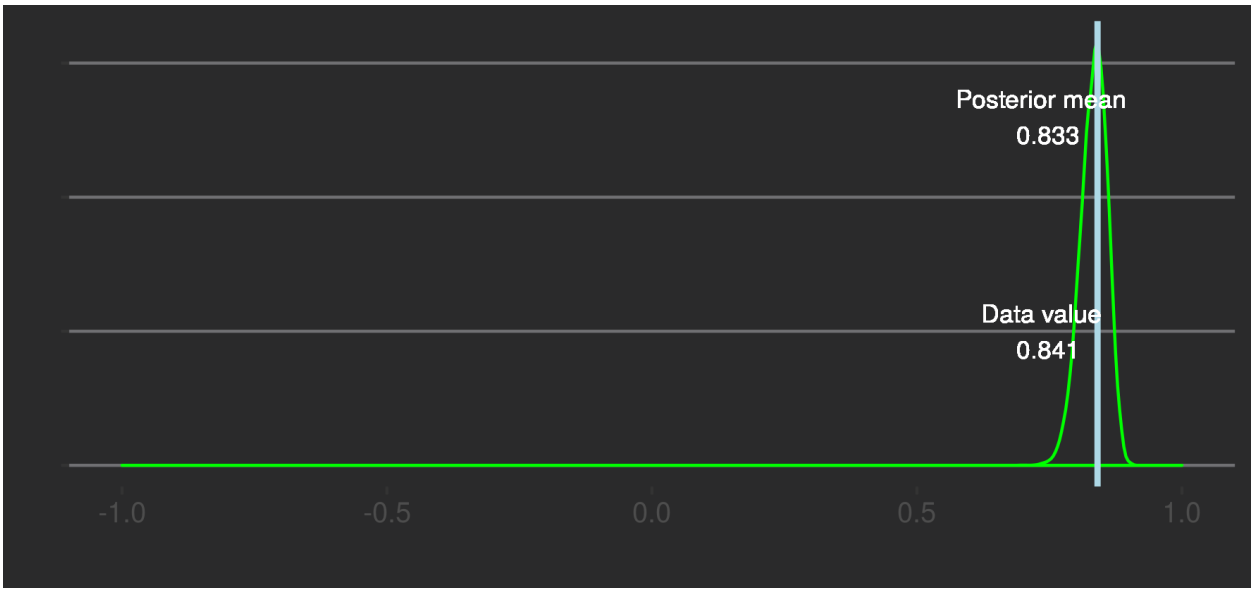


Figure 2: O prior verde supõe maiores probabilidades para valores baixos. O prior amarelo é pouco, informativo atribuindo probabilidades semelhantes em todo o intervalo

Outras funções poderia ser escolhidas (e.g. Heaviside step do capítulo anterior). Isto depende do 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_density(aes(x=low),color="#009E73")+
ylab("Densidade")+xlab("Priors: informativo(verde) ou incerto (amarelo)")+
theme_hc(style="darkunica")+theme(axis.text.y=element_blank())
```

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 Recrutador[3]

Um recrutador técnico, ao descobrir que os caminhos dos hackers Unix lhe eram estrahos, 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?"

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".

[3] http://www.catb.org/~esr/writings/unix-koans/recruiter.html

```
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 - Verossimilhanca
  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: amostra 2, N: tamanho amostral).

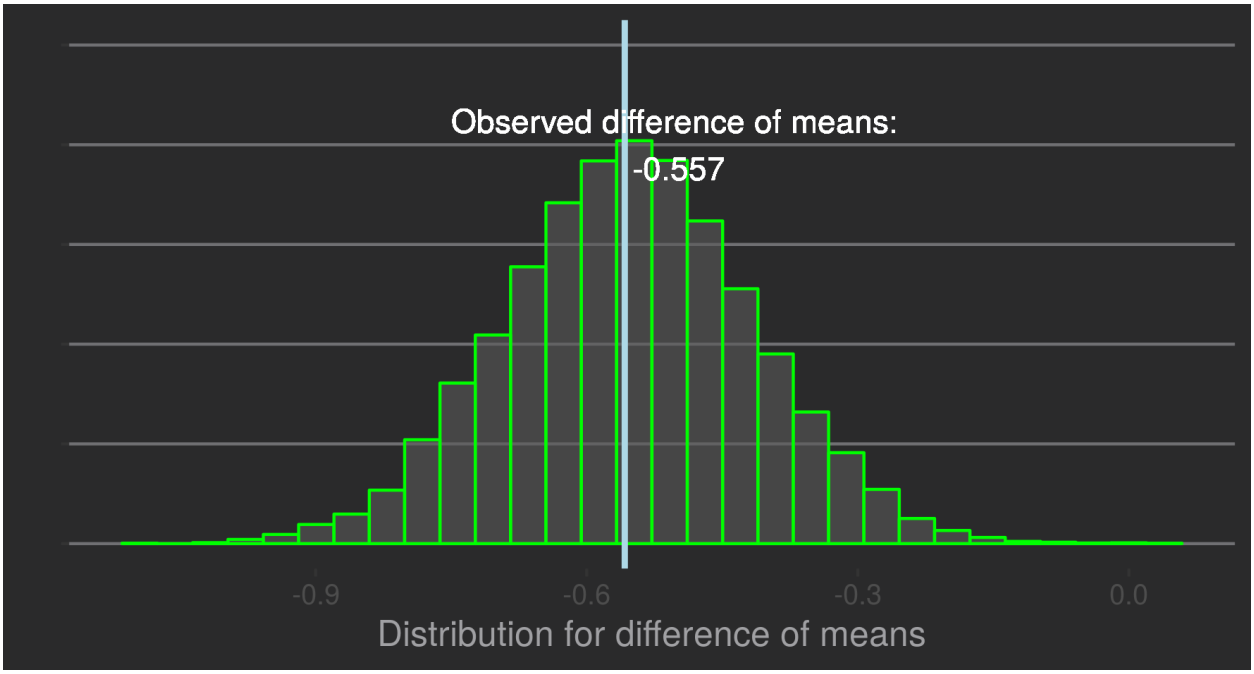
```
> a <- rnorm(n = 100, sd = 1, mean = 0)
> b <- rnorm(n = 100, sd = 1, mean = 0.6)
> sample_data <- list(y_1=a,y_2=b,N=length(a))
> fit <- rstan::stan(file="aux/bayes-t.stan",
        data=sample_data,
        iter=3000, warmup=100, chains = 6)
 SAMPLING FOR MODEL 'bayes-t' NOW (CHAIN 1).
 (...)
```

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 idea 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

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

Então,

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

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;
```

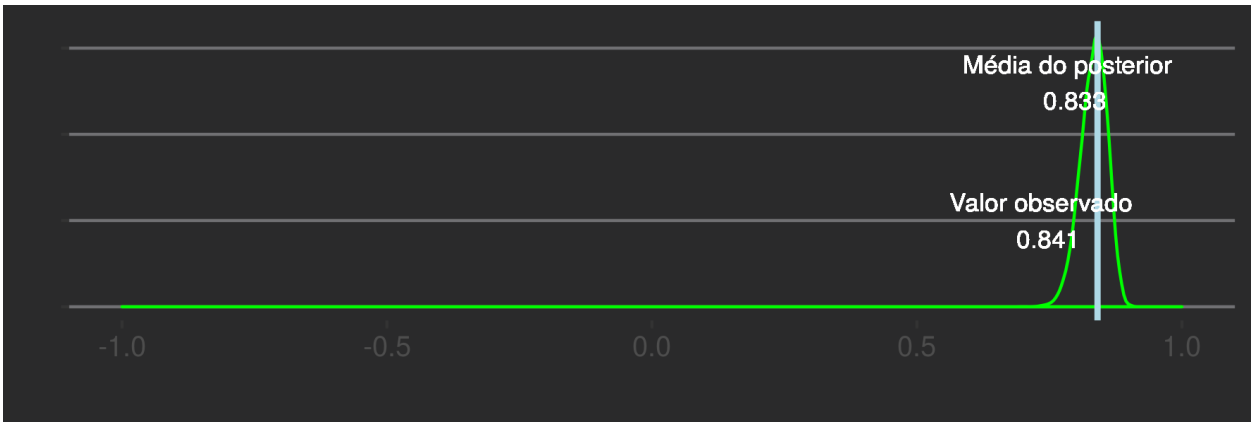
```
  x_rand = multi_normal_rng(mu, cov);
}
```

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())
```



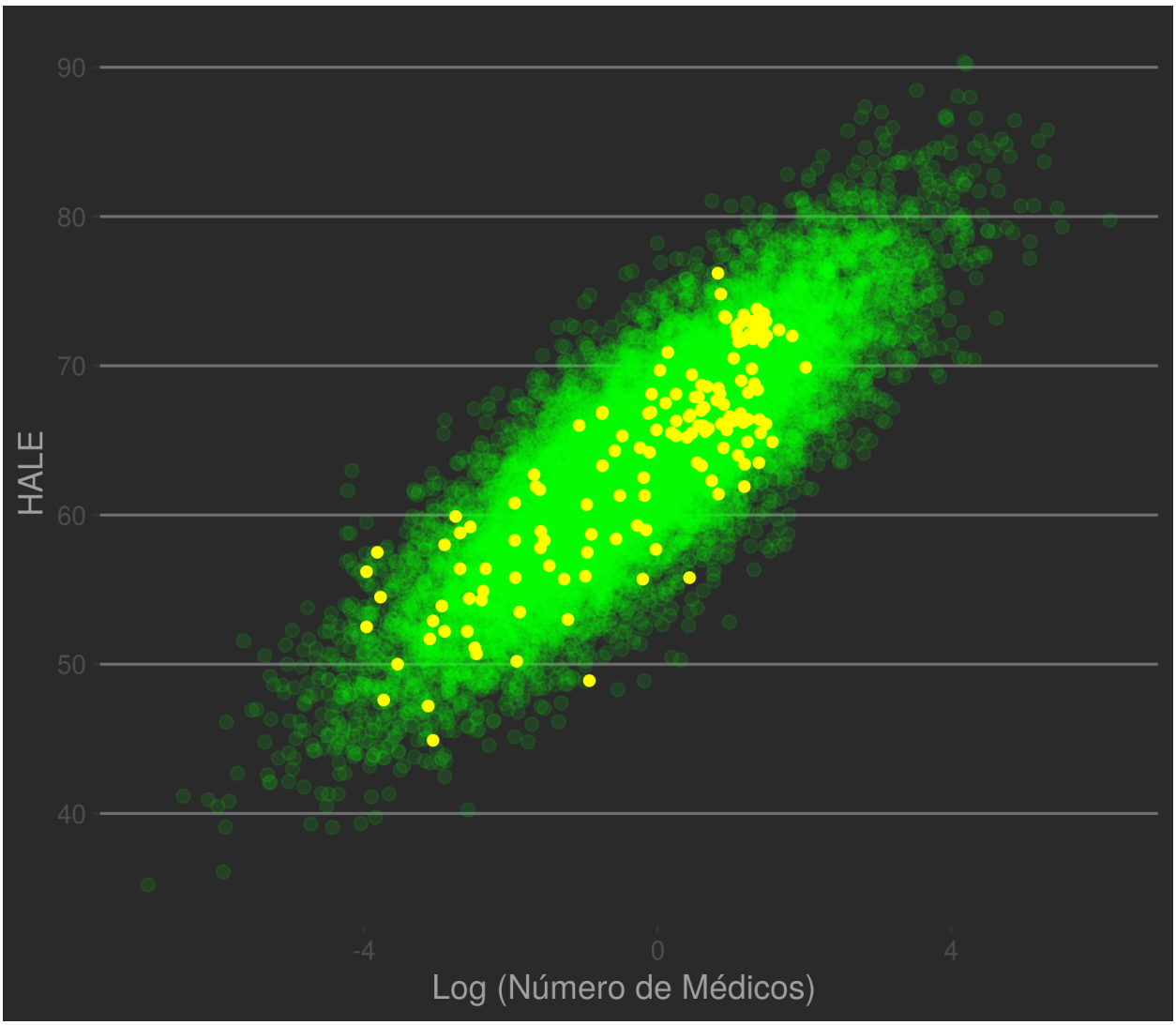
Notamos que as estimativa posterior para ρ foram razoavelmente distruídas ao redor do valor empiricamente calculado na amostra. Podemos ainda observar na distribuição intervalos com alta densidade de probabilidade (HDI, High density intervals) ou ainda outros fins.

```
> 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.

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}, P(X) \neq 0$$

Usando Stan, informamos priors, a função de verossimilhança, observações e todo o trabalho sujo é realizado sem mais esforços.
A estimativa de $P(\theta \mid X)$ 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

[4]

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

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

Por exemplo, o coeficiente angular (β_1) é

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

Gradient Descent

No capítulo 4, mostramos outra maneira de estimar parâmetros, analisando uma função de perda. Usando derivadas parciais, calculamos o gradiente, análogo à *inclinação* 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*).

$$(g \circ f)' = (g' \circ f)f'$$

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, $P(\theta\ midX)$, usamos um algoritmo que permite obter amostras representativas de $P(\theta\ midX)$. Para isso, a condição é de que exista uma função $f(x)$ proporcional à densidade de $P(\theta\ midX)$ e que possamos calculá-la.

[4] 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,**

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*). Reavaliamos 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, $f(x) \sim \text{dens}(P(\theta\ midX))$, 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:

$$P(X|\mu_i) =$$

$$\prod_{j=1}^{n} P(x_j|N(\mu_i, 1)) =$$

$$\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j-\mu_i)}{2}}$$

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).

$$\log(\text{Posterior}) \propto log(\text{Prior} \times \text{Likelihood})$$

$$f : L(s_i) = \log(P(X|\mu_i, 1) \times N(0, 1))$$

$$\log(\prod_{j=1}^{n} P(x_j|N(\mu_i,1)) \times N(0,1)) =$$

$$\log(\prod_{j=1}^{n} P(x_j|N(\mu_i,1))) + \log(N(0,1)) =$$

O segundo termo é uma distribuição normal com média e variância conhecidas. Precisaremos apenas usar valores transformados por logaritmo.
O primeiro termo é[5] :

$$\sum_{j=1}^{n} \log(P(x_j|N(\mu_i,1))) =$$

$$= -\frac{n}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}\sum_{j=1}^{n}(x_j-\mu_i)^2$$

Finalmente, podemos calcular para cada estado um valor para os parâmetros μ_i, σ_i, aceitá-los ou rejeitá-los.

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 artesenalmente 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 e esse número será normalizado pelo tamanho de nossa amostra para permitir passos numa escala razoável.

```
>log_norm <- function(n,mu,sigma){
  require(magrittr) # para o operador %>%
  # Truque para obter distribuicao ~ uniforme em [-Inf,+Inf]
```

[5]Dedução em https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood

```
  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:

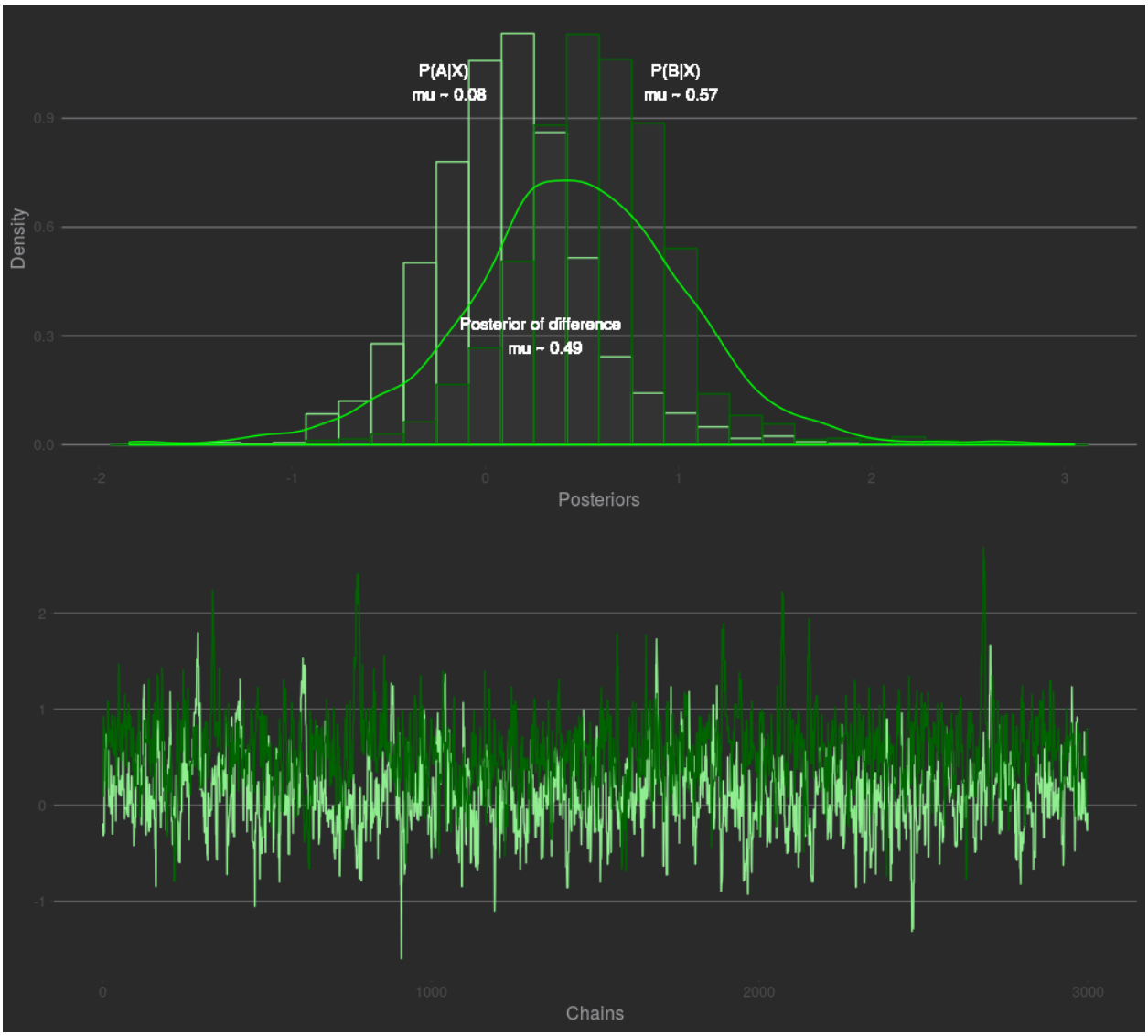
```
# MCMC chain
>mc_chain <- function(obs,iter=4000,n_obs=length(obs)){
  # seeds e objetos
  sample <- matrix(nrow = iter, ncol = 2)
  s1_mu <- rnorm(n=1,mean=0) # media inicial
  s_sigma <- 1 # variancia = 1
  s1_lik <- 2000
  for (i in 1:iter){
    # Salva estado
    s0_mu <- s1_mu
    s0_lik <- 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 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,1]) # Descarta as primeiras 1000 amostras (warm-up)
}
```

Podemos então obter nossas distribuições posteriores:

```
>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)
```



A visualização destaca distribuições posteriores de A e B, assim como da diferença. simulações Markov Chain Monte Carlo. Poderíamos extrair regiões de alta densidade para nossa estimativa usando `quantile((posterior_b - posterior_a),probs = c(0.025,0.5,0.975))`.