Laboratório 2: Probabilidade

Mãos Quentes

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

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

Salvando seu Código

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

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

Preparações

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

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

load("kobe.RData")

head(kobe)
```

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

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

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

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

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

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

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

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

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

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

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

barplot(table(kobe_streak))
```

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

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

Comparado a quê?

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

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

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

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

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

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

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

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

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

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

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

Simulações no R

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

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

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

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

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

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

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

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

```
sim_fair_coin

table(sim_fair_coin)
```

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

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

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

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

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

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

```
?sample
```

Simulando o Arremessador Independente

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

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

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

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

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

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

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

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

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

```
kobe$basket

sim_basket
```

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

Sua vez

Comparando Kobe Bryant ao Arremessador Independente

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

1. Descreva a distribuição das sequências de arremessos. Qual é o comprimento de sequência típico para o arremessador independente simulado com um percentual de arremesso de 45%? Quão longa é a sequência mais longa de cestas em 133 arremessos?

2. Se você rodasse a simulação do arremessador independente uma segunda vez, como você acha que seria a distribuição de sequências em relação à distribuição da questão acima? Exatamente a mesma? Mais ou menos parecida? Completamente diferente? Explique seu raciocínio.

3. Como a distribuição dos comprimentos de sequência de Kobe Bryant, analisada na página 2, se comparam à distribuição de comprimentos de sequência do arremessador simulado? Utilizando essa comparação, você tem evidência de que o modelo das mãos quentes se ajusta aos padrões de arremessos de Kobe? Explique.

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