Inclusão, exclusão e a tentativa de forjar o acaso

1) Um jogo sobre acaso para brincar em sala de aula.

A brincadeira descrita aqui é adaptada de uma ideia de Pál Révesz, matemático húngaro que trabalhou muitos anos na Áustria.

Uma turma de alunos é dividida dividia em dois grupos. O grupo 1 lança uma moeda justa *n* vezes e anota os resultados como 0 (cara) e 1 (coroa). O grupo 2 inventa resultados (falsos) do que seriam os lançamentos da moeda. O professor receberá os resultados sem saber qual é qual; o objetivo dele é descobrir quais são os resultados realmente aleatórios do grupo 1 e quais são os forjados. Já os alunos querem que o professor se engane.

Em geral o professor vence, com a estratégia descrita logo abaixo. A razão para isso é que a intuição dos alunos em geral não bate com a realidade do que é uma sequência aleatória.

O professor Révesz observou que, na prática, há uma estratégia simples para separar o real do forjado. Defina:

$$R_n = \text{ comprimento da maior sequência consecutiva de resultados iguais em n lançamentos.}$$

Na intuição dos alunos, valores grandes de R_n parecem "suspeitos" de não-aleatoriedade. Na verdade, o contrário ocorre: em geral a sequência verdadeira tem um valor de R_n *maior* do que o da forjada. Ou seja, o valor forjado de R_n tende a ser o menor.

De fato, como veremos a seguir, o R_n de uma sequência aleatória tende a estar bem próximo de $\log_2 n$. Ou seja: em 1024 lançamentos, deve haver alguma sucessão de cerca de 10 valores iguais consecutivos. Você pode testar isso e se convencer que é verdade usando o gerador de números aleatórios do site: http://www.random.org/bytes (use a opção "binary").

No sítio web

https://faculty.math.illinois.edu/~hildebr/fakerandomness/

você mesmo pode se arriscar a gerar uma sequência de 200 lançamentos para fazer o computador pensar que se tratam de lançamentos aleatórios. Se você resolver forjar, o mais provável é que o computador descubra, usando uma série de testes estatísticos.

2) O teorema por trás do jogo

Teorema

Considere n lançamentos de uma moeda justa. Chame de R_n o comprimento da maior sequência de resultados iguais consecutivos nestes lançamentos. Então:

$$\mathbf{Pr}\left(0.99\ \log_2 n \le R_n \le 1.01 \log_2 n\right) \to 1$$

quando $n \to +\infty$.

No restante do texto, veremos um esboço da prova deste teorema. Primeiro temos de formulá-lo, definindo um espaço amostral e uma probabilidade adequados.

$$\Omega := \{ \textit{sequências de n zeros e uns} \} = \{0,1\}^n$$
$$P(\omega) = \frac{1}{2^n} \textit{para todo } \omega = \left(\omega_1, \omega_2, \ldots, \omega_n\right) \in \Omega.$$

Um subconjunto de Ω é chamado de evento e sua probabilidade é a sua cardinalidade dividida por 2^n.

É um exercício verificar que, para provar o teorema, basta mostrar que:

$$(?)\ P\{R_n \ge 1.01 \log n + 1\} \to_? 0 \ e \ P\{R_n \ge 0.99 \log n\} \to_? 1$$

A seguir, faremos todas as contar tratando $1.01 \log n + 1$ e $0.99 \log n$ como inteiros.

De modo geral, o evento $\{R_n \ge k\}$ é o conjunto de todos $\omega = \left(\omega_1, \omega_2, \ldots, \omega_n\right) \in \Omega$ que tem k ou mais coordenadas consecutivas iguais em algum lugar. Pode-se verificar que a partir daí que, definidos

$$E_i := \left\{\omega \in \Omega \;\middle|\; \omega_i = \omega_{i+1} = \ldots = \omega_{i+k-1}\right\} \textit{para } i = 1, \ldots, n - k + 1,$$

vale que $\{R_n \ge k\} = \bigcup_{i=1}^{n-k+1} E_i$.

Veja ainda que $\# E_i = 2^{n-k+1}$. Com efeito, escolher $\omega \in \Omega$ é equivalente a escolher cada coordenada ω_j com $j \le i$ ou $i + k \le j \le n$ arbitrariamente (as outras coordenadas devem valer 1). Portanto, há um total de $n - k + 1$ escolhas de 0 e 1 a serem feitas.

Portanto, $P(E_i) = 2^{1-k}$. Temos

$$P\{R_n \ge k\} = P(\bigcup_{i=1}^{n-k+1} E_i) \le \sum_i P(E_i) = (n - k + 1) 2^{1-k},$$

já que a probabilidade de uma união é no máximo a soma das probabilidades. Com $k = 1.01\log n$, vê-se que a expressão acima vai a 0, logo $P\{R_n \geq 1.01\log n + 1\} \rightarrow_? 0$.

3) A parte mais difícil da prova e o método do segundo momento

Como provar que $P\{R_n \geq 0.99\log n\} \rightarrow_? 1$?

Esta parte da prova é mais difícil. Em princípio, como estamos lidando com uma união de eventos, deveria ser possível usar o Princípio da Inclusão-Exclusão, mas isso parece bastante complicado.

A alternativa que empregaremos será usar o "método do segundo momento", que é aplicável a qualquer espaço de probabilidade. (Ele é discutido por exemplo no livro de Alon e Spencer, "The Probabilistic Method".)

(2M) Para quaisquer subconjuntos $E_1, E_2, \ldots, E_m \subset \Omega$

$$P(E_1 \cup E_2 \cup \ldots \cup E_m) \geq \frac{s_1^2}{2s_2 + s_1}.$$

onde $s_1 = \sum P(E_i)$ e $s_2 = \sum P(E_i \cap E_j)$

Note que tanto s_1 quanto s_2 aparecem na fórmula de Inclusão-Exclusão; a vantagem da desigualdade acima é que todos os outros termos "chatos" daquela fórmula não estão presentes.

Para aplicar a fórmula, precisamos determinar (ou pelo menos estimar) s_1 e s_2. Já vimos que, no nosso caso, $s_1 = (n - k + 1)2^{1-k}$. Para calcular os termos $P(E_i \cap E_j)$ de s_2, temos dois casos a considerar.

Quando $j \geq i + k$, é fácil ver que a escolha de um $\omega \in E_i \cap E_j$ pode ser feita selecionando as coordenadas ω_k de maneira completamente arbitrária, exceto pelos índices $k = i + 1, \ldots, i + k - 1, j + 1, \ldots, j + k - 1$. Como há 2k-2 destes índices,

$$\#(E_i \cap E_j) = 2^{n-2k+2} \; e \; P(E_i \cap E_j) = 2^{2-2k}.$$

Note ainda que o número de pares $i < j$ do tipo acima é no máximo $(n - k + 1)^2/2$.

Quando $j \leq i + k - 1$, podemos cotar $P\left(E_i \cap E_j\right) \leq P(E_i) = 2^{1-k}$. A cota não é muito boa, mas isso importa pouco porque há poucos pares *i,j* deste tipo: no máximo $(n - k + 1)$ escolhas para i e k escolhas para j dado *i*.

Da discussão acima, deduzimos que:

$$s_2 \leq \frac{(n-k+1)^2}{2} 2^{2-2k} + (n-k+1)k\,2^{1-k} = \frac{s_1^2 + 2ks_1}{2}.$$

Algumas contas nos revelam que:

$$P(R_n \geq k) = P(E_1 \cup E_2 \cup \dots \cup E_{n-k+1}) \geq \frac{1}{1 + (2k+1)/s_1}.$$

No caso que nos interessa, $k = 0.99 \log n$ e

$$s_1 = (n-k+1)2^{1-k} = 2n^{0.01}\left(1 - O\left(\frac{\log n}{n}\right)\right)$$

Logo, $\frac{2k+1}{s_1} \to 0$ e deduzimos que $P(R_n \geq k) \to 1$.

4) Qual a moral da história?

Nesta aula, vimos um jogo para ser feito em aula que nos mostra o quanto a nossa intuição é falha quando lidamos com probabilidades. Também apresentamos um teorema que fala algo relacionado ao jogo e vimos que a prova dele pode ser feita através da aproximação para probabilidades chamada de "método do segundo momento".

Mais informações sobre sequências de repetições em lançamentos de moedas são dadas em:

Mark F. Schilling, "The Surprising Predictability of Long Runs." Mathematics Magazine, Vol. 85, No. 2 (April 2012), pp. 141-149

Um ponto fundamental sobre nosso problema é que achar probabilidades exatas é difícil, mas desigualdades são mais fáceis. O fato de que Análise, Probabilidade e outras áreas de Matemática avançada fazem uso amplo de estimativas por desigualdade não tem paralelo no Ensino Médio. Uma pergunta para reflexão: há espaço para falar mais de desigualdades e estimativas na escola? Isso seria boa ideia?