Esenciais USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SIMULACIÓN NO CÁLCULO DE PROBABILIDADES

unha aplicación a ránkings de páxinas web

Alberto Rodríguez Casal, Alejandro Saavedra Nieves e Paula Saavedra Nieves (Departamento de Estatística, Análise Matemática e Optimización)

O termo probabilidade adoita usarse coloquialmente para referirnos ao grao de certeza no resultado dun experimento antes da súa realización. Para cuantificar a probabilidade podemos empregar a experiencia empírica previa a ese experimento. Así, se xa se ten observado antes, poderiamos calcular as frecuencias relativas asociadas aos resultados posibles (número de veces que obtemos un resultado dividido entre o número de repeticións) como fonte de información sobre a certeza do que ocorrerá nunha realización futura do experimento.

Formalmente, a lei de estabilidade de frecuencias asegura que se un experimento se repite moitas veces e en idénticas condicións, as frecuencias relativas dos seus posibles resultados tenden a estabilizarse arredor das probabilidades teóricas. De feito, a noción frecuentista de probabilidade define a probabilidade dun suceso como o límite das súas frecuencias relativas cando realizamos o experimento moitas veces. Neste contexto, a simulación estatística é unha ferramenta moi útil para dar evidencia numérica a determinados resultados de probabilidade. Véxase o Exemplo 1 en que se simula o lanzamento dun dado.

Exemplo 1

Todos sabemos que a probabilidade de obter un 5 ao lanzar un dado é 1/6. Podemos simular esta situación tan simple, por exemplo, co software estatístico R (R Core Team, 2022). É dicir, podemos repetir o experimento de lanzar un dado centos de veces. Se dividimos o número de veces que obtemos un 5 entre el número total de lanzamentos observaremos que, a medida que o número de lanzamentos aumenta, aproximámonos cada vez máis á solución teórica.

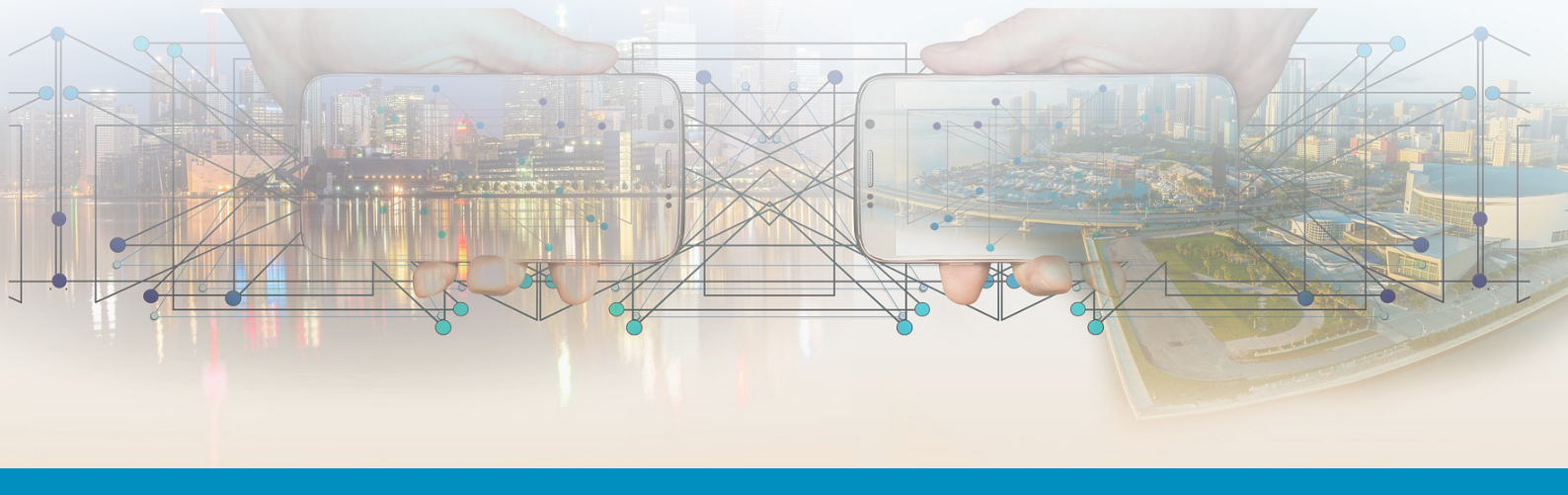
Execución con 10 lanzamentos

```
> resultados=sample(1:6,size=10,replace=TRUE)
> table(resultados)/10
    1     2     3     4     5     6
0.200 0.100 0.300 0.300 0.100 0.000
```

Execución con 500 lanzamentos

```
> resultados=sample(1:6,size=500,replace=TRUE)
> table(resultados)/500
    1     2     3     4     5     6
0.174 0.172 0.178 0.164 0.142 0.170
```

A función `sample` de R sortea os valores dentro dun vector con substitución e sen substitución empregando catro argumentos de entrada. O primeiro argumento contén o vector cos elementos que se van sortear. O segundo argumento etiquetado como `size` denota o número de elementos a xerar. Por defecto, a mostra xérase sen substitución. Se queremos que o proceso se realice con substitución, necesitaremos utilizar o terceiro argumento `replace = TRUE`. O cuarto argumento denotado por `prob` úsase para establecer pesos de probabilidade diferentes asociados aos elementos do vector que se van sortear. Se, coma neste exemplo, os posibles resultados do experimento son equiprobables, pódese omitir. Nos dous cadros superiores simulamos o lanzamento dun dado 10 e 500 veces. Os resultados (cor negra) amosan as frecuencias relativas dos resultados obtidos. Pode observarse que, con 500 lanzamentos, as frecuencias estabilízanse arredor do valor 1/6 que é a probabilidade teórica de cada un dos posibles resultados.



Baixo este enfoque, é posible experimentar con modelos máis complexos en que calcular a probabilidade teórica pode resultar relativamente complicado. Na actualizade, o posicionamento de sitios web atendendo á sua relevancia é un problema de interese para grandes compañías como Google ou Bing. No Exemplo 2, a simulación estatística emprégase para establecer empiricamente ránkings de páxinas web. Formalmente, se X_n denota a web visitada na etapa n, a situación descrita pode verse como unha camiñada aleatoria que se pode modelar a través dunha cadea de Markov homoxénea. As cadeas de Markov son procesos estocásticos $\{X_n, n=0, 1, 2, \dots\}$ con tempo e espazo de estados S discretos caracterizados porque a probabilidade de que ocorra un evento depende só do evento inmediatamente anterior (ver capítulo 1 en Lawler, 2018, para máis detalles). Matematicamente, en termos de probabilidades condicionadas,

$$P(X_{n+1}= x_{n+1} \mid X_n= x_n,\dots, X_o= x_0)= P(X_{n+1}= x_{n+1} \mid X_n= x_n)$$

para calquera enteiro n positivo e para calquera colección de estados $x_{n+1}, x_n,\dots, x_0 \in S$. Se, ademais, a distribución condicional de X_{n+1} dado X_n é independente de n, dise que a cadea de Markov é homoxénea indicando que as probabilidades de transición son estacionarias. Se k denota o número de posibles estados diferentes, a **matriz P de probabilidades de transición nun paso** dunha cadea de Markov homoxénea vén dada por:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix},$$

onde o elemento na fila i e na columna l, $p_{il} = P(X_{n+1} = l \mid X_n = i)$, representa a probabilidade de transición do estado i na etapa n ao estado l na etapa $n+1$. A matriz P é unha matriz estocástica porque todos os seus elementos son non negativos e a suma dos elementos de cada fila vale 1.

Exemplo 2

O seguinte grafo orientado representa as ligazóns que teñen entre si 6 páxinas web. Se asumimos que ao navegar pola rede prememos ao azar nalgunha das ligazóns da páxina na que nos atopamos, a importancia dunha páxina virá dada pola probabilidade de visitar cada páxina por este paseo aleatorio. Deste xeito, poderiamos establecer un ránking das páxinas consideradas de acordo coa súa relevancia. Supoñamos que comezamos no nodo 1 e estamos navegando aleatoriamente por un total de 10000 páxinas. Cal sería a frecuencia relativa de cada nodo ao longo do percorrido? Os resultados serían moi diferentes se comezamos no nodo 4? Cal é nese caso o nodo máis visitado?

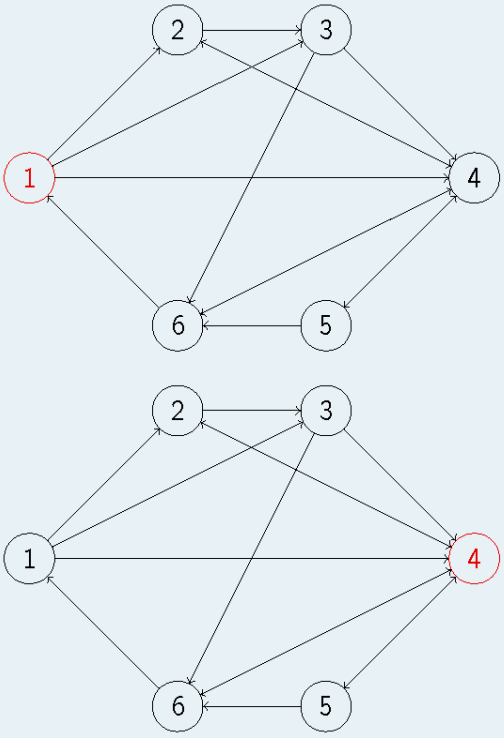


Figura 1: Ligazóns existentes entre 6 páxinas web

Execución con comezo no nodo 1

```
> P=matrix(0,nr=6,nc=6)
> P[1,2:4]=1/3
> P[2,3:4]=1/2
> P[c(3,5),c(4,6)]=1/2
> P[4,c(2,5,6)]=1/3
> P[6,1]=1

> consultas=10000
> paseo=numeric(consultas)
> paseo[1]=1
> for (i in 2:consultas){
+ paseo[i]=sample(1:6,size=1,prob=P[paseo[i-1],])
+}

> table(paseo)/consultas
     1      2      3      4      5      6
 0.195  0.147  0.134  0.248  0.082  0.195
```

Para comezar o camiño aleatorio no nodo 4, é preciso substituír a segunda liña do código por `paseo[1]=4` e repetir a execución.

```
     1      2      3      4      5      6
 0.191  0.148  0.137  0.251  0.083  0.191
```

A función `matrix` é empregada para definir a matriz P contendo as probabilidades de transición nun paso.

A estrutura de control básica `for` permite modelar as ligazóns existentes entre as 6 webs e determinar o percorrido do paseo aleatorio. A comparación das frecuencias relativas (cor negra) asociadas a cada páxina web para os dous puntos de inicio do percorrido amosan que o ránking web obtido, encabezado polo nodo 4, é independente do nodo de partida seleccionado mostrando a existencia dunha distribución estacionaria.

De xeito análogo, defínese a **matriz $P(n)$ de probabilidades de transición en n pasos**:

$$P(n) = \begin{pmatrix} p_{11}(n) & p_{12}(n) & \dots & p_{1k}(n) \\ p_{21}(n) & p_{22}(n) & \dots & p_{2k}(n) \\ \dots & \dots & & \dots \\ p_{k1}(n) & p_{k2}(n) & \dots & p_{kk}(n) \end{pmatrix},$$

onde o elemento na fila i e na columna l, $p_{il}(n) = P(X_n = l \mid X_0 = i)$, representa a probabilidade de transición do estado i ao estado l en n pasos. A partir do Teorema da probabilidade total, a ecuación de Chapman-Kolmogorov garante que para calquera r, n enteiros positivos con $0 \leq r \leq n$ e para calquera estados i e $l \in S$ cúmprese

$$p_{il}(n) = \sum_{k \in S} p_{ik}(r) p_{kl}(n-r).$$

Como consecuencia, a probabilidade de transición n en pasos, $p_{il}(n) = (P^n)_{il}$, está dada polo elemento (i, l) da potencia n-ésima da matriz P de probabilidades de transición nun paso. De novo, polo Teorema da probabilidade total, a probabilidade de acadar o nodo i $(i = 1, \dots, 6)$ na Figura 1 despois de n consultas vén dada pola suma dos elementos da columna i da matriz $P(n)$. Dado que n=10000 no Exemplo 2, é útil diagonalizar a matriz P para determinar a súa potencia n-ésima e, así, poder calcular de forma exacta as probabilidades de transición en n pasos para cada web. Ademais, a teoría do punto fixo garante a existencia da distribución estacionaria. Este resultado teórico xa foi anticipado no experimento realizado no Exemplo 2. A simulación estatística permite evitar a complexidade do modelo matemático exposto e, polo tanto, convértese nunha ferramenta chave no cálculo de probabilidades.

Referencias:

Lawler, G. F. (2018). Introduction to stochastic processes. Chapman and Hall/CRC.

R Core Team (2022). R: A language and environment for statistical computing. En: https://www.R-project.org/.