TD n°08

29 mars 2018

Exercice 1 - Approximation de Poisson

Consider the *Balls and Bins* model once again : suppose m balls are thrown into n bins independently and uniformly at random. Let X_i be the number of balls in the i-th bin where $1 \leq i \leq n$. (X_i are not independent, intuitively, since $X_1 + \cdots + X_n = m$.) We would like to approximate the *Balls and Bins* model with the Poisson distribution. Here, $Y_1, \ldots, Y_n$ are independent random variables each following Poisson distribution with parameter (i.e. expected value) $\mu = m/n$ (Y_i can be viewed as a simplified version of X_i).

1.1 Show that $Y = \sum_{i=1}^{n} Y_i$ follows Poisson distribution and determine its parameter.

1.2 Show that the distribution of $(Y_1, \ldots, Y_n)$ conditioned on $Y = m$ is the same as the distribution of $(X_1, \ldots X_n)$.

Note : One can, in fact, obtain a slightly more general result. If $(X_1, \ldots X_n)$ represents the load (charge) of n bins after throwing k balls at random, and Y_i are n independent random variables that follow Poisson distribution with parameter m/n, then the distribution of $(Y_1, \ldots, Y_n)$ conditioned on $Y = k$ is the same as the distribution of $(X_1, \ldots X_n)$ independent of value m.

1.3 Let f be a function of n variables that takes values in $\mathbb{R}_+ \cup \{0\}$. Show that

$$\mathbf{E}\left\{f(X_1, \ldots, X_n)\right\} \leq e\sqrt{m}\mathbf{E}\left\{f(Y_1, \ldots, Y_n)\right\} \ .$$

You may use the fact that $m! < e\sqrt{m}\left(\frac{m}{e}\right)^m$.

1.4 Call the *Poisson case* the set of events that occur when the number of balls in the bins are taken to be independent Poisson random variables with mean m/n. Call the *Balls and Bins case*, the set of events when m balls are thrown into n bins independently at random. Which function f would you apply to the above result to conclude :
Any event that takes place with probability p in the Poisson case takes place with probability at most $pe\sqrt{m}$ in the Balls and Bins case.

1.5 Re-establish the lower-bound on the maximal load in case $m = n$ using Poisson approximation. More precisely, show that if n balls are thrown independently into n bins, the maximal load will be at least $(\ln n)/(\ln \ln n)$ with probability at least $1 - 1/n$ for sufficiently large n.

Exercice 2 - Numéros de Sécurité Sociale

American Social Number is a number composed of 9 digits, and the last 4 digits serve as a secret code. Imagine that these digits were chosen uniformly at random and independent, without checking whether a particular number was already used.

2.1 We have n people s.t. there are at least 2 people with the same secret code. For which value of n this event is more likely to happen than not to happen (during the computation one is able to use approximations) ?

2.2 The same question with the following event : there are at least 2 people with the same social security number.

Exercice 3 - Uniformisation

Suppose we have a device that generates random bits that are guaranteed to be independent and have the same Bernouilli (p) distribution, except that we do not know the value of p. Design an algorithm that uses this source to produce a uniform bit and analyze the expected number of uses of the device that are needed to generated one uniform bit.

Exercice 4 - Bloom Filters

Consider once again a password checker. It prevents people from using common, easily cracked passwords by keeping a dictionary of unacceptable passwords. When a user tries to set up a password, the application would like to check if the requested password is a part of the unacceptable set.

A Bloom filter consists of an array of n bits, $A[0]$ to $A[n-1]$ initially all set to 0. A Bloom filter uses k independent random hash functions $h_1, \ldots, h_k$ with range $\{0, \ldots, n-1\}$. We make the assumption that these hash functions map each element in the universe to a random number uniformly over the range $\{0, \ldots, n-1\}$. Let $F = \{f_1, \ldots, f_m\}$ be the set of unacceptable passwords. The pre-processing step is the following : for each element $f \in F$, the bits $A[h_i(f)]$ are set to 1 for all $1 \leq i \leq k$. A bit location can be set to 1 multiple times, but only the first change has an effect.

To check if a query element x is in F, we check whether for all array locations $A[h_i(x)] = 1$ (for $1 \leq i \leq k$). If not, we conclude that $x \notin F$. It is easy to verify that we cannot have false-negatives. If all $A[h_i(x)] = 1$, we conclude that $x \in S$, although we might be wrong.

4.1 Let X be the number of positions in A that remain 0 after pre-processing. What is $\mathbf{E}\{X/n\}$?

4.2 To simplify, assume $X = pn$ for $p = e^{-km/n}$. What is the probability P of a false-positive (it may falsely declare a match when it is not an actual match) ? Choose k that minimizes P and p that minimizes P.

4.3 Reconsider (and justify) our assumption that $X = pn$. Use Poisson approximation to bound $\mathbf{P}\{|X - np| \geq \varepsilon n\}$.

Exercice 5 - Nombre chromatique

Recall that the chromatic number $\chi(G)$ is the smallest number of colors needed to color the vertices of G such that two adjacent vertices never share the same color. It might seem reasonable to believe that if the graph does not have short cycles, then $\chi(G)$ should not be too large. This however turns out not to be true. Prove that for any integer $k \geq 2$, there exists a graph G with no triangles and that has a chromatic number $\chi(G) \geq k$.

Exercice 6 - Graphe connexe

An undirected graph on n vertices is *disconnected* if there exists a set of $k < n$ vertices such that there is no edge between this set and the rest of the graph. Otherwise, the graph is said to be *connected*. Show that if $p = (2 + \varepsilon) \log n / n$ for $\varepsilon > 0$, then the probability that a graph chosen randomly from $G_{n,p}$ is connected tends to 1 for $n \to \infty$.