Fun with Kulback and Leibler

As discussed, the Kullback-Leibler divergence is a good way to compare a measured discrete distribution (P) with some known discrete distribution (Q). In both cases, the distributions are normalized with

$$\sum_i p_i = \sum_i q_i = 1 \ .$$

The definition of D_{KL} is simply

$$D_{KL}(p||q) = \sum_i p_i \ln \left(\frac{p_i}{q_i} \right)$$

To understand the physical interpretation of the KL divergence, we consider the most probable result of N independent random draws on P which has $n_i = Np_i$. The probability of drawing this result on the distribution P is Π_P and the probability of drawing the same result on the distribution Q is Π_Q with

$$\Pi_P = N! \prod_i \frac{p_i^{n_i}}{n_i!} \quad \text{and} \quad \Pi_Q = N! \prod_i \frac{q_i^{n_i}}{n_i!} \ .$$

The KL divergence is thus seen to be

$$D_{KL}(p||q) = -\frac{1}{N} \ln \left(\Pi_P / \Pi_Q \right) \ .$$

As a specific example, let us consider what the KL divergence allows us to say about the digits of $\pi - 3$. Specifically, It seems reasonable to assume that the individual digits, 0–9, of this number are drawn independently and at random (i.e., drawn on the distribution $p_i = 1/10$). **What can you say about this assumption using the KL divergence?**

The data is as follows: For several values of N, the number of times the digits 0 to 9 appearing in $\pi - 3$ is given as:

N = 10^3 93 ~~3,~~ 116, 103, 102, 93, 97, 94, 95, 101, 106.
N = 10^4 968, 1026, 1021, 974, 1012, 1046, 1021, 970, 948, 1014.
N = 10^5 9999, 10137, 9908, 10025, 9971, 10026, 10029, 10025, 9978, 9902.
N = 10^6 99959, 99758 ~~999758~~, 100026, 100229, 100230, 100359, 99548, 99800, 99985, 100106.

Things to try:

1. Calculate $D_{KL}(p||q)$ for each of the 4 data sets above assuming that each element of p is $p_i = 1/10$.

2. Now consider the case $N = 10^3$. Make a random draw of 10^3 digits and calculate the KL divergence with p. Do this roughly 1000 times, and determine the fraction of times that this value is greater than that found for the digits of $\pi - 3$.

3. Are such tests conclusive for deciding if the digits of π are randomly distributed? If not, what other tests could you imagine performing.

Something to think about: Consider a string of N random digits, and determine one number — the longest unbroken string of, e.g., the digit 3. Repeat this process many times, and determine the average value of the longest unbroken string of 3's (N_3). Believe, it or not, the answer is[1]

$$\langle N_3 \rangle = \log_{1/r}([1-r]N) - \frac{\gamma}{\ln(1/r)} - \frac{1}{2} \pm \left[\frac{\pi^2}{6\ln^2(1/r)} + \frac{1}{12} \right]^{1/2} .$$

Here, $\gamma = 0.577\ldots$ is Euler's gamma and $r = 1/10$ is the probability of drawing the number 3. Appreciate how fantastic this result is: This average grows logarithmically with N, and the variance is *independent* of N in the large N limit.

For $N = 10^7$, the average value of the longest unbroken string of any number is thus 6.70 ± 0.63. I have checked the first 10^7 digits of $(\pi - 3)$. The longest unbroken string of the digits 0–9 are $(7, 7, 6, 7, 6, 7, 7, 7, 7, 7)$ with an averge value of 6.8 ± 0.4.

For For $N = 2 \times 10^9$, the average value of the longest unbroken string of any number is thus 9.01 ± 0.63. I have checked the first 2×10^9 digits of $(\pi - 3)$. The longest unbroken string of the digits 0–9 are $(8, 9, 9, 8, 9, 8, 10, 9, 9, 9)$ with an averge value of 8.8 ± 0.4.

You can find data for this at `http://www.subidiom.com/pi/pi.asp`.

A. D. Jackson
6 February 2019

[1] In case you are not familiar with logarithms arbitrary base a, $\log_a(x) = \log(x)/\log(a)$ for any choice of a.