

A Needle in the Haystack

Gary Nan Tie, Feb 11, 2009

A Parable

Suppose we have a haystack consisting of 100 Red straws and 1,000 Green straws where:

$\Pr(\text{a Red straw contains a needle}) = 9/1,000$ and

$\Pr(\text{a Green straw contains a needle}) = 1/10,000$.

So the (fewer) Red straws are more likely to contain a needle than the (many) Green straws.

If we sample *with replacement* to screen for a needle, intuitively we should screen more of the Red draws than the Green draws, as they are more likely to contain a needle, but how much more?

For example, in sampling straws with replacement, suppose we screen 9 in 1,000 Red draws and 1 in 10,000 Green draws for a needle, then the mean number of screenings per found needle is 1,100.

Now, if instead we screen 5 in 1,000 Red draws and 5 in 10,000 Green draws for a needle, then the mean number of screenings per found needle is 380. Why the difference? Can we do better?

If resources are scarce or expensive, then fewer the better, the mean number of screenings per found needle, so is there an optimal sampling (with replacement) strategy?

An optimal sampling application of Cauchy's inequality

Cauchy's inequality for real numbers: $\sum a_j b_j \leq \sqrt{(\sum a_j^2)} \sqrt{(\sum b_j^2)}$

(If vector b is unequal 0, equality holds iff \exists constant λ such that $a_j = \lambda b_j$ for $j = 1, 2, \dots, N$.), can for positive weights $w_j, j = 1, 2, \dots, N$, be restated as a bound on the square of a general sum:

$$(\sum a_j)^2 \leq (\sum 1/w_j) (\sum a_j^2 w_j)$$

(Equality iff $\exists \lambda$ such that $1/w_j = \lambda a_j, j = 1, 2, \dots, N$.)

In the context of (1), let p_j denote the prior probability of individual j being a malfeasant, and q_j denote the sampling probability determined by public policy.

Let $a_j = p_j^{1/2}$ and $w_j = 1/q_j$, in Cauchy's inequality restated above, then

$$(\sum p_j^{1/2})^2 \leq (\sum q_j) (\sum p_j / q_j) = \sum p_j / q_j, \text{ and equality occurs iff}$$

$$q_j = p_j^{1/2} / \sum p_j^{1/2}, j = 1, 2, \dots, N, \text{ since } \sum q_j = 1.$$

Thus Cauchy's inequality gives the optimal choice for the sampling q_j 's and the mean number of tests per found malfeasant, found in (1). So *square-root biased sampling* is a sharp instance of Cauchy's inequality.

Moreover, by Carlson's inequality: $(\sum p_j^{1/2})^2 \leq \pi \sqrt{(\sum j^2 p_j)}$

i.e. with the optimal choice of sampling probabilities, the mean number of tests per found malfeasant, has a bound above, involving a second moment under the prior probabilities.

So, to find a needle in a haystack, square-root biased sampling (1) is optimal.

1. Press WH (2009) Strong profiling is not mathematically optimal for discovering rare malfeasants. PNAS vol.106 no.6 1716-1719.