



Why

A natural distribution to associate with a dataset is to assign to each outcome a probability which reflects the number of times it appears in the dataset.

Definition

The *empirical distribution* of a dataset is the function which associates to each outcome the proportion times it appears in the dataset. Since the proportions are nonnegative and sum to one, the function is a probability distribution.

Notation

Let A be a non-empty set and (a^1, \dots, a^n) be a dataset in A . The empirical distribution $q : A \rightarrow \mathbf{R}$ of (a^1, \dots, a^n) is

$$q(a) = \frac{1}{n} |\{k \in \{1, \dots, n\} \mid a^k = a\}|.$$

In other words, to give the probability of outcome $a \in A$, we count the number of times it appeared in the dataset of size n , and then divide by n .

