Empirical Distribution

1 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.

2 Definition

The *empirical distribution* of a dataset is the distribution which associates to each outcome a probability which is the proportion of its appearance in the dataset. The proporitions are nonnegative and sum to one, so the function so described is indeed a probability distribution.

2.1 Notation

Let A be a non-empty set. Let $(a^1, \ldots, a^n)$ be a data set in A. Let $q : A \to \mathsf{R}$ be defined by

$$q(a) = \frac{1}{n} \left| \left\{ k \in \{1, \ldots, n\} \;\middle|\; a^k = a \right\} \right|$$

Then q is the empirical distribution of $(a^1, \ldots, a^n)$. In other words, to give the probability of outcome a, we count the number of times it appeared in the dataset of size n, and then divide by n.