



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

## Notation

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

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

Then  $q$  is the empirical distribution of  $(a^1, \dots, a^n)$ . 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  $n$ .



