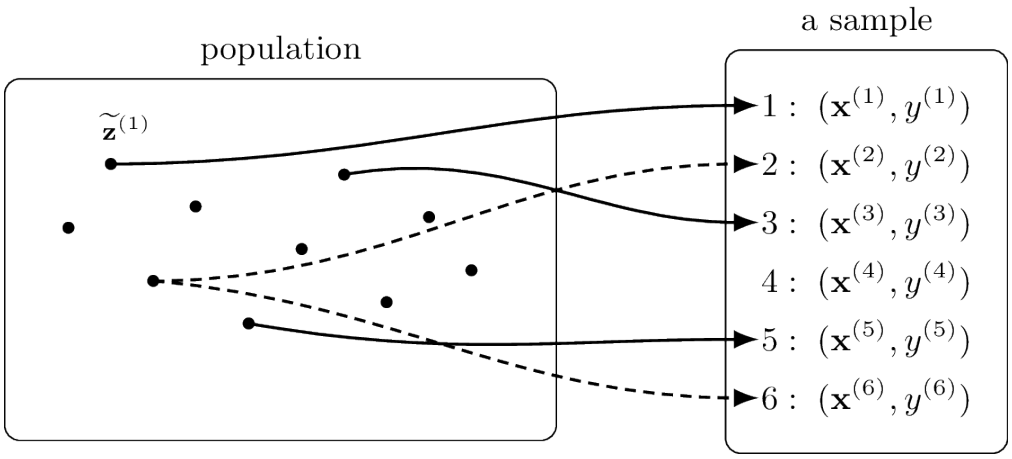

Figure 1: Illustration of a sample as a finite sequence. Each sequence element consists of the feature vector and the label of some data point which belongs to an underlying population. Depending on the application, the same data point is used to obtain multiple sample elements.

Sample

In the context of machine learning (ML), a sample is a finite sequence (of length m) of data points, $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$. The number m is called the sample size. Empirical risk minimization (ERM)-based methods use a sample to train a model (or learn a hypothesis) by minimizing the average loss (the empirical risk) over that sample. Since a sample is defined as a sequence, the same data point may appear more than once. By contrast, some authors in statistics define a sample as a set of data points, in which case duplicates are not allowed [**?**, **?**]. These two views can be reconciled by regarding a sample as a sequence of feature–label pairs, $\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(m)}, y^{(m)}\right)$. The r-th pair consists of the features $\mathbf{x}^{(r)}$ and the label $y^{(r)}$ of an unique underlying data point $\widetilde{\mathbf{z}}^{(r)}$. While the underlying data points $\widetilde{\mathbf{z}}^{(1)}, \ldots, \widetilde{\mathbf{z}}^{(m)}$ are unique, some of them can have identical features and labels. For the analysis of machine learning (ML) methods, it is common to interpret a sample as the realization of a stochastic process indexed by $\{1, \ldots, m\}$. A widely used assumption is the independent and identically distributed assumption (i.i.d. assumption), where sample elements $\left(\mathbf{x}^{(r)}, y^{(r)}\right)$, for $r = 1, \ldots, m$, are independent and identically distributed (i.i.d.) random variables (RVs) with a common probability distribution.

See also: data point, realization, independent and identically distributed (i.i.d.), random variable (RV), probability distribution, sample size, empirical risk minimization (ERM).