熵与不确定性

熵的概念最早来源于热力学，衡量的是能量在空中分布的均匀程度，分布的越均匀，熵值越大。随后，Shannon（1948）[1]将热力学中的熵引入到信息论中，提出了“信息熵”的概念，并指出每条接受的信息都存在冗余，每条信息中排除了冗余后的平均信息量就是“信息熵”。由此，“信息熵”成为了度量不确定性的一种方式。

经济生活中，不确定性是一种常态，衡量不确定性的大小可以根据其出现的概率来度量。不确定越高，越不可能发生，概率越小。换言之，一个不可能发生的事情，当它发生了，会提供更多的信息。一个可行的表示不确定性大小函数应该有两个基本特征：第一，不确定性函数是概率的减函数；第二，两个独立符号所产生的不确定性应等于各自不确定性之和，称之为可加性。能同时满足这两个条件的函数是对数函数。随机变量X有n中可能的取值，分别为，出现的概率分别为,且各事件彼此独立，则信息熵的大小为单个符号不确定性的统计平均数，记为。若随机X为连续时，则信息熵的大小记为。故，将信息熵定义为：

换言之，信息熵是消除不确定性所需信息量的度量，也即未知事件可能含有的信息量。一个事件或一个系统，准确的说是一个随机变量，它有着一定的不确定性。这个随机变量的不确定性很高，要消除这个不确定性，就需要引入很多的信息，这些很多信息的度量就用“信息熵”表达。需要引入消除不确定性的信息量越多，则信息熵越高，反之则越低。

在信息论中，交叉熵（cross entropy）描述了表示两个概率分布p、q（其中p表示真实分布，q表示非真实分布）在相同的一组事件中，用非真实分布q来表示某个事件发生所需要的平均信息。当变量为离散型时，其表达式为：

相对熵（relative entropy）又叫KL离散度（Kullback-Leibler divergence），其含义是：相对于真实分布p，用非真实分布q来表示某个事件发生时所需的平均额外的信息，其表达式为：

KL离散度（Kullback-Leibler divergence）既然可以用来量化两个概率分布之间的差异，那么可以将其作为损失函数（loss function）来使用。KL离散度越小，则表示两个分布越接近。

[1]Shannon C E . A Mathematical Theory of Communication[J]. The Bell System Technical Journal, 1948, 27.