Under TheHood

BY JUSTIN B. KINNEY, PhD

Mutual Information: A Universal Measure of Statistical Dependence

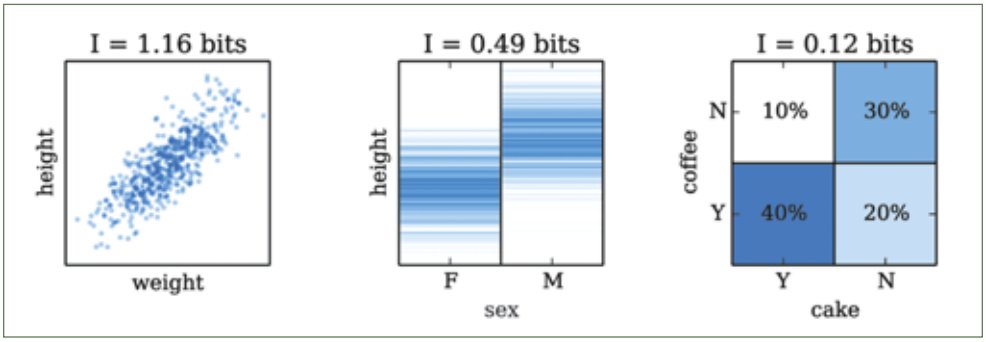
A deluge of data is transforming science and industry. Many hope that this massive flux of information will reveal new vistas of insight and understanding, but extracting knowledge from Big Data requires appropriate statistical tools. Often, very little can be assumed about the types of patterns lurking in large data sets. In these cases it is important to use statistical methods that do not make strong assumptions about the relationships one hopes to identify and measure.

In this tutorial we consider the specific problem of quantifying how strongly two variables depend on one another. Even for data sets containing thousands of different variables, assessing such pairwise relationships remains an important analysis task. Yet despite the simplicity of this problem and how frequently it is encountered in practice, the best way of actually answering it has not been settled.

One standard approach is to compute the Pearson correlation coefficient. Unfortunately, Pearson correlation has severe limitations. First, it only applies to variables that are continuous real numbers; it cannot be used when either variable represents a discrete category, such as gender. Second, the assumptions underlying Pearson correlation are violated by relationships that are nonlinear or have many outliers. Such violations can result in correlation values that conflict with more intuitive notions of dependence.

A more general way of quantifying statistical dependencies comes from the field of information theory. This branch of mathematics arose from a classic 1948 paper by Claude Shannon titled "A Mathematical Theory of Communication." Although Shannon's immediate purpose was to describe information transmission in telecommunications systems, his work illuminated deep truths that have since had a profound impact on fields as diverse as engineering, physics, neuroscience, and statistics.

Shannon argued that the concept of "information" can be formalized by a mathematical quantity now known as "mutual information." Mutual information quantifies the amount of information that the value of one variable reveals about the value of another variable. It is measured in units called "bits:" A value of zero corresponds to no dependence whatsoever, while larger values correspond to stronger relationships.



Data from three hypothetical relationships with corresponding mutual information values shown. Mutual information can quantify dependencies regardless of whether one or both of the variables in question are continuous numbers (e.g., a person's height and weight) or discrete categories (e.g., a person's gender or after-dinner food preferences).

Importantly, mutual information retains its fundamental meaning regardless of how nonlinear a relationship is. Mutual information can also be computed between variables of any type, be they continuous or discrete. Some hypothetical relationships illustrating this are shown in the accompanying figure.

Computing mutual information from data is complicated, however, by the difficulty of estimating a continuous probability distribution from a limited number of samples. Fortunately, there are algorithms that can solve this problem well enough for many practical purposes, and estimating mutual information becomes easier the more measurements one has.

Mutual information therefore provides a sensible alternative to Pearson correlation in many Big Data settings. As better ways of estimating mutual information are developed, this important concept from information theory is likely to become increasingly useful in data analysis efforts, both in science and in industry. □

DETAILS

Justin Kinney is a Quantitative Biology Fellow at Cold Spring Harbor Laboratory.

His research combines theory, computation, and experiment in an effort to better understand quantitative sequence-function relationships in molecular biology. An expanded discussion of mutual information and its merits as a statistic can be found in the recent paper, Kinney, JB and Atwal, GS (2014) Equitability, mutual information and the maximal information coefficient, *PNAS* 111(9):3354-3359.