THE PROPER INTERPRETATION OF SIGNIFICANCE TESTS IN RISK ASSESSMENT

Stephen Spielman

Department of Philosophy
Lehman College and the Graduate Center of the City University of New York
Bronx, New York 10468

In experimental and epidemiological assessments of potential carcinogens, two complementary techniques of data analysis are commonly employed. Data are cranked through an appropriate test of significance to determine whether or not the substance or factor is risk increasing. If the data are statistically significant, the magnitude of the risk to humans is estimated. Such estimates are highly controversial if the study involves extrapolation from data of a large dose animal study. In those cases where the substance tested is a food additive, the main burden falls on the first pattern of analysis under the so-called Delaney clause of the Food, Drug and Cosmetic Act. This is especially true when the substance is weakly, if at all, carcinogenic. In such cases, no clear dose response curve is exhibited by the data, and significance tests play a crucial role.

Four general rules are usually employed in the interpretation of tests of significance in the area of risk assessment. Some definitions are required before they can be described. Let E be an experiment or epidemiological survey designed to test the hypothesis $H_R$ of a positive correlation between exposure and response rate against the null hypothesis $H_N$ of no correlation. E involves a comparison of a test group against a control. Let D be the set of all possible data points for the outcome of E and X a suitable test statistic. X is usually either the familiar chi-square statistic or the z-statistic which is a measure of rate differences between test and control samples. If $d^*$ is the actual data point obtained, $P_N(X \geq X(d^*))$ is the probability that X would take on a value equal to or greater than the actually observed X-value if $H_N$ is true. This probability may be called the "observed significance probability" or simply, the P-value for the test. Since $H_R$ is an hypothesis of increased
risk or positive correlation, the test is one-tailed.

**Rule 1.** (a) If the P-value for E is "statistically significant" (usually $\leq 0.05$), the data point $d^*$ so strongly disconfirms the null hypothesis $H_N$ that it may be rejected. In other words, $d^*$ provides very strong evidence against the hypothesis that the observed differences between test and control groups are due to chance. 
(b) Moreover, the lower P is, the stronger is this evidence.

The next rule is thought to be a corrollary of Rule 1.

**Rule 2.** (a) If the P-value for E is statistically significant, the hypothesis of increased risk is strongly confirmed by the data $d^*$. (b) The lower the P-value the stronger the confirmation.

The next two rules govern "negative" results.

**Rule 3.** (a) If the P-value exceeds the set level of significance, $d^*$ provides some weak evidence for the view that the null hypothesis is close to the truth. The larger the sample size, the stronger is this degree of confirmation. Very large sample sizes are required for "negative" data to have the evidential weight that "positive", ie. significant, data have. (b) Moreover, the larger P is, the stronger the evidence in favor of $H_N$ is.

**Rule 4.** (a) If the P-value is not significant, the data weakly disconfirm $H_R$, the degree of disconfirmation increasing with sample size and (b) increasing with the size of P.

The concept of evidence is central to these rules. As a result, modern theory does not support them. The authors of the currently received theory of statistical testing, Jerzy Neyman and E. S. Pearson, explicitly rejected use of the concept of evidence. In an important early paper they remark:

"We are inclined to think that as far as a particular hypothesis is concerned no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis...Without hoping to know whether each separate hypothesis is true or false we may search for rules to govern our behavior with regard to them, in following which we will insure that, in the long run of experience we shall not be too often wrong."

As Neyman elaborates this idea in subsequent works, significance tests are to be viewed as a branch of decision theory, of the art of gambling with truth, and not as a branch of the art of weighing evidence. We gamble with truth by making decisions in