StatNews #64

Evaluating Statistical Interactions
September 2004

In court, evidence is examined differently depending on the type of trial. In a criminal trial, we first assume innocence, and reject it only if the evidence is strong enough to infer guilt beyond a reasonable doubt. In contrast, in a civil trial, we first assume that neither party is correct. We weigh the evidence, and award the verdict to the party for which there is a preponderance of evidence.

Hypothesis testing in statistics is analogous to a criminal trial. We first assume that there is no effect. Then, we reject this assumption if the evidence is strong enough to infer that the observed effect was very unlikely to have occurred if the assumption was true. We cite as evidence a high value of a test statistic, such as a t-statistic, or a low p-value (often $p < 0.05$). This is, for example, the approach taken when testing main effects in regression models: we assume that the effect is zero and test whether this can be rejected.

When evaluating statistical interactions, however, sometimes the situation is like that of the criminal trial and sometimes is it more like that of the civil trial. In the former situation, we hypothesize that a specific interaction of interest is important. To evaluate this hypothesis, we assume that there is no interaction then test for the interaction in the usual way. The p-value encodes the probability that the magnitude of the observed interaction was due to chance assuming there was, in fact, truly no interaction.

In many other situations, however, we have not hypothesized specific interactions. We might even prefer to describe the data without the complications that interactions bring, but we want to ensure that it is reasonable to ignore possible interactions. In this situation, we want to screen for possible interactions and then decide whether it is better to describe the data with or without an interaction. If we proceed as usual and assume no interaction, then we may have poor power to detect an interaction that is truly there. That is, this screening method for interactions has poor sensitivity (to use language from epidemiology), and will likely miss detecting some interactions that are substantively important.

A solution to this problem of poor power or sensitivity is to screen using a p-value that is higher than 0.05. This is because the power of a test is a function of the p-value we have specified to be used for evaluation, with a higher p-value corresponding to higher power. To illustrate this, we consider a simple example where a p-value of 0.05 corresponds to power of 77%. If the p-value for screening for interactions is increased to 0.15, with all other things being equal, then the power increases to 90%.

Consequently, when the situation calls for screening for possible interactions rather than testing for a hypothesized interaction, then a useful procedure is to screen using a p-value of 0.15 or 0.20. Each interaction with a p-value lower than 0.15 or 0.20 should then be investigated to determine if it is substantively important.

Author: Edward Frongillo