Testing the Bayesian confidence hypothesis
Wei Ji Ma (New York University)

Asking subjects to rate their confidence is one of the oldest procedures in psychophysics. Remarkably, quantitative models of confidence ratings have been scarce. The Bayesian confidence hypothesis (BCH) states that an observer's confidence rating is monotonically related to the posterior probability of their choice. I will report tests of this hypothesis in two visual categorization tasks: one requiring rapid categorization of a single oriented stimulus, the other a deliberative judgment typically made by scientists, namely interpreting scatterplots. We find evidence against the Bayesian confidence hypothesis in both tasks.

Model. Let s be the world state of interest and x a set of noisy visual observations that follow a distribution $p(x|s)$. A Bayes-optimal observer would compute the posterior over s, denoted by $p(s|x)$. We model the observer's decision as a maximum-a-posteriori (MAP) estimate, $\hat{s} = \underset{s}{\operatorname{argmax}}\, p(s \mid x)$, and the observer's confidence rating as a monotonic function F of the posterior distribution evaluated at that estimate, $\gamma = F\left(\max_s p(s \mid x)\right)$. Noise can be added before or after applying F. Even though F has to be postulated on a task-by-task basis, this model always makes two strong predictions: (1) experimental manipulations that leave the posterior $p(s|x)$ unchanged should leave the distribution of confidence ratings unchanged as well; (2) decision and confidence will be correlated in a specific way due to their common dependence on the random variable x.

Results. In Task 1, observers classified an orientation as coming from a narrow or a wide Gaussian distribution with the same mean (Fig. A), and reported their confidence on a scale from 1 to 4. We jointly fitted category reports and confidence ratings (Fig. B-C), across a range of contrasts. We modeled F nonparametrically. The BCH does not account for these data. We discuss a heuristic decision rule that does account for the data.

In Task 2, subjects saw one or two scatterplots representing data drawn from one of two possible linear trends corrupted by noise (Fig. D). Subjects judged which of the trends the data came from and reported confidence on a continuous scale. Matching log posteriors between 1-plot and 2-plot displays, we found that the number of plots affected confidence rating (Fig. E), contradicting the BCH. We discuss a modified model that does account for the data.