Dorst 2023, Overconfidence

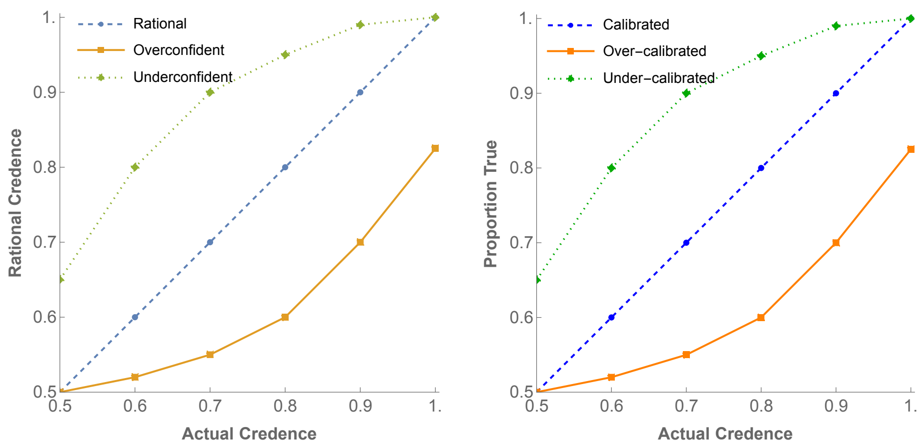
Kevin Dorst

24.223 Rationality

I. Calibration tests

Do people tend to be overconfident—i.e. more confident than it's *rational* to be, given their evidence?

Note: this is a question about the relationship between an empirical quantity ($\overline{C}$) a normative one ($\overline{R}$).



Calibration studies. 2AFC tests.

vs. interval-estimation

If we observe $\overline{T} << \overline{C}$, when is that evidence that $\overline{R} << \overline{C}$?
$\rightarrow$ Only if we have reason to think $\overline{T} \approx \overline{R}$.

When should we?
Short answer: when we *defer* to $\overline{R}$

Bayesians expect *themselves* to be calibrated.

But we are not them; often don't expect rational people to be calibrated:

1. Rajat the BIV
2. Georgie the geographer
3. When is my mother's birthday?
4. Flukey coins
5. Double-sided coins
6. Set of answers you're wrong/right about.

In what sense are these cases abnormal, i.e. do rational opinions *tend* to be right?

II. Deference and Independence

Calvin does a calibration test. What to make of it? Focus on 80%-opinions.

Analogy: Magic Mary and bias-busting Bianca.

If Bianca is calibrated, we get good evidence that she can decipher the coins' biases; if poorly calibrated, we get good evidence she can't. Why?

Principal Principle: defer to the biases of the coins, and the bias screens off the outcomes from each other ("Independence").

Analogy:

Bias of coin $\rightsquigarrow$ rational credence for Calvin to have.
Heads or tails $\rightsquigarrow$ opinion true or false
Align credence with bias $\rightsquigarrow$ align credence with rational credence
Defer to biases $\rightsquigarrow$ defer to rational credences

So we need:

Deference: Upon learning that the average rational confidence for Calvin to have in his 80%-opinions is x%, you should be x% confident in each of them. $P(g_i|\overline{R} = x) = x$.

For all g_i, x.
Fails with Rajat, Georgie, Fluke, & $\mathcal{W}$.

Independence: Given that the average rational confidence for Calvin to have in his 80%-opinions is x%, learning that certain of these opinions are true or false shouldn't affect your opinion in the others. $P(g_{i_0}|\overline{R} = x, g_{i_1}, ..., g_{i_l}, \neg g_{i_{l+1}}, ..., \neg g_{i_k}) = P(g_{i_0}|\overline{R} = x)$.

For all $g_{i_0}, ..., g_{i_k}, x$.
Fails with misprinted coins.

Together, guarantee that conditional on $\overline{R} = x$, *your* distribution for the number of g_i that are true is binomial with parameters x, n.

$\Rightarrow$ conditional on $\overline{R} = x$, you're confident $\overline{T} \approx x$. Thus confident that $\overline{R} \approx \overline{T}$. Inference goes through.

Note: if Independence false, Deference still sets *expectation* of $\overline{T}$ to x—but not necessarily confident it's close.

III. The limits.

This inference is *fragile*: hard to avoid evidence that breaks Deference or Independence. E.g. *hit rate*.

Claim (1): hit rates don't provide direct evidence about rationality.

Sketchy argument for this using monotonicity.

Claim (2): hit rates distort deference.

Like learning it's a tricky test.

So eg $P(g_i|\overline{R} = x, \textit{hit-rate is low}) = x - 0.1$.
Then expect *if rational*, 70% of 80%-opinions will be true.

$\Rightarrow$ miscalibration evidence for rationality.

What to conclude?

Empirical generalization is **hard-easy effect.** Evidence for irrationality?

No—to be expected even if people are rational. Consider Bianca: amongst sets of tablets where her hit rate is low, expect over-calibration. Vice versa if high.