Noisy retrieval of experienced probabilities underlies rational judgment of uncertain multiple events

Leonidas Spiliopoulos[*] and Ralph Hertwig[*]

[*]Max Planck Institute for Human Development

Motivation

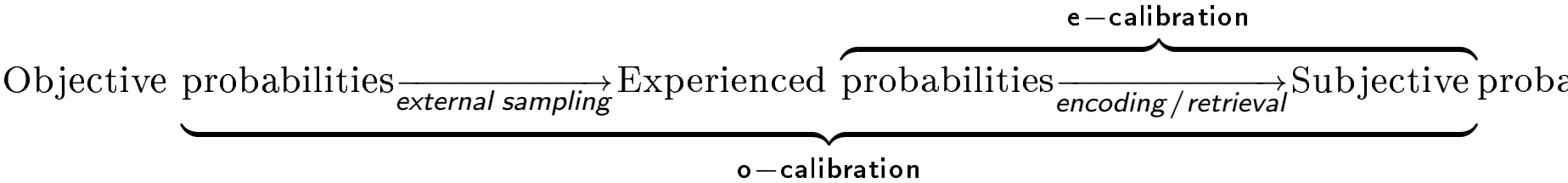
- Learning event likelihoods through experience (not description) is a core competency of any organism, as it is a necessary prerequisite for adaptation to an environment
- Peterson and Beach (1967) considered the "mind as an intuitive statistician"
  - that is, that humans are mostly quite competetent at learning the stochastic properties of an environment
  - Ward Edward's work in the 50s/60s involved learning probabilities through experience

Motivation

- In the 70s, experimental work shifted towards the examination of statistical competence through descriptive tasks and the paradigm changed to one of heuristics and biases
  - e.g., representativeness, anchoring and availability (e.g., Tversky and Kahneman 1973; Kahneman and Tversky 1972)

Research questions

- Do people rationally learn environmental probabilities?
- How accurate or well calibrated are subjective (elicited) probabilities w.r.t.:
  - experienced probabilities (through sampling/observation)
  - objective probabilities



Does observing systematic deviations or miscalibration necessarily imply irrational behavior?

- Consider that no physical systems (the human mind included) can function error-free
- Any descriptive model of behavior must then consistent of a deterministic component and a stochastic/noisy component (e.g., QRE versus NE)

- In non-linear models, errors may not cancel out on average
  - thereby leading to systematic deviations from the normative predictions of a rational model
- However, I would argue that final behavior cannot be used as a judgement of rationality
  - Instead, only the deterministic component of the behavioral model should be used
    - in line with the notion of Simon's bounded rationality

Main findings from the literature

Zacks and Hasher (2002) lit. review:

- people automatically encode relative frequency information
- judged relative frequencies are highly correlated with the actual frequencies
- there is a tendency for over-estimation of low frequency events and under-estimation of high frequency events.

- This pattern is called under-extremity (Griffin and Brenner 2004) and is found in a wide range of tasks:
  - elicited bliefs about opponents behavior in repeated games (Spiliopoulos 2012)
  - visual tasks, guessing the proportion of white versus blacks dots moving randomly (Zhang, Ren, and Maloney 2020)
  - judgments of card combinations in poker (Zhu, Newall, et al. 2022)
  - decisions from experience literature (Hau et al. 2008);(Ungemach, Chater, and Stewart 2009);(Ert and Trautmann 2014)

Basic setup

The majority of existing studies with elicited beliefs typically investigate learning the likelihood of one event (out of two possible events)

- We are interested in investigating how people simultaneously learn a probability distribution of more than two events (three or four).
- There is some evidence that the number of events influence probability judgments in specific ways, but this has not been previously thoroughly investigated

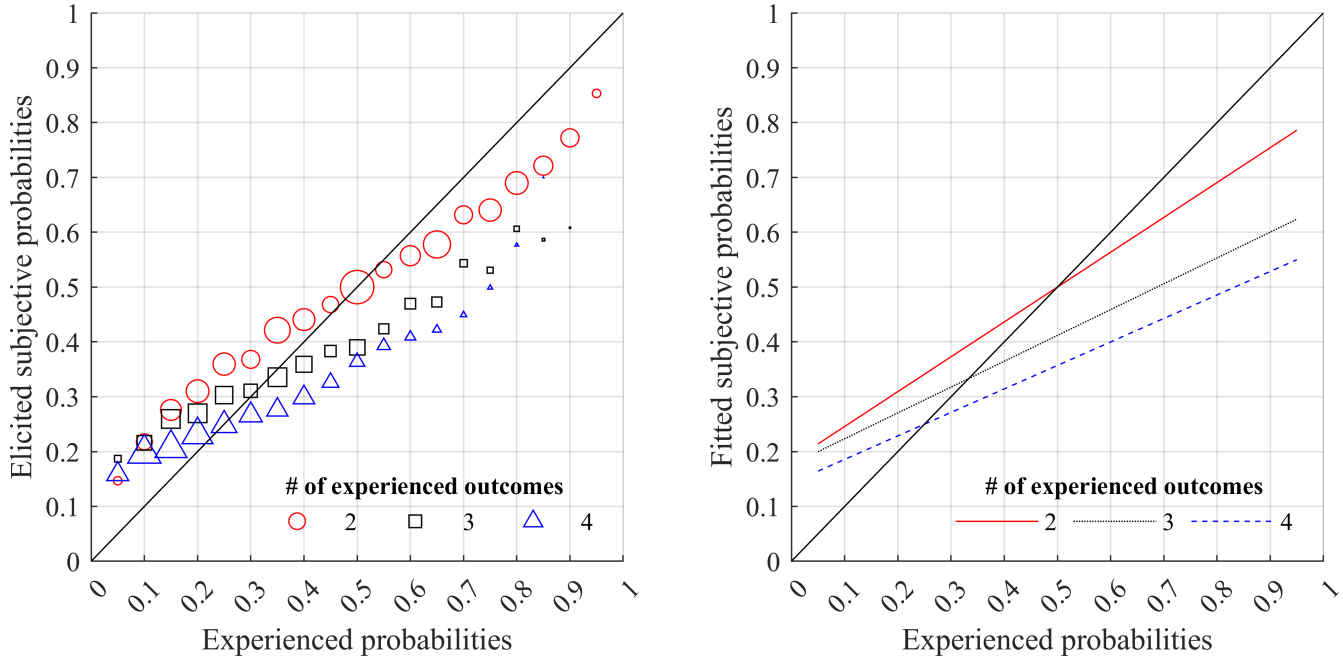
- e.g., estimating the relative frequency of the 26 letters of the alphabet Attneave (1953)
  - the crossover point of subjective and objective probabilities was close to 1/26
  - 1/N effect hypothesized by Fox and Y. Rottenstreich (2003) and See, Fox, and Y. S. Rottenstreich (2006)

Methods

- 96 participants
- 120 pairs of prospects ranging from 1 to 4 outcomes each
  - lottery outcomes and probabilities were drawn uniformly
  - participants did not know how many outcomes each prospect had
- Three stages (conducted on different days)
  - Decisions from description
  - decisions from experience
    - subjects sample as much as they like
  - Belief elicitation (always the final stage)
    - subjects report probabilities of the experienced outcomes (summing to 1)

e-calibration

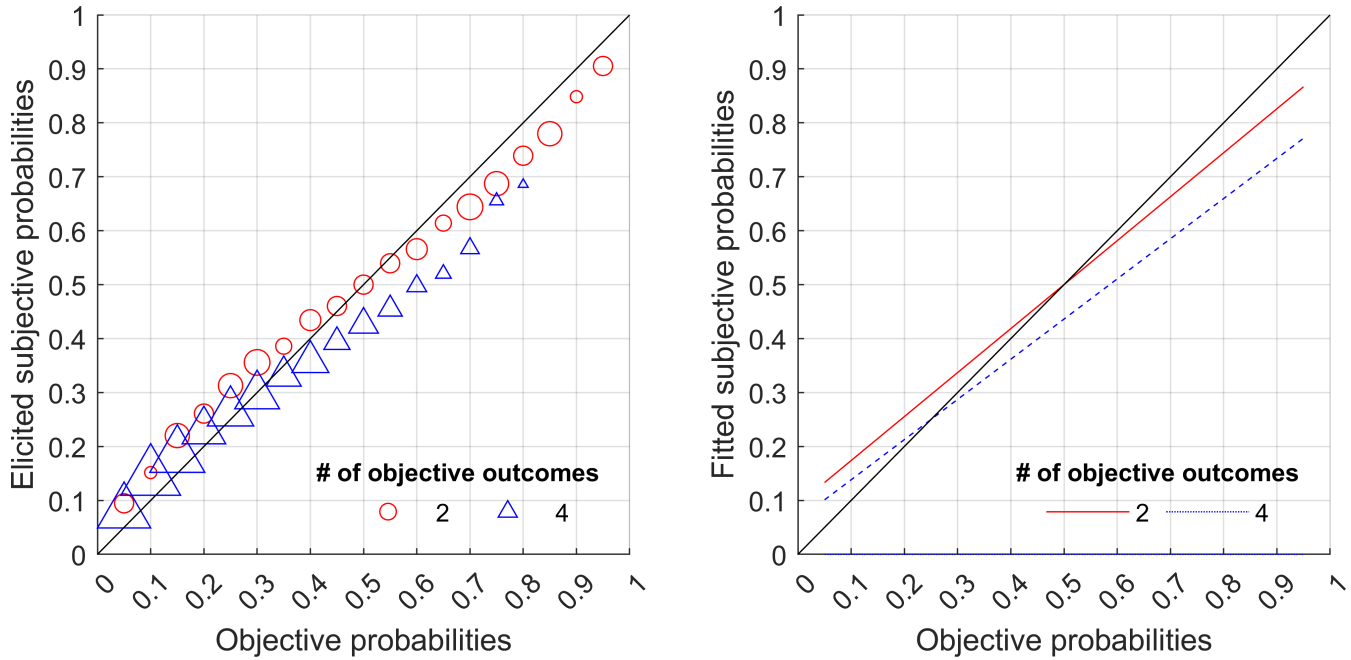
Figure: Subjective versus experienced probabilities conditional on the number of experienced outcomes



- Cross-over points are 0.334 and 0.2491 for N=3 and 4
- Gradient: 0.636 [95% CI: 0.627–0.644], 0.471 [0.457–0.485] and 0.428 [0.412–0.445] for N=2, 3 and 4.
- Cross-over and inverse-sensitivity hypotheses are not rejected

o-calibration

Figure: Subjective versus objective probabilities per number of outcomes



- Cross-over point for N=4: 0.263
- Gradient: 0.815 [0.805–0.826] and 0.745 [0.734–0.756], for N=2 and 4
- Cross-over and inverse-sensitivity are not rejected for o-calibration

Comparison between *e*- and *o*-calibration

- Note from the graphs, and estimated gradients that o-calibration is better than e-calibration
  - the sensitivity of o-calibration curves is closer to the ideal value of 1
  - the difference is due to the effects of limited sampling, as outcomes are sometimes never observed (with implied subjective probaility of zero)
  - this ameliorates the over-estimation of low probability events observed in e-calibration

The model

- During sampling, each outcome is encoded in memory as an instance
  - e.g., if you observe the sequence 100, 120, 100, 200
- To construct subjective probabilities, these instances are retrieved from memory with errors:
  - a fixed ϵ probability of erroneous recall per instance
  - an erroneous recall is then attributed to one of the other events/outcomes
    - errors are more likely to be attributed to events that are more similar
    - similarity = the distance in outcome magnitude
  - this ensures also that the subjective probabilities sum up to 1

The special case without similarity

$$E\left[s_n|e_n\right] = \underbrace{\frac{\bar{\epsilon}_{i,N}}{N-1}}_{\textit{constant}} + \underbrace{\left[1 - \left(\frac{N}{N-1}\right)\bar{\epsilon}_{i,N}\right]}_{\textit{gradient}} e_n \quad (1)$$

- Linear in expected probabilities e_n
- The crossover point occurs when $\bar{s}_n = e_n$, which is at $1/N$.
- Sensitivity, or the gradient of the calibration curve is inversely related to N
  - assuming error rates are independent of N
- The deterministic component is rational, i.e., let $\epsilon = 0$
  - $E\left[s_n|e_n\right] = e_n$

Model estimation and prediction (per individual)

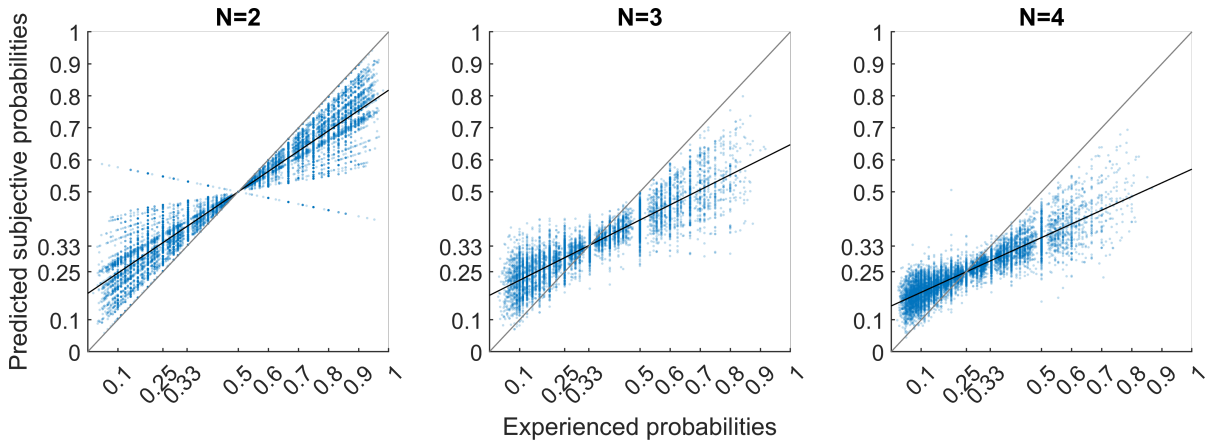
	$\hat{\epsilon}_{i,N}$	$\hat{\lambda}_{i,N}(\times 10^{-2})$	
# of outcomes	Median	Median	$p\left(\hat{\lambda}_{i,N}\right) > 0$
2	0.180	–	–
3	0.358	0.29	0.76
4	0.462	0.36	0.79

- Error increasing with N for 72% of subjects
- λ is significantly greater than zero for 30% and 40% of subjects (N=3 and 4)

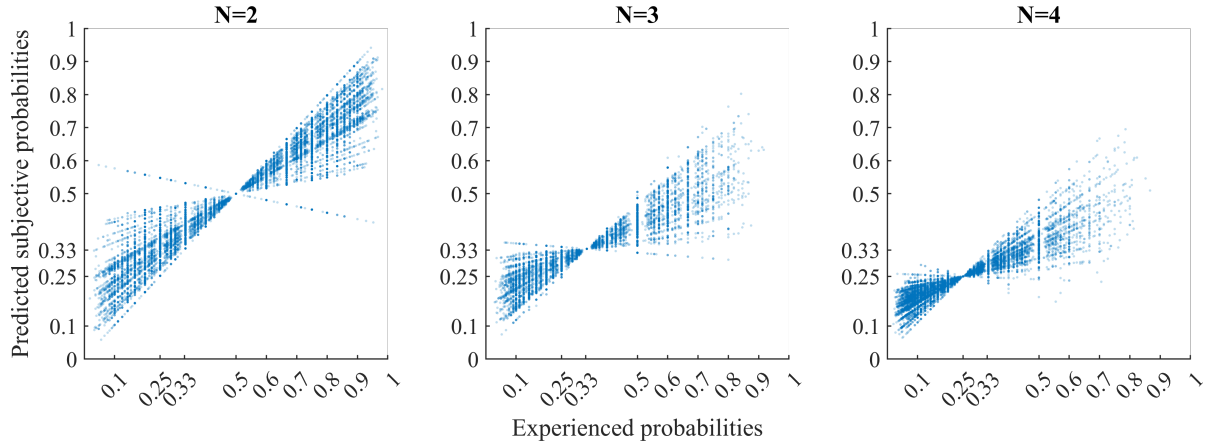
Model estimation and prediction (per individual)

- However, although statistically significant, similarity leads to a very small change in economic significance.
- Essentially, we can use the simplified model where $\lambda = 0$ without significant loss in performance for this dataset
  - but, differences may be larger for more specific prospects that may be more diagnostic
- This means that the 1/N crossover effect essentially still holds, as can be seen in the following graphs.

Figure: Predicted subjective probabilities for SEDM and EDM



(a) SEDM



(b) EDM

Discussion

Empirical regularities affecting calibration:

- 1/N crossover effect
- Inverse sensitivity effect
- A single parameter, the error rate, is sufficient to model both regularities
  - error rates increase with N
- Similarity-based reasoning is found for 30-40% of subjects, but not economically important

Probability judgments are rational, but tempered by noisy retrieval

Discussion

This contrasts other possible accounts:

- Partition-dependent Support Theory Fox and Y. Rottenstreich (2003)
  - judgments are biased toward a uniform (1/N) ignorance prior
  - but in our task, subjects did not know N before sampling
- Bayesian updating (Zhu, Sanborn, and Chater 2020; Zhu, Newall, et al. 2022)
  - no evidence of a negative relationship between # of samples and the relative weight attributed to the prior compared to the sampled likelihood

Discussion

O-calibration is closer to the ideal than e-calibration

- the former is what determines an organism's fitness or adaptability to the environment
- also, it appears to be an improvement over calibration in decisions from description (e.g., over-estimation arising from the probability weighting function in Prospect Theory)

Discussion

- Similarity based reasoning may be extended to think about the detection of possible temporal patterns, i.e., possible correlation across samples and also to account for forgetting
  - e.g., as found in beliefs about opponents' behavior in repeated games (Spiliopoulos 2012; Spiliopoulos 2013a; Spiliopoulos 2013b)
  - e.g., learning conditional probabilities (Plonsky, Teodorescu, and Erev 2015; Plonsky and Erev 2017)

- Attneave, Fred (1953). "Psychological probability as a function of experienced frequency". In: *Journal of Experimental Psychology* 46.2, pp. 81–86. ISSN: 0022-1015. DOI: 10.1037/h0057955.
- Ert, Eyal and Stefan T. Trautmann (2014). "Sampling experience reverses preferences for ambiguity". In: *Journal of Risk and Uncertainty* 49.1, pp. 31–42. ISSN: 0895-5646. DOI: 10.1007/s11166-014-9197-9.
- Fox, Craig R. and Yuval Rottenstreich (2003). "Partition Priming in Judgment Under Uncertainty". In: *Psychological Science* 14.3, pp. 195–200. ISSN: 0956-7976. DOI: 10.1111/1467-9280.02431.

- Griffin, Dale and Lyle Brenner (2004). "Perspectives on probability judgment calibration". In: *Blackwell handbook of judgment and decision making*, pp. 177–199.
- Hau, Robin et al. (Dec. 2008). "The descriptionâexperience gap in risky choice: the role of sample size and experienced probabilities". English. In: *Journal of Behavioral Decision Making* 21.5, pp. 493–518. ISSN: 1099-0771. DOI: 10.1002/bdm.598. URL: http://onlinelibrary.wiley.com/doi/10.1002/bdm.598/abstract;jsessionid=1285E815BFB8BF1B4EBA71E1F9A46A98.f02t02.
- Kahneman, Daniel and Amos Tversky (Jan. 1972). "Subjective probability: A judgment of

representativeness". In: *Cognitive Psychology* 3.3, pp. 430–454. ISSN: 0010-0285. DOI: 10.1016/0010-0285(72)90016-3.

Peterson, Cameron R and Lee R Beach (1967). "Man as an Intuitive Statistician". English. In: *Psychological Bulletin* 68.1, p. 29. ISSN: 0033-2909. DOI: 10.1037/h0024722.

Plonsky, Ori and Ido Erev (2017). "Learning in settings with partial feedback and the wavy recency effect of rare events". In: *Cognitive Psychology* 93, pp. 18–43. ISSN: 0010-0285. DOI: 10.1016/j.cogpsych.2017.01.002.

Plonsky, Ori, Kinneret Teodorescu, and Ido Erev (Oct. 2015). "Reliance on small samples, the wavy recency effect, and similarity-based learning.". English. In: *Psychological Review*

122.4, pp. 621–647. ISSN: 0033-295X. DOI: 10.1037/a0039413.


See, Kelly E., Craig R. Fox, and Yuval S. Rottenstreich (2006). "Between Ignorance and Truth: Partition Dependence and Learning in Judgment Under Uncertainty". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32.6, pp. 1385–1402. ISSN: 0278-7393. DOI: 10.1037/0278-7393.32.6.1385.


Spiliopoulos, Leonidas (2012). "Pattern recognition and subjective belief learning in a repeated constant-sum game". English. In: *Games and Economic Behavior* 75.2, pp. 921–935. ISSN: 0899-8256. DOI: 10.1016/j.geb.2012.01.005.


Spiliopoulos, Leonidas (2013a). "Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching". English. In: *Games and Economic Behavior* 81, pp. 69–85. ISSN: 0899-8256. DOI: 10.1016/j.geb.2013.04.005.


— (2013b). "Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game". In: *Autonomous Agents and Multi-Agent Systems* 27.1, pp. 131–160. ISSN: 1387-2532. DOI: 10.1007/s10458-012-9203-z.


Tversky, Amos and Daniel Kahneman (1973). "Availability: A heuristic for judging frequency and probability". In: *Cognitive Psychology* 5.2, pp. 207–232.


Ungemach, Christoph, Nick Chater, and Neil Stewart (2009). "Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)?" In: *Psychological Science* 20.4, pp. 473–479. ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2009.02319.x.


Zacks, R T and L Hasher (2002). "Frequency processing: A twenty-five year perspective". In: *ETC frequency processing and cognition*. Ed. by P Sedlmeier and T Betsch. New York: Oxford University Press, pp. 21–36.


Zhang, Hang, Xiangjuan Ren, and Laurence T. Maloney (2020). "The bounded rationality of probability distortion". In: *Proceedings of the National Academy of Sciences*

117.36, pp. 22024–22034. ISSN: 0027-8424. DOI: 10.1073/pnas.1922401117.


Zhu, Jian-Qiao, Philip W.S. Newall, et al. (2022). "Clarifying the relationship between coherence and accuracy in probability judgments". In: *Cognition* 223, p. 105022. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2022.105022.


Zhu, Jian-Qiao, Adam N. Sanborn, and Nick Chater (2020). "The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments". In: *Psychological Review* 127.5, pp. 719–748. ISSN: 0033-295X. DOI: 10.1037/rev0000190.