A Toolbox of Methods for Probabilistic Inference

Charles Kemp (ckemp@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Caleb Eddy (cweddy@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

We propose that probabilistic inference is supported by a mental toolbox that includes sampling and symmetry-based reasoning in addition to several other methods. To flesh out this claim we consider a spatial reasoning task and describe a number of different methods for solving the task. Several recent process-level accounts of probabilistic inference have focused on sampling, but we present an experiment that suggests that sampling alone does not adequately capture people's inferences about our task.

Keywords: probability judgment; probability estimation; reasoning; sampling; symmetry

Certainty is often unattainable, and people must therefore maintain degrees of belief. A prominent tradition in cognitive science explores where these degrees of belief come from and how they are updated given evidence. One line of work focuses on normative accounts of reasoning under uncertainty, and many of these accounts rely on probability theory. A distinct but related line of work focuses on process-level accounts that attempt to characterize how probabilistic inference is implemented by the mind and brain.

Recent work on process-level accounts has emphasized the idea that the mind approximates probabilistic inference by sampling (Griffiths, Vul, & Sanborn, 2012; Sanborn & Chater, 2016; Bonawitz, Denison, Griffiths, & Gopnik, 2014). We believe, however, that sampling is just one among many methods that people use for probabilistic inference. Other possible methods depend on symmetry-based reasoning (Strevens, 1998; Vasudevan, 2012), counting events (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Fox & Levav, 2004), computing sums (Fischbein, 1975), products (Fischbein, 1975) and ratios (Zhao, Shah, & Osherson, 2009), and ignoring irrelevant information (Grove & Koller, 1991). This paper lays out an initial proposal about a mental toolbox of such methods. The long term challenge is to understand which methods belong in the toolbox, when they are applied, and how they are flexibly combined. Addressing this challenge is far from straightforward, but essential in order to understand probabilistic inference at the process level.

A longstanding debate in the reasoning literature pits model-based approaches against those that rely on mental proofs. Model-based inference relies on representations of concrete states of affairs, and mental proofs are constructed by applying abstract rules. We believe that both approaches have their merits, and that people draw on both in different contexts. Our toolbox of methods therefore includes model-based approaches such as sampling alongside alternatives that require the construction of mental proofs. A pluralist approach, of course, does not immediately resolve the issues at stake in the debate about models and proofs. Detailed work is needed to establish when people rely on model-based approaches and when they construct mental proofs.

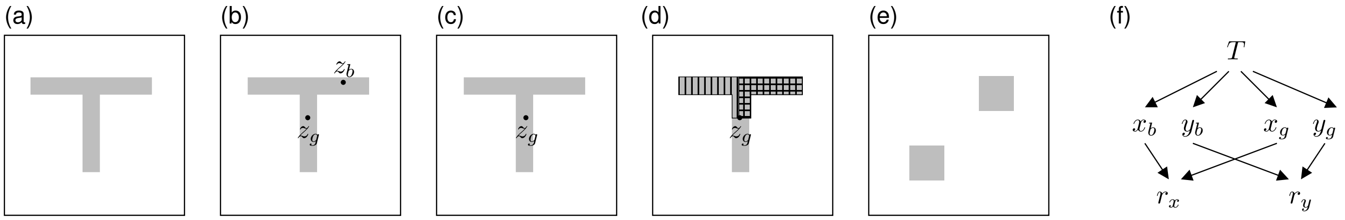
Spatial reasoning task

Because different tasks may elicit different reasoning methods, a comprehensive theory of probabilistic inference should be able to account for a wide range of tasks. As a starting point, we focus here on one simple task. Figure 1a shows a T-shaped rock in a square pond. Suppose that a blue beetle and a gold beetle are both located somewhere on the rock. If the blue beetle is north of the gold beetle, what is the probability that the blue beetle is also east of the gold beetle?

The inferences we consider can be formalized using the graphical model in Figure 1f. Variable T specifies the topography of the pond, and $z_b = (x_b, y_b)$ and $z_g = (x_g, y_g)$ indicate the positions of the blue and gold beetles respectively. These positions depend on T because both must fall on a rock rather than in the water. Variables r_x and r_y indicate relations between the beetles along the x and y axes respectively. For example, $r_x = 1$ indicates that blue is east of gold, $r_x = -1$ indicates that blue is west of gold, and $r_x = 0$ captures the rare case in which neither beetle is east of the other.

In this setting, a *model* is a pair (z_b, z_g) that specifies the locations of both beetles. One example is shown in Figure 1b. Our task is deliberately chosen so that it is impossible to enumerate all possible models. In contrast, some previous research on probabilistic inference focuses on problems for which the set of models is discrete and relatively small, which allows inference methods that depend on enumeration or counting (Fox & Levav, 2004).

Our task has several other appealing properties. It is closely related to a family of tasks known as *three-term series* problems that have been prominent in the reasoning literature (Clark, 1969; Jahn, Knauff, & Johnson-Laird, 2007). One such problem asks "if A is west of B and B is west of C, is A west of C?" Compared to these problems, one advantage of our task is that it allows for a wide range of normative responses. For example, the normative responses to the questions in Figures 1a and 1e are 0.5 and 0.75 respectively, and by varying the shape of the rock it is possible to create a ques-



The blue beetle is north of the gold beetle. How likely is it that the blue beetle is east of the gold beetle?

Figure 1: (a) Our task requires participants to reason about the relative locations of two beetles in a pond. The rock surface is shown in grey, and both beetles are located somewhere on the rock. (b) A complete model that specifies the locations of both beetles. (c) A partial model that specifies the location of the gold beetle only. (d) The normative answer to the canonical question below the ponds can be computed by dividing the area marked with horizontal lines by the area marked with vertical lines. (e) A pond that produces a normative response of 0.75. (f) Graphical model showing the relationships between variables described in the text.

tion with any desired probability as the normative response. A second advantage of our task is that it admits a range of variants that can potentially provide insight into probabilistic inference. One such variant is to ask the same question but to display the position of one beetle, as shown in Figure 1c. Finally, the next section illustrates that our task is useful for exploring probabilistic inference because it can be solved in principle by several methods.

A toolbox of probability estimation methods

We now describe a toolbox that contains eight methods for estimating a conditional probability $P(r_x|r_y)$. This probability corresponds to the strength of an argument in which the premise is r_y (e.g. "blue is north of gold") and the conclusion is r_x (e.g. "blue is east of gold"). To simplify our notation we treat the topography T as background knowledge and drop it from our equations.

Each method is intended to represent a family of related approaches rather than a single precisely-defined algorithm. After introducing each method, we describe one concrete instantiation of the method, but other instantiations of each method are possible.

1. Sample complete models. One way to estimate the probability $P(r_x|r_y)$ is to think of a number of models that make the premise true, and to consider how many of these models also make the conclusion true (Johnson-Laird et al., 1999). We refer to this approach as *complete sampling*, because each model considered provides a complete specification of the positions of the beetles. A normative version of complete sampling is:

$$\begin{aligned} P(r_x|r_y) &= \int_{z_b,z_g} P(r_x|z_b,z_g)P(z_b,z_g|r_y)dz_b dz_g \\ &\approx \frac{1}{m}\sum_{i=1}^{m} P(r_x|z_b^i,z_g^i), \end{aligned} \tag{1}$$

where each pair (z_b^i, z_g^i) is a sample from $P(z_b,z_g|r_y)$. Equation 1 shows how sampling m models is a way to approximate an integral over all possible locations of the beetles, and the approach is normative in the sense that the approximation approaches the probability $P(r_x|r_y)$ as the number of samples becomes large.

2. Sample partial models. An alternative to complete sampling is to work with partial models such as the example in Figure 1c that specify the location of one beetle only. For example, a reasoner might imagine several possible locations of the gold beetle, and assess the probability of the conclusion in each case. A normative version of this method is:

$$\begin{aligned} P(r_x|r_y) &= \int_{z_g} P(r_x|r_y,z_g)P(z_g|r_y)dz_g \\ &\approx \frac{1}{m}\sum_{i=1}^{m} P(r_x|r_y,z_g^i), \end{aligned} \tag{2}$$

where each z_g^i is a sample from $P(z_g|r_y)$. In Equation 1, the term $P(r_x|z_b^i,z_g^i)$ is either 1 or 0, and can be computed by inspecting whether location z_b^i lies to the east or the west of z_g^i. The analogous term in Equation 2 is $P(r_x|r_y,z_g^i)$, which can be computed using the ratio in Equation 3 below, or one of the other methods in the toolbox.

A premise such as "blue is north of gold" locates a figure object (blue beetle) with respect to a ground object (gold beetle), and Equation 2 could be used by reasoners who focus on the ground object. Another approach is to sample possible locations of the figure object. This approach can be formalized using a variant of Equation 2 in which z_g is replaced by z_b.

3. Construct a model-based proof. If the process of sampling models (complete or partial) is accessible to awareness, then reflecting on this process may be enough to derive some conditional probabilities. Suppose for example that a reasoner samples the mental model shown in Figure 1b — a model in which blue is north of gold (as required by the premise) and in which blue is east of gold (as stated by the conclusion). An alert reasoner may notice that this model can be paired with a twin that is identical except that the positions of blue and gold are reflected about the rock's axis of symmetry. In the twin model the premise still holds but the conclusion does not. Further reflection may establish the conviction that each model can be paired with a twin in this way, which means that every model that supports the conclusion is paired with a twin that rejects the conclusion. The predictions based on the models and their twins therefore "cancel out," revealing that $P(r_x|r_y) = 0.5$. The overall chain of reasoning can be formalized as a mathematical proof that refers to models chosen "without loss of generality."

4. Exploit symmetry. The proof sketched in the previous section has two distinctive characteristics: it refers to specific models and it makes use of symmetry. Symmetry, however, can also be used to derive probabilities without needing to consider any specific models. In Figure 1a, a reflection in the T-shaped rock's axis of symmetry maps east onto west and vice versa, but leaves the shape of the rock unchanged. As a result, inverting east and west in any probability statement concerning the rock leaves the probability unchanged. For example, $P(\text{blue east of gold}|\text{blue north of gold})$ must equal $P(\text{blue west of gold}|\text{blue north of gold})$, and because these two probabilities sum to one both must equal 0.5.

Symmetry can also be used to derive an unconditional probability such as $P(\text{blue east of gold}) = 0.5$. It is vanishingly improbable that the beetles have identical x coordinates, which means that blue is either east or west of gold. Given that no available information distinguishes between these states, the *principle of indifference* (Strevens, 1998) implies that both must have a probability of 0.5.

5. Ignore irrelevant information. A basic strategy for simplifying probabilistic inference is to ignore information that has no bearing on the conclusion. If the pond contains an upright square rock, for example, the x and y coordinates of a beetle are statistically independent—knowing one of these coordinates places no constraints on the other. It follows that $P(r_x|r_y) = P(r_x) = 0.5$, where the final step follows from the principle of indifference as described in the previous section.

6. Apply the ratio rule. Suppose that z_g (the position of the gold beetle) is known, as shown in Figure 1c. The conditional probability $P(r_x|r_y, z_g)$ can be computed using

$$P(r_x|r_y,z_g) = \frac{P(r_x,r_y|z_g)}{P(r_y|z_g)}. \tag{3}$$

Equation 3 is simple to compute by estimating the area of two regions in a diagram like Figure 1c. The denominator $P(r_y|z_g)$ is proportional to the area of rock that is north of z_g (indicated with vertical lines in Figure 1d). The numerator $P(r_x,r_y|z_g)$ is proportional to the area that is north and east of z_g (indicated with horizontal lines in Figure 1d).

7. Apply Bayes rule. Bayes rule can be applied as follows:

$$P(r_x|r_y) = \frac{P(r_y|r_x)P(r_x)}{P(r_y)} = P(r_y|r_x), \tag{4}$$

where the final step follows from the observation above that $P(r_y) = P(r_x) = 0.5$. In general $P(r_y|r_x)$ will be no easier to compute than $P(r_x|r_y)$, so applying Bayes rule may not be useful. There may be cases, however, in which one of these probabilities is easier to compute than the other.

8. Enumerate cases. One general strategy for solving a difficult problem is to break it down into a set of simpler subproblems. In Figure 1e, a reasoner may estimate $P(r_x|r_y)$ by considering 4 cases: either both beetles are on the bottom left rock, both are on the top right rock, blue is bottom left and gold is top right, or blue is top right and gold is bottom left. This strategy can be captured formally by introducing a variable v that indicates which of the 4 cases obtains:

$$P(r_x|r_y) = \sum_v P(r_x|v,r_y)P(v|r_y). \tag{5}$$

Each of the sub-problems is simpler than the original. For example, if both beetles are on the same rock, then $P(r_x|r_y) = 0.5$, as argued in our discussion of method 5.

Using the toolbox. We suspect that all eight methods in the toolbox and possibly others are available to human reasoners. Given a problem, a reasoner must therefore decide which method or methods to try. Sometimes two or more methods will need to be combined: for example, methods 2 and 8 ("sample partial models" and "enumerate cases") both express the original probability as a function of several probabilities, which must be estimated in turn.

At present, a detailed mechanistic understanding of probability estimation seems remote. Establishing that people rely on one method for a given task is difficult, because numerous other methods must be ruled out. Establishing that people do not rely on a given method may be more tractable, because only one hypothesis must be ruled out. Given the recent emphasis on sampling as a mechanism for probabilistic inference, we designed a study to explore whether sampling is a plausible account of inference in our setting.

Experiment

We suspect that people rely on sampling when other methods are unavailable, but are able to exploit symmetry when relevant. If so, then people's responses to symmetric ponds might be systematically different from their responses to other ponds. Our experiment was designed to test this possibility.

Participants. 36 participants were recruited using Amazon Mechanical Turk and paid for their participation.

Materials. Participants were asked to reason about the 26 ponds shown in Figure 2. The first 19 ponds are categorized as "double symmetry," "single symmetry" or "no symmetry" ponds depending on whether they have both vertical and horizontal symmetry, only one of these symmetries, or neither vertical nor horizontal symmetry. The next 5 ponds are "non-50" ponds, or ponds for which the normative response is other than 50.

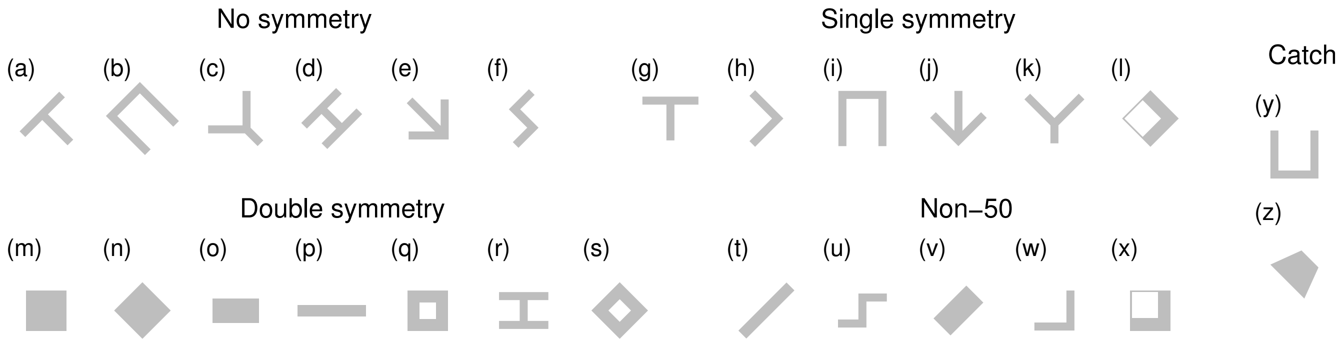


Figure 2: Ponds used in the experiment. The normative response for all ponds in the first three categories (no symmetry, single symmetry, and double symmetry) is 50. Ponds (l), (q), (s) and (x) all have rocks that enclose a body of water.

Procedure. Participants read an introduction that described an eccentric businessman who owned many square ponds. Each pond was said to contain a single gold beetle and a single blue beetle. Participants were told that the beetles could not swim, so each beetle was located somewhere on a rock. They then answered three simple questions that tested their comprehension of what they had just read. They remained on the introductory screen until they had answered all three questions correctly.

Each participant then saw the 24 ponds in Figures 2a-2x in a random order. For each pond, they read that "In this pond the blue beetle is $r1$ of the gold beetle." They were then asked "How likely is it that the blue beetle is $r2$ of the gold beetle?", and required to give their answer on a 0-100 scale with labels at 0 ("Not likely") and 100 ("Very likely"). For each pond and each participant, $(r1, r2)$ was a pair of perpendicular directions (e.g. (north, east), (north, west)) randomly drawn from the set of 8 such pairs.

After the 24 ponds participants responded to two catch trials that had unambiguous answers. One stated that "the blue beetle is east of the gold beetle" and asked participants to rate the likelihood that the blue beetle is west of the gold beetle. The second was similar but used the north-south instead of the east-west axis. The rocks used for these questions are shown in Figures 2y and 2z.

Results. We computed normative responses for each pond by assuming that the location of each beetle was generated from a uniform distribution over the rock surface. For single symmetry and double symmetry ponds, the normative response is always 50. Normative responses for the non-50 and no symmetry ponds were computed by using complete sampling and drawing 100,000 samples. When creating the no symmetry ponds, the dimensions of the ponds (e.g. the relative lengths of the two T-segments in Figure 2a) were adjusted until complete sampling returned a normative result between 49.5 and 50.5.

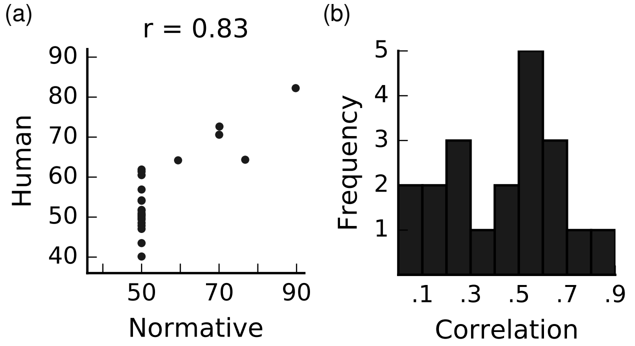


Figure 3: (a) Mean human responses versus normative responses (b) Histogram of correlations achieved by individuals

Because the question associated with each pond was randomized, all responses were converted to responses to the canonical question $P(\text{blue north of gold}|\text{blue east of gold})$. Our conversion assumed that $P(\text{blue north of gold}|\text{blue east of gold}) = 1 - P(\text{blue south of gold}|\text{blue east of gold})$, and similarly for other pairs of opposite directions. We also assumed that $P(r_x|r_y) = P(r_y|r_x)$, as discussed in Method 7 above. We expect that intuitive judgments do not always respect or even approximate this latter identity, but assuming that they do allows for a simple first look at our data.

16 participants failed to give ratings of 0 on both catch trials, and were dropped from all subsequent analyses. Figure 3a shows that mean responses among those who remained roughly tracked normative responses. Each point in the scatter plot corresponds to a pond. For example, the point at the top right of the plot corresponds to Figure 2t. The overall correlation between human and normative responses is 0.83, and Figure 3b shows the correlations achieved by individual participants. Some participants had correlations near zero, but half had correlations exceeding 0.5. Overall, Figure 3 suggests that humans perform relatively well at the task.

The comparison of primary interest is between no sym-

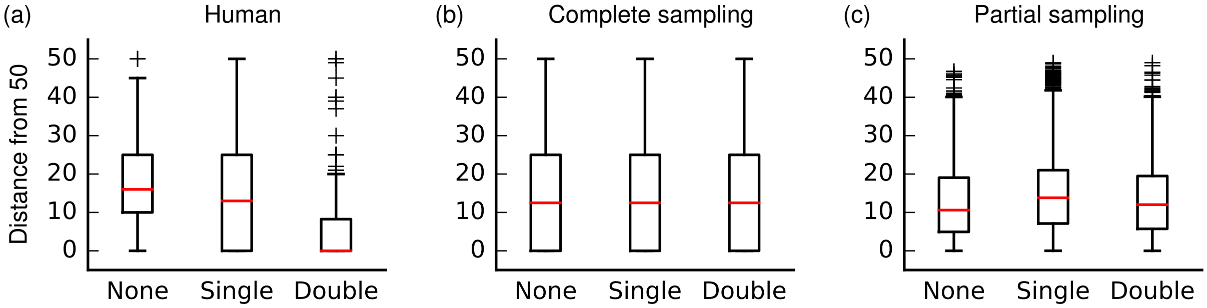


Figure 4: Mean distance from 50 for no symmetry, single symmetry and double symmetry ponds. Results are shown for (a) experimental data (b) complete sampling model with $m = 8$ (c) partial sampling model with $m = 3$. These m values were chosen to approximately match the variability in the human data.

metry ponds, single symmetry ponds and double symmetry ponds. The normative response for these ponds is always 50, and we therefore analyzed the extent to which responses differed from 50. Figure 4a shows that responses for the double symmetry ponds tended to be closer to 50 than responses to the other two kinds of ponds. A Mann-Whitney test indicated that the distance from 50 was greater for no-symmetry ponds (median = 16, n = 120) than for single-symmetry ponds (median = 13, n = 120), U = 5998, $p = 0.012$. A second test indicated that the difference between single-symmetry ponds and double-symmetry ponds (median = 0, n = 140) was also statistically significant (U = 5260, $p < 0.001$). A natural interpretation of these results is that some participants relied on symmetry-based reasoning.

Complete and partial sampling can both be implemented in different ways, but the implementations suggested by Equations 1 and 2 are especially appealing. These implementations are relatively simple, and both approximate the normative response as the number of samples becomes large. Figures 4b and Figures 4c show results for these two implementations. In both cases, the number of samples is chosen so that the model matches the average distance from 50 in the human data. Although matched to humans in this respect, the two sampling models do not account for the special status of the double symmetry ponds in the human data. For example, the complete sampling model predicts no difference between the no symmetry and double symmetry ponds.

A second challenge for a sampling model is whether it can account for the human data given a psychologically plausible number of samples. For the sake of argument, assume that each of our participants is using complete sampling, and that each draws the same number of samples m in Equation 1. Figure 5 shows how the predicted variability in the human data decreases as m increases. If each participant drew one sample only, then some would give responses of 0 and others would give responses of 100, and the average distance from 50 would be 50 for no symmetry, single symmetry and double symmetry ponds alike. If m were very large, then each participant would give a response very close to 50. Figure 5a shows that setting m to 5 or 6 is enough to account for the variability in responses to the no symmetry and single symmetry ponds.

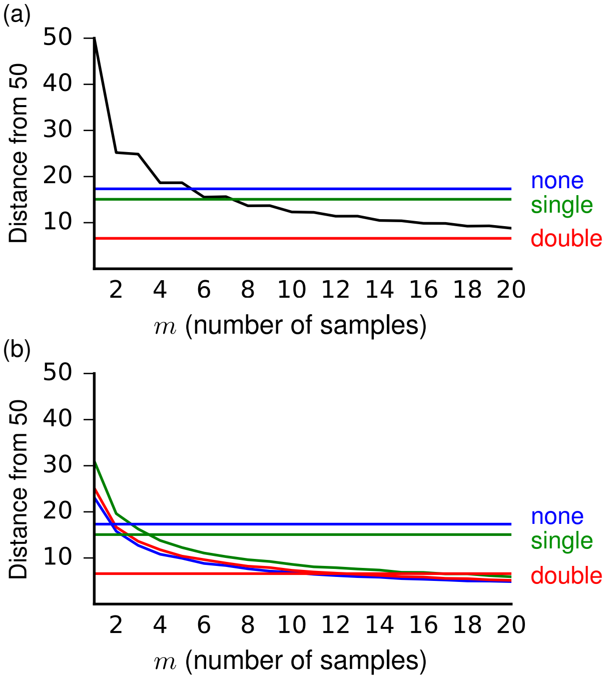


Figure 5: Distance from 50 predicted by (a) complete and (b) partial sampling as the number of samples increases. In (a) a single model curve is shown in black because model predictions are identical for no symmetry, single symmetry and double symmetry ponds. In (b) three model curves are shown because the model predictions for the three classes of ponds are close but not identical.

For double symmetry ponds, however, m must be set higher than 20 in order to match the human data. A value this high does not seem psychologically plausible, and challenges the hypothesis that people rely on complete sampling when reasoning about double symmetry ponds.

Figure 5b shows the analogous plot for partial sampling. In this case, setting m to 10 or so is enough to account for the variability in responses to the double symmetry ponds. This value seems high, but perhaps not high enough to definitively rule out partial sampling as a psychological account. The difference in the human data between double symmetry and both single and no symmetry ponds, however, remains a challenge for models that rely on partial sampling.

Discussion. Although our implementations of complete and partial sampling do not account well for our data, it is possible that other implementations of these methods will perform better. Our implementations assume that models are randomly sampled from the set of all models consistent with the premise of a given argument, but in reality people may sample some kinds of models more often than others. For example, perhaps people prefer to locate the beetle mentioned first towards the left of the pond (Jahn et al., 2007) or towards the top (Levelt & Maasen, 1981). Previous accounts of spatial reasoning have documented effects like these (Jahn et al., 2007), and it seems likely that similar effects will emerge in our setting.

In addition to left-right and up-down preferences, people may also prefer to sample models in which the beetles are located along axes of symmetry. A preference of this kind could help to explain results that are also consistent with symmetry-based reasoning. In Figure 1a, for example, a partial sampling method that uses just one sample will generate the normative response of 50 provided that the single sample locates the gold beetle along the rock's axis of symmetry.

Throughout we have mostly considered inference methods that compute or approximate normative responses. Our data suggest that people's responses to our task are roughly consistent with normative inference, but in other settings people make inferences that are far from normative. For example, base-rate neglect may occur if people apply Bayes rule without including the prior (Kahneman & Tversky, 1973). In other cases people may rely on sampling but sample from the "wrong" distribution—for example, some of our participants may have sampled from $P(z_g)$ rather than $P(z_g|r_y)$ in Equation 2. Each method in the toolbox can be applied in normative and non-normative ways, and detailed work is required to understand how a method is applied in any given setting.

Conclusion

We suggested that people make use of a mental toolbox that includes several qualitatively different methods for probabilistic inference. Each of these methods has several variants, and some methods can be combined with each other. We therefore believe that people can draw on a set of inference methods that is relatively large, which makes understanding probabilistic inference at the process level very challenging indeed.

Like previous researchers we believe that behavioral experiments can provide some insight into the processes that support probabilistic inference. We described a spatial reasoning task that appears to be a natural candidate for inference by sampling, but our results suggest that any simple sampling method is unlikely to fully capture the way in which people approach the task. Ruling out one simple hypothesis about inference is one thing, but providing a comprehensive account of probabilistic inference is another thing entirely. We confess to some scepticism about whether behavioral data alone are enough to reveal the mind's algorithms for probabilistic inference.

Acknowledgments

Caleb Eddy was supported by a fellowship in computational neuroscience that was awarded by the Center for the Neural Basis of Cognition at CMU and the University of Pittsburgh and funded by NIH R90 DA023426.

References

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500.

Clark, H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, *76*(4), 387.

Fischbein, H. (1975). *The intuitive sources of probabilistic thinking in children*. D. Reidel Publishing Company.

Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, *133*(4), 626–642.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263-268.

Grove, A. J., & Koller, D. (1991). Probability estimation in face of irrelevant information. In *Proceedings of the seventh conference on Uncertainty in Artificial Intelligence* (pp. 127–134).

Jahn, G., Knauff, M., & Johnson-Laird, P. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*(8), 2075–2087.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*(1), 62-88.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Levelt, W., & Maasen, B. (1981). Lexical search and order of mention in sentence production. In W. Klein & W. Levelt (Eds.), *Crossing the boundaries in linguistics* (pp. 221–252). D. Reidel.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Strevens, M. (1998). Inferring probabilities from symmetries. *Noûs*, *32*(2), 231–246.

Vasudevan, A. (2012). *Symmetry and probability*. Unpublished doctoral dissertation, Columbia University.

Zhao, J., Shah, A., & Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, *113*(1), 26–36.