Finding hidden types: Inductive inference in long-tailed environments

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Abstract

Making inference in everyday life often requires people to make inferences about low frequency events. In the most extreme case, some types of object or event may have never been previously observed. An experiment is presented in which participants needed to infer the existence and number of unobserved event types, based solely on the frequency distribution of a set of observed events. Results indicate people's inferences are sensitive to the shape of the distribution over the observed events, even when the number of observed events and event types is held constant, and that people are able to infer abstract rules that describe entire classes of event distributions. Human inferences are shown to be similar to those made by a hierarchical Bayesian model.

Keywords: inductive inference, Bayesian cognition, frequency effects, concept learning

Imagine you are walking through the bushlands in a foreign land. You are accompanied by a local guide, who comments on the plant life around you. So far she has described 20 plants as *alba*, 20 plants as *glabra* and another 20 as *eburnia*. On this basis it is tempting to think that *albas*, *glabras* and *eburnias* are the only types of plants around, or at least the only plant types that your guide is intending to label for you. You could not be certain that this is the correct inference of course, but it seems sensible.

Contrast this with a slightly different scenario, in which your guide refers to 58 of the plants as *albas*, points to one example of a *glabra* and one example of an *eburnia*. Again, it is impossible to be sure what to believe, but it seems much less reasonable to conclude that these are the only three plant labels that your guide is ever going to use. Both scenarios involve 60 plants and 3 category labels, yet they do not feel equivalent.

The logic behind this intuition is relatively straightforward. In the second example, you have evidence of the existence of low-frequency types, whereas in the first example you do not. The fact that some types are relatively rare suggests that there may be other rare types that you have not yet seen. In other words, the *shape* of the frequency distribution plays a powerful role in shaping our inductive inferences in this problem. This is illustrated in Figure 1.

In essence, this is a category learning problem: the learner has encountered a new kind of object (the foreign plants) and is attempting to learn the extension of the category with respect to a particular feature (the labels). Viewed as a category learning problem, the different inferences drawn in the two cases are an example of a *frequency effect*, though of a rather different character than the usual exemplar frequency effects. The key difference is that the effect does not pertain to a specific exemplar, but instead is an effect that pertains to the overall frequency distribution. In the first case, the learner has evidence that the frequency distribution is homogeneous: the observed exemplars have equal frequency. In the second case, the evidence implies that the frequency distribution is *long-tailed*, meaning that there are a small number of items that are very common, but most observations are quite rare.

Frequency effects in categorization and choice

Exemplar frequency effects are well-established in the categorization literature: for instance, high-frequency exemplars are classified more accurately, and are judged to be more typical of the category than are low-frequency items (Nosofsky, 1988). However, although the role of item frequency is well-studied (Nosofsky, 1988; Barsalou, 1985; Barsalou, Huttenlocher, & Lamberts, 1998), the inductive inference described earlier is rather different to exemplar frequency effects as they are traditionally conceived. In both examples the *observed* frequency of blue, purple, white or any other color flower is zero, yet they differ in terms of the expected *subjective* frequency. That is, changing the *distribution* of the same set of three types (*albas*, *glabras*, *eburnias*) alters the expectations about the probabilities associated with as-yet-unobserved types.

Frequency effects of a different kind arise in the judgment and decision making literature. In this literature the focus is on how much weight people place on low-frequency outcomes when evaluating possible options, whereas the concept learning literature tends to focus on the role of high-frequency items. Although much of the early evidence (Kahneman & Tversky, 1979; Tversky & Fox, 1995) suggested that people tend to overweight low-frequency events, there is some evidence indicating that this applies primarily to *described* frequencies, and not to *experienced* ones (e.g. Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004), though much of this difference can be attributed to the different information and feedback available to participants (e.g. Rakow, Demes, & Newell, 2008; Camilleri & Newell, 2011). As with the category learning literature, these studies have focused on events whose observed frequency is at least one, rather than looking at the inferences people make about never-observed events.

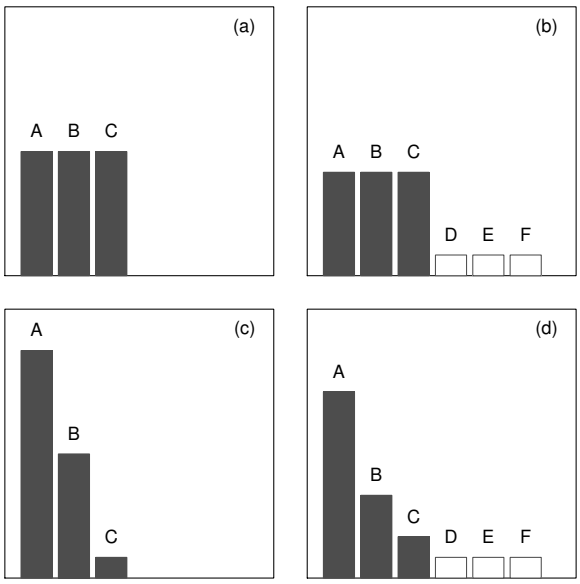

Figure 1: Illustration of why the shape of the type-token distribution matters. Suppose you had observed several observations of types A, B, and C, all of which are equally frequent (panel a). In order to believe that there are more hidden types D, E and F, one is required to postulate that the true distribution looks like panel b. If, however, the empirical frequencies observed were asymmetric (panel c), then in order to believe in hidden types D, E and F, one is required only to postulate a rank-frequency distribution like the one in panel d. To the extent that the distribution in panel b feels less natural than those in panels a, c and d, people should be expected to draw different inferences about unobserved types when presented with uniform data than when they are presented with asymmetric data.

Learning kinds of feature distribution

A recent topic of interest in the concept learning literature is how people learn abstract rules[1] that guide inductive inference in new situations (e.g. Kemp, Goodman, & Tenenbaum, 2010; Perfors & Tenenbaum, 2009). Applied to the current context, the idea would be that people do not merely learn that a single category shows a skewed frequency distribution over object types. Instead, people can learn that "skewness" is a property that is possessed by multiple categories. For example, if we know that the distributions of flood and fire severity are long-tailed (two categories of natural disaster for which a reasonable of data are available to people), we might also guess that the distribution of asteroid strikes (a category of natural disaster largely unknown to people) has a similar shape. One goal of the current work is to see whether people are willing to draw abstract inferences about distributional shape, and use these inferences to alter their guesses about unobserved event types.

Overview

The goal of this paper is to investigate how people infer the existence of unobserved event types, and whether people are sensitive to distributional form when doing so. The structure of this paper is as follows. An experiment is described in which participants were asked to guess how many types of marbles exist in a bag that is only partially observed, where the distribution of observations is manipulated. Human responses in this task are compared to the predictions of a hierarchical Bayesian model that learns both the number of types and the shape of the distribution over types. The implications of the results for the black swan problem that motivated the experiment are discussed.

[1]Throughout this paper, the term "rules" is used informally, and in this context refers to any regularity that people rely upon to guide inference. It is not intended to imply that the regularities in question correspond to explicitly represented, verbalizable rules.

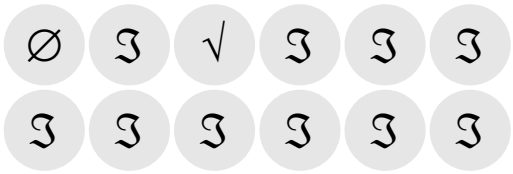

Figure 2: Sample stimulus in the pencil-and-paper version of the task. This was the first trial in the skewed condition (lower left panel in Figure 3). There are 10 tokens of the ℑ type, 1 token of the √ type, and 1 token of the ∅ type. The computerized task was the same, but types were differentiated by color as well as by symbol, and the assignment of symbols was randomized.

Experiment

Method

Participants 101 participants (68% female) were recruited from the University of Adelaide community: 33 were undergraduates participating for course credit, 57 were recruited through a paid participant list, and 11 were graduate students. The 57 paid participants did a computerized version of the task, while the other 44 participants completed a pencil and paper version.

Materials & Procedure The task took the form of a guessing game involving 7 trials. On each trial participants were shown 6, 12 or 18 marbles, and told these had been drawn from a bag containing 100 marbles in total. Each marble belonged to one of several types, indicated by a symbol displayed on the marbles surface, and the participant was asked to guess how many types were represented in the full set of 100 marbles. No feedback was given as to the true number of types. Figure 2 illustrates how a set of 12 marbles belonging to 3 types was displayed.

Participants were randomly assigned into one of two conditions, referred to as the "uniform data" condition and the "skewed data" condition. The number of marbles observed and the number of types they belonged to was identical across conditions. For example, the first trial always showed 12 marbles (tokens) belonging to 3 types, and the second trial always showed 18 tokens that represented 4 types of marble regardless of condition. The conditions differed only in the frequency distribution over types. In the uniform condition, the tokens were evenly divided among types: on trial 1, for instance, there were 4 marbles of each of the 3 types (i.e., a 4-4-4 split). In the skewed condition, the split was highly uneven, with most marbles belonging to a single type: on trial 1, the split was 10-1-1. The complete set of frequency dis-

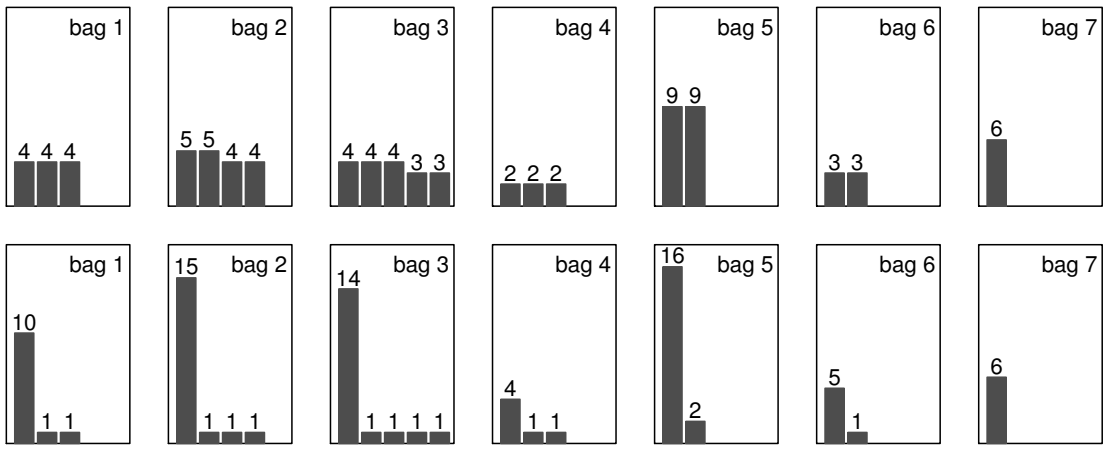

Figure 3: Experimental design. Each panel shows a rank-frequency plot of the marbles on a single trial. The top row shows the type-token distribution for all 7 bags in the uniform condition. The bottom row shows the corresponding distributions for the skewed condition.

Table 1: Descriptive statistics. On the left is a summary of the observations shown to participants on each trial. The middle columns show the 5% trimmed mean response broken down by bag number and condition. The right columns show the proportion of "extrapolative" responses, namely the proportion of responses that imply the existence of at least one unobserved type.

			Mean		Extrapolation	
Bag	Tokens	Types	Unif.	Skew.	Unif.	Skew.
1	12	3	4.35	6.38	0.34	0.47
2	18	4	4.40	7.64	0.18	0.47
3	18	5	5.75	10.54	0.34	0.63
4	6	3	4.58	8.64	0.45	0.70
5	18	2	2.42	2.79	0.14	0.30
6	6	2	3.02	4.13	0.43	0.49
7	6	1	1.32	1.77	0.25	0.47

tributions used in the experiment is shown in Figure 3. Note that the final trial was identical in both conditions.

Exclusions Data from 7 participants were excluded either because they gave impossibly large or impossibly small responses, indicating that they did not understand the task. An 8th participant was excluded for omitting responses. An additional 6 participants gave sensible but qualitatively different responses[2] to the remaining 87. As such, the data from these two groups should not be aggregated, but the minority group is too small to analyze separately.

Results

Table 1 presents an overview of the data. For all seven trials, the average number of types estimated by participants was larger in the skewed distribution condition than in the uniform distribution condition. Moreover, if we classify responses into two categories – those "extrapolative" responses in which participants inferred the existence of at least one hidden type, and responses in which they did not – we observe the same pattern. Participants were more likely to infer the existence of hidden types when the observed frequency distribution was skewed.[3]

To determine if the tendency to estimate more types in the skewed condition represents a significant effect, it is convenient to code the responses in terms of the number of *unobserved* types the participant predicted, rather than the total number of types estimated for the bag. When coded in this fashion, a response of "3 types" on the first trial is treated the same as a "1 type" response on the last one, because in each case the participant has indicated that he or she does not believe there are any hidden types. This has the advantage that a "0 hidden types" response always represents "no extrapolation", and all other responses represent "the extent of the extrapolation" from the sample shown to the participant.

Once the data are coded in this fashion, they can be analyzed using linear mixed effects models, which are well-suited for describing data with a repeated measures structure. In addition to including a fixed effect of condition, the model includes a random effect of bag for each participant in order to capture individual differences in responding.[4] Moreover, because the responses are skewed due to the presence of a floor effect (i.e., "zero" hidden types is a natural lower bound on responses), a Poisson error distribution was used instead of assuming normality.[5] The key result is that the Wald test for the

[2] The responses for these 6 rose monotonically across trials. This pattern makes sense if one assumes the bags are constrained to contain the same set of types. One participant spontaneously reported having made this assumption. This was not the intended interpretation of the task, but it is not an unreasonable one.

[3] One reviewer noted that the gap between skewed and uniform does not increase across trials, and took this to imply that participants were not learning across trials. This is not correct: the trials differ systematically in tems of the number of types and tokens, making it difficult to draw any such inference. The key test of whether cross-trial learning takes place is to look at bag 7: if no cross-trial learning occurs, then responses should be identical for this bag in both conditions, because this stimulus was identical in the two conditions.

[4] Bag was coded as a categorical variable, and the random effect of bag-by-subject subsumes the random effect of subject.

[5] Analyses were run in `R` version 2.15.2 using the `lme4` package version 0.999999-0. Several other model specifications were

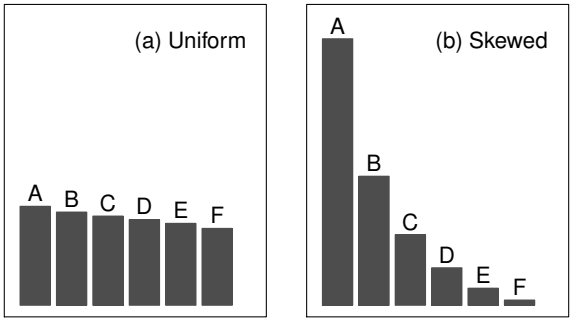

Figure 4: Two different biases that the model can learn, for bags containing $k = 6$ types. In panel (a), the type frequencies are highly uniform ($\alpha = 100$), and the expected rank-frequency plot is quite flat (it becomes perfectly flat as $\alpha \to \infty$). In panel (b), the type frequencies are highly variable ($\alpha = .5$) and the expected rank-frequency plot is very skewed.

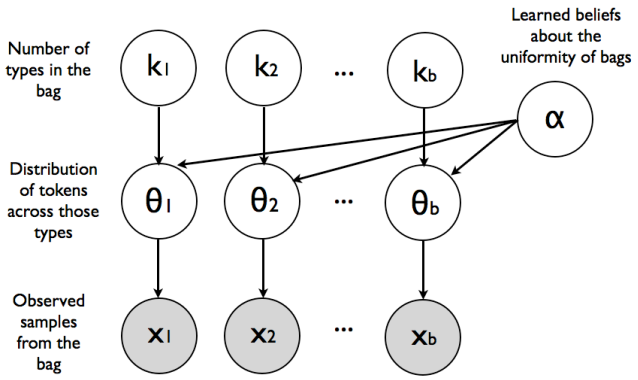

Figure 5: Structure of the model. Shaded circles denote variables that had been observed by participants on or before trial b of the experiment. Unshaded circles denote variables whose values must be inferred. The question asked of participants on trial b corresponds to the value of k_b.

main effect of condition was significant ($z = 3.11, p = .002$): participants did in fact guess that more unobserved types existed in the skewed condition than in the uniform condition.

The previous analysis demonstrated that participants in the skewed condition tended to estimate more hidden types than participants in the uniform condition. In addition to showing that this effect exists across the whole experiment, it is particularly useful to focus on bag 7, as this represents the purest test of whether people were forming theories about bags in general. A two sample Wilcoxon test[6] applied to the bag 7 data revealed a significant difference ($Z = -2.09, p = .037$). Despite the fact that the final bag was identical in both conditions, participants estimated more unobserved types when the preceding bags had revealed a skewed distributional shape.

A probabilistic model of the task

This section outlines a computational analysis of the induction problem used in the experiment. The analysis relies on a probabilistic model of how bags of marbles are generated and how observations are sampled from those bags. It is related to the Bayesian concept learning model used by Kemp, Perfors, and Tenenbaum (2007), but differs in a key respect. Kemp et al. (2007) assume the learner knows the true number of object types in advance, whereas the model used here treats the number of types as an unknown quantity that must be inferred. As with most computational analyses, the model does not describe the processes people use to arrive at estimates. Rather, it provides a sensible standard against which human judgments in this task can be assessed.

Generative model for bags

Suppose that a bag contains k types of marbles, and let θ_i denote the probability that a particular marble will be of the i-th type. We may characterize the bag itself using a vector of type probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. A set of n observed marbles from the bag $\boldsymbol{x}$ can be treated as a multinomial sample of size n generated with probabilities $\boldsymbol{\theta}$. The unobserved marbles can be viewed as a multinomial sample of size $100 - n$ from the same distribution.[7] This model implies that, in a sample of size n, the learner should expect to see $n\theta_i$ exemplars of type i. As such, if n and θ_i are both small, it is quite possible that zero exemplars of type i appear in the learner's observations; it therefore becomes an unobserved type.

This formalism can be extended to provide a generative model for bags, which comes in two parts. First, the number of types k is sampled from some distribution. This paper uses a binomial distribution for this purpose, though this choice is somewhat arbitrary. Second, once k is sampled, the vector of type probabilities $\boldsymbol{\theta}$ is generated. A convenient choice is a Dirichlet distribution with symmetry parameter α. This distribution is widely used by Bayesian concept learning models (e.g. Anderson, 1991; Kemp et al., 2007), and allows the learner to have strong beliefs about the shape of the frequency distribution without knowing a priori which types are more common. If α is small, the learner has a strong expectation that some types of marble will be frequent (Figure 4b) while others will be rare. In contrast, if α is large, the learner possesses a strong expectation that all types of marble should occur with approximately equal frequency (Figure 4a).

An important characteristic of this model is that it satisfies the intuitive constraint illustrated in Figure 1. The uniform distribution in panel a is the expected pattern when $k = 3$ and α is large. The skewed distributions in panels c and d are the expected patterns produced by small α values, with $k = 3$ and $k = 6$ respectively. In contrast, although the distribution shown in panel b is possible within the model, it is not highly likely under any choice of k and α.

Formally, the model is written as follows: if bags are generated with symmetry parameter α, then we obtain the following sampling model for the observations $\boldsymbol{x}$:

$$k|\lambda \quad \sim \quad \text{Binomial}(\lambda, n)$$

tried: none had lower BIC. Inspection of residuals suggests this model provides a good fit to the data. Nevertheless, it is important to note that the effect of condition is robust: it was significant in all model specifications tried, including several that analyzed only the binary version of the response variable (i.e., extrapolative vs non-extrapolative).

[6] The `coin` package (version 1.0-21) in `R` was used to compute an exact p value in the presence of ties.

[7] Strictly speaking, the samples should be constrained such that each type appears at least once among the n observed marbles or the $100 - n$ unobserved ones. For simplicity I have avoided introducing this additional constraint in this paper.

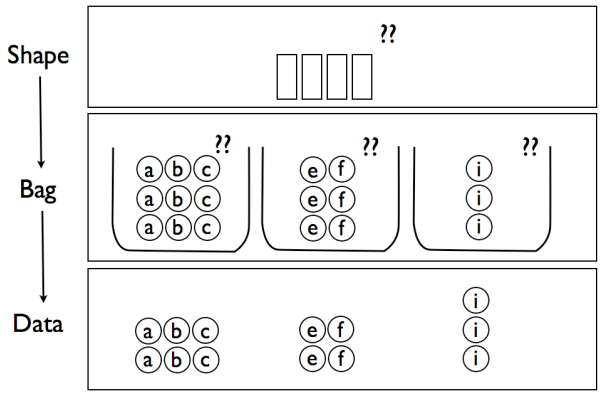

Figure 6: Schematic illustration of the inferences drawn by the model when the observed data are uniform. Things marked "??" refer to values that are inferred by the model rather than observed.

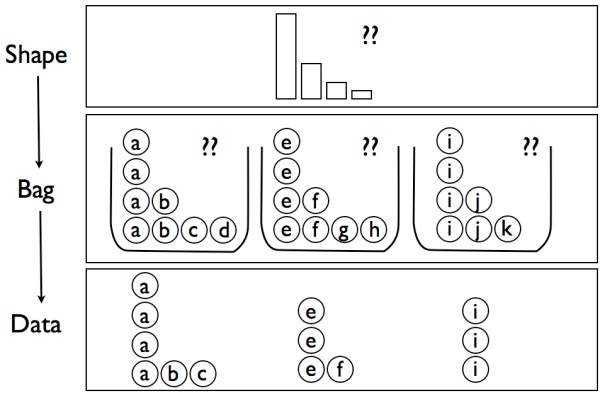

Figure 7: Schematic illustration of the inferences drawn by the model when the observed data are skewed. Things marked "??" refer to values that are inferred by the model rather than observed.

$$\begin{aligned} \boldsymbol{\theta}|k,\alpha &\sim \text{Dirichlet}(\alpha \mathbf{1}^{(k)}) \\ \boldsymbol{x}|\boldsymbol{\theta} &\sim \text{Multinomial}(\boldsymbol{\theta}, n) \end{aligned}$$

where $\mathbf{1}^{(k)}$ denotes a vector of length k that contains only 1s, and n is treated as a fixed property of the experiment and is not part of the generative process over observations. The prior over α is assumed to be a gamma distribution.

The structure of this model as a whole is illustrated graphically in Figure 5. On trial b of the experiment, the learner has access to the samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_b$ from the first b bags (shaded circles). The task as stated is to estimate the the number of types in the b-th bag, k_b, which is one of the several unobserved variables (white circles) whose value is inferred via Bayesian inference.

Learning abstract rules about bags

One of the important patterns in the empirical data is the fact that participants give different responses to bag 7 in the two conditions. The model reproduces this pattern because the symmetry parameter α describes an abstract regularity that attaches to all bags. As such, the model is able to learn the value α across trials. If the model is shown several samples with uniform distributions over observed types, the model will gradually raise the value of α. The value of α tends to decrease when the observed type frequencies are consistently non-uniform.

The consequences of this learning are illustrated in Figures 6 and 7. In the bottom row of Figure 6, the observed samples are evenly split across types, so the model infers a large value for α (top row). The most plausible way to have uniform distributions and remain consistent with the raw data is to have no unobserved types (middle row). Contrast this with the skewed-data scenario in Figure 7. Here the model infers a small value for α and assumes that all of the frequency distributions are also skewed (middle row). Skewed distributions over types imply that at least some types are low probability, so it is entirely plausible to believe that unobserved types exist. As a consequence, the model makes different predictions about the final bag in Figure 7 than it does for the exact same bag in Figure 6.

Model implementation

Although the model specifies many latent variables, the quantity of interest for the b-th bag is $P(k_b|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_b)$, the posterior probability that bag b contains k_b types of marbles, given all of the samples observed so far. This posterior probability cannot be computed analytically: given this, the model was implemented in JAGS (version 3.1.0) and numerical estimates were obtained using Markov chain Monte Carlo. For each bag b, samples were drawn from the joint posterior distribution over all latent variables, and these samples were used to approximated the posterior probabilities of interest.

Because the data presented to participants is different on each trial, fitting the model to the data requires 14 separate model runs. Each of these 1750 model runs involved drawing 1000 samples from the posterior distribution over k after a burn in of 1000 samples. Moreover, because the model predictions depend on the choice of priors, a grid search using 125 different parameter sets was tried. The value of λ was varied from .05 to .25, and the shape and scale parameters for the prior over α were both varied from 1 to 5. The best performing parameter values correspond to a prior over k that is Binomial$(0.15, 100)$ for all bags, and a prior over α that is Gamma$(4, 2)$.

Modeling human data

The model predictions are generally in close agreement with human responses, but there are some differences. The main one is that the model never generates extremely large estimates: human participants occasionally guessed that a bag contained more than 5-6 hidden types, whereas guesses of this kind do not appear at all among the 1000 samples from the model posterior. In other words, although the model produces a distribution over responses that is qualitatively in agreement with human responses, it contains fewer very large values. This difference appears to be due to the fact that the model does not incorporate individual differences: it assumes that all participants have the same priors and rely on the same probabilistic assumptions about the task.[8] Nevertheless, there

[8] In principle there is no reason why a model with individual differences should be avoided: in practice, the computational difficulties in estimating such a model are somewhat severe.

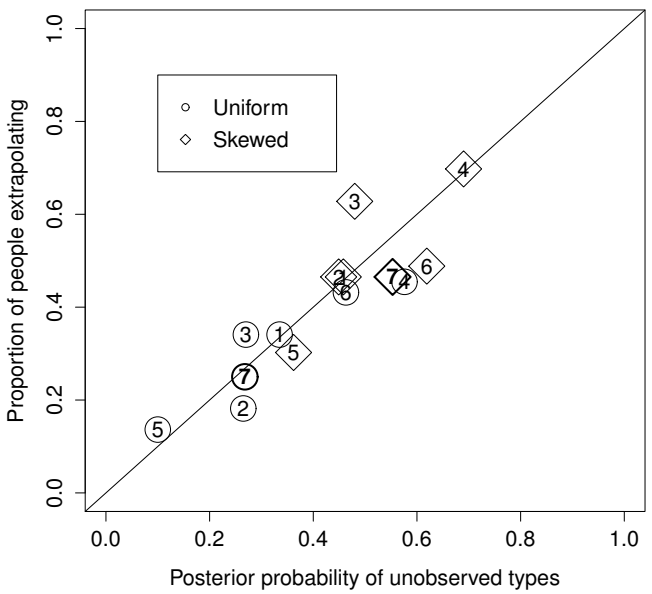

Figure 8: Comparing the model fits against human responses. The Pearson correlation between model and human is 0.89 ($p < .001$).

are individual differences in how people solve the task. There are a few people who consistently estimate large numbers of hidden types, but most do not. This makes it difficult to directly compare model estimate of k against the raw human responses.

A simple solution to the problem is to compare the qualitatively important distinction in the task, namely whether or not a particular response implies the existence of at least one hidden type. That is, instead of fitting the model to the mean number of types estimated by participants (middle columns in Table 1), it is fit to the proportion of human responses in which the number of estimated types was larger than the number of types revealed in the same (right columns in Table 1). These responses are "extrapolative" in that they indicate that the participant has extrapolated beyond the observed data and guessed that there exists at least one hidden type.

Figure 8 plots model estimate of the probability that a bag contains at least one hidden type against the proportion of extrapolative responses in the empirical data. Circles denote bags in the uniform condition, and diamonds represent bags in the skewed condition, and the text denotes bag number. The correlation between model predictions and human responses is 0.89 ($p < .001$) for the best fitting parameter values. However, the model fit is robust: the average correlation across all 125 parameter sets was 0.84, never fell below 0.66, and was significant at $p < .01$ in all cases.

Discussion

The close agreement between model predictions and human responses implies that people are sensitive to the information contained in the *shape* of the distribution of events they have experiences when making inferences about types of events they have never seen. Moreover, the fact that systematic differences existed on the final trial of the experiment, and that these differences are captured via a hierarchical Bayesian model, implies that people are able to use the information from one context (i.e., one bag) to inform the inferences they draw in another one.

One potential extension to this work is to is to consider the role of information search. In the current study, the number of observations sampled from each bag was fixed by the experimenter. However, in many real world decision making problems, people have some degree of control over how much information they collect before making choices. It seems plausible to think that, when the true event distribution is very uneven, people will adopt a very different search strategy than if the frequency distribution is uniform. As such, the constraint that the number of types and tokens observed needed to be matched across experimental conditions, although important from a methodological perspective, may obscure one of the key differences in how people make inferences and choices more generally. In preliminary work investigating this question, we have found some evidence that information search process is indeed influenced by distributional shape, but this is work in progress.

Acknowledgements

This research was supported by ARC grant DP110104949, with salary support from ARC grant FT110100431. I thank Frances Nettle and Natalie May for their assistance with data collection, and Amy Perfors and Nancy Briggs for many helpful discussions.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409429.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215233.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*(3), 203272.

Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic bulletin & review*, *18*(2), 377384.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534539.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263291.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 11851243.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307321.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54.

Perfors, A. F., & Tenenbaum, J. B. (2009). Learning to learn categories. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 136–141). Austin, TX: Cognitive Science Society.

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*(2), 168179.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological review*, *102*(2), 269.