Proceedings of Machine Learning Research 103:367–376, 2019 ISIPTA 2019

Incompletely Known Sample Spaces: Models and Human Intuitions

Michael Smithson MICHAEL.SMITHSON@ANU.EDU.AU
Research School of Psychology, The Australian National University, Canberra, Australia

Abstract

This paper surveys models and human intuitions about incompletely known "sample spaces" (Ω). Given that there are very few guidelines for how best to form such beliefs when Ω is incompletely known, and there is very little research on the psychology behind beliefs about Ω, this survey is preliminary and brings in ideas and models from probability and statistics, biology, and psychology. Pilot experimental studies of how people estimate the cardinality of Ω when given sample information from it are presented, demonstrating that to a surprising extent their estimates correspond with those produced by normative statistical models. The paper concludes by outlining future directions for a research program on this topic.

Keywords: sample space, cardinality, capture-recapture sample, Dirichlet process, imprecise Dirichlet model, human intuition

1. Introduction

Many real-world decisions must be made when we do not know all of the possible relevant states beforehand or the outcomes that could result. The collection of possible states is called the "sample space" and often denoted by Ω. Rational decision-making frameworks require that we base our decisions on whatever beliefs we have about Ω, and these frameworks often assume we have a complete elaboration of Ω. However, there are very few guidelines for how best to form such beliefs when Ω is incompletely known, and there is very little research on the psychology behind beliefs about Ω. In practice, we drastically delimit the set of possibilities. Humans do so in a context-dependent and localized way, even for simple situations where they believe they fully know Ω. For example, when modelling the outcome of a coin toss, Ω will typically taken to be {heads, tails}; and when modelling the outcome of a die toss, Ω will typically be $\{1, 2, 3, 4, 5, 6\}$. Neither of these sets is exhaustive. The coin could land on its edge; the die could fall down a drain before it finishes rolling. Yet we happily discount such possibilities, even if one of them occurs. At the other extreme where we do not know Ω, for example at the advent of a new technology we may not foresee all of its eventual uses and applications, but we should not neglect possibilities such as it being weaponized.

Why is this topic important? Ignoring the potential occurrence of unanticipated possibilities can have serious consequences. For instance, the Australian Department of Agriculture is concerned with all the ways that Australia's crops could be endangered. We may do our best to list them, but our list will inevitably be incomplete. Probability theory and expected utility theory give us no advice on what this Ω should be or how to construct it. Standard frameworks for risk management, such as ISO 31000, do not address this type of unknown. Even if an event in our Ω never has been observed (e.g., foot-in-mouth disease invading Australia), we still should want to estimate how likely it is to occur. Likewise, if we become newly aware of a possibility that should be added to Ω (e.g., a new plant disease), we need to know how to adjust the probabilities already assigned to the other possibilities in Ω. Finally, we should be interested in the likelihood that threats unknown to us in Ω might nonetheless occur. Again, standard probability theory is of little use here.

A reasonable task to pose in the face of ignorance about Ω is estimating how large it might be. We may also want to make predictions about samples from Ω, such as the probability of observing a novel state in the next sample, the expected number of observations to occur before the next novel state is observed, or the number of novel states expected to be observed in the next n samples. Likewise, we may want to estimate probabilities of hypothetical states that have yet to be observed in Ω. This paper surveys prescriptive methods for estimates and predictions regarding an incompletely known Ω and presents preliminary empirical comparisons of some properties of these methods with human intuition.

The main interest in our survey of methods for estimates and predictions about Ω is the nature of the assumptions underpinning these methods. Various models differ on crucial assumptions, and these can result in substantial differences among model estimates, prescriptions for decision makers, and predictions.

We will examine two sources of human intuition about Ω. The first is the literature on the topic of species diversity estimation, containing debates among biologists and biostatisticians. These debates also have underpinned the development of the alternative models discussed in the next section. The second source is laypeople's responses in experimentally controlled tasks. We will focus on two primary questions:

1. What assumptions about Ω are people inclined to make, and what are the properties of their resulting priors? How do their assumptions compare with those in formal models?

2. Humans tend to be highly suggestible when they must make judgments or decisions in an informational vacuum. How severely are people's predictions or estimates about Ω affected by "extraneous" influences such as priming and/or anchoring, versus sample information from Ω itself?

First, however, we will survey statistical approaches to estimating the size and related properties of Ω when samples are taken from it.

2. Methods for Estimating Properties of a Partially Known Sample Space

There are three well-established types of statistical models that may be applied to estimating properties of an Ω whose states are only partially known. Capture-recapture sampling models employed in biology provide methods of estimating the cardinality of Ω [3]. Sample prediction models such as the Pitman-Yor process model [19] provide estimates of the probability that a novel state will be observed in future samples. Imprecise sample prediction models such as Walley's imprecise Dirichlet model [31] provide lower and upper estimates of such probabilities. An exhaustive review of these models is beyond the scope of this paper, so instead we shall focus on comparing the assumptions entailed by the models and forming a program of research on human intuitions that ascertains which assumptions are compatible with human cognition.

2.1. Capture-Recapture Sampling Models

Capture-recapture models estimate population sizes and/or species abundance from so-called "capture-recapture" sampling processes, i.e., sequential sampling with replacement. Sampled species are "marked" so that they are identifiable as having been observed previously when they turn up in subsequent samples. The questions about the structure of Ω typically addressed are its size, and the expected number of observations (or amount of time) required before an exhaustive list of species is accumulated. Beginning with the simplest model, we have K_j distinct species captured on the j^{th} sampling occasion out of a total sample size n_j, with π_j the proportion of the total number of species that would be expected to be retrieved on this occasion. Denoting the total number of species existing on this occasion by κ_j, if we have an estimate of π_j then we may estimate κ_j by $\hat{\kappa}_j = K_j/\hat{\pi}_j$. The classic Lincoln-Petersen [16]; [18] estimate of π_j exploits the capture-recapture process, under the assumption that the sample observations are independent and identically distributed. Given a sample whose species were marked and replaced into the population, the proportion of marked species turning up in the next sample, M_2/K_1 gives an estimate of π_2, so we now have

$$\hat{\kappa}_2 = K_2/\hat{\pi}_2 = K_2K_1/M_2. \tag{1}$$

This model also provides an estimate of the number of new species, U, we should expect to be still undiscovered. Let $S_2 = K_1 + K_2 - M_2$, the total number of species observed in the first and second samples. Then the expected number of species still to be discovered is $U = \hat{\pi}_2(\hat{\kappa}_2 - S_2)$.

Generalizations of the Lincoln-Petersen estimator allow multiple samples and capture probabilities that differ among species, sites, and/or over time. These models have a log-linear or logistic form [3], so they can be expressed via a logistic regression approach. The most sophisticated among them also relax the assumption that Ω is a closed population, thereby modeling rates of species introduction and disappearance.

2.2. Sample Prediction Models

Sample prediction models, unlike capture-recapture models, begin with a single sample of observed species and predict the prevalence of observed and as-yet unobserved species in future samples. These models date back to at least the mid-twentieth century ([11] and [13]), with more concentrated efforts beginning in the 1970's (e.g., [2]; [5]; [14]). The most popular sample prediction models are based on Dirichlet process models as introduced in [8].

The Dirichlet pdf is defined as

$$f(x,\lambda) = \frac{1}{B(\lambda)}\prod_{i=1}^{K} x_i^{\lambda_i - 1}, \tag{2}$$

where $B(\lambda) = \frac{\prod_{i=1}^{K}\Gamma(\lambda_i)}{\Gamma\left(\sum_{i=1}^{K}\lambda_i\right)}$, $\lambda_i > 0$, and $\sum_{i=1}^{K} x_i = 1$. Dirichlet process (DP) models split the λ_i parameters so that they are the product of a "base" distribution, H, and "concentration" parameter $\alpha > 0$. Random samples, $X_1, X_2, \ldots$ are drawn from H in a so-called "size-biased" fashion, via the following steps. First, draw X_1 from H. For $n > 1$:

1. With probability $n_k/(\alpha + n - 1)$ set $X_n = x_{n_k}$, where n_k is the number of times x_{n_k} has been observed so far; otherwise

2. With probability $\alpha/(\alpha + n - 1)$ draw X_n from H.

The simplest DP model is the "Chinese Restaurant Process" (CRP), wherein $n-1$ customers occupy K tables and the n^{th} customer arrives. The probability that the n^{th} customer will sit at an occupied table is $n_k/(\alpha + n - 1)$, where n_k is the number of customers at the k^{th} table, and because the n_k sum to $n-1$, the probability of the n^{th} customer sitting at a new table is $\alpha/(\alpha + n - 1)$.

The probability of generating a new table (i.e., observing a new state) in the CRP is not influenced by the number of observed states, K. There are situations where this restriction is counter-intuitive and criticisms have been raised in several places regarding its validity. The Pitman-Yor process (PYP) generalizes the CRP so that larger K, e.g., greater species diversity, increases the probability of observing a new state. In the PYP, the probability of observing a new state on the sample is $(\alpha + K\beta)/(\alpha + n - 1)$, where $0 \leq \beta \leq 1$ and $\alpha \geq -\beta$. The PYP model boosts this probability most strongly when n is small and when K is close to n. However, the PYP is insensitive to the distribution of the n_k across the states. For instance, it does not distinguish between $n_k = \{20, 20, 20, 20, 20\}$ and $n_k = \{96, 1, 1, 1, 1\}$.

A more fundamental criticism raised against DP models generally is that they do not represent the prior state of ignorance adequately. The Bayesian prior proposed by [22] is the limiting DP as $\alpha \to 0$. This prior is not a non-informative prior, and it also assigns zero probability to any unobserved species. Imprecise DP models have been proposed to overcome this criticism.

2.3. Imprecise Prediction Models

Imprecise DP models were initiated by [30] and [31] and have been extended by [1]. The key idea is to provide a fixed value for α and let the base distribution H span the set of all probability measures. Walley's imprecise Dirichlet model (IDM) is readily generalizable to the CRP and PYP models. It provides a non-Bayesian way of updating lower and upper probabilities for previously observed events and for the catch-all.

Coolen and Augustin [4] present a nonparametric inference model (NPI) that differs from the IDM in several important respects, which I briefly describe here without going into detail about their model. First, lower-upper probabilities of categories in Ω may be influenced by the advent of new categories, regardless of whether those new categories are evidentially relevant to the categories already in Ω. The NPI also permits the difference between the lower and upper probabilities (i.e., the precision of these probability assignments) to be influenced by the advent of new categories. In the NPI, the effect of new categories on lower-upper probability assignments differs depending on whether the refinement occurs within a focal or a non-focal category. Finally, lower-upper probability assignments may be differently affected by whether we are considering the prospect of a heretofore undiscovered category being added to Ω or the prospect of observing a defined but heretofore unobserved category already belonging to Ω.

Our brief survey of methods for estimating properties of an Ω whose states are only partially known provides a rich set of alternative assumptions and hypotheses to compare with human intuitions in this domain. For instance, estimates of the cardinality of Ω may or may not be sensitive to the diversity of observed states and/or the distribution of sampled cases across the states. These and other testable characteristics will be elaborated in the next section.

3. Human Intuitions Regarding Partially Known Sample Spaces

This section begins by outlining a research program, and then presents results from preliminary experiments. The psychological literature contains three primary findings about how people deal with a partially known Ω:

1. They tend to under-estimate the probability that undiscovered alternatives exist in Ω. This literature refers to this phenomenon as the catch-all underestimation bias (CAUB; see [10]; [23]).

2. They tend to anchor on the number of salient states, and their probability assignments are influenced by this (partition dependence; see [12]).

3. The greater the number of ways they think an outcome could occur, the higher the probability they assign to it (support theory; see [29]; [21]).

The third claim has been the most widely disputed (e.g., [24]). There also is some debate over whether the explanations of the first and third findings lie in actual biases among judges, artefacts of sampling error (e.g., [9]), or partition-dependence due to judges applying the principle of indifference.

There are two empirical papers addressing sample-space ignorance directly: [27] and [25]. Briefly, [27] finds that sample space ignorance is aversive, and that people are slower to become more precise in their probability assignments as they acquire data than some imprecise probability assignment schemes recommend. Smithson and Segale [25] demonstrate experimentally that lower-upper probability assignments exhibit partition dependence to a similar degree as precise probability assignments (again, contrary to recommendations from some imprecise probability frameworks).

Other research literature in psychology that is related to our topic includes the large body of work comparing human choices in the presence of "described" versus "experienced" samples from a population. An informative overview of findings and theory development in this domain is provided by [20]. The main theme in this line of research has been highlighting the differences between decisions and preferences under described samples versus experienced sampling. For instance, several researchers such as [7] have claimed that ambiguity aversion disappears when people are able to sample from a population. However, [28] produce evidence that ambiguity aversion does not disappear when ambiguity is located in the sampled cases themselves rather than in their probabilities, which was the focus of the

earlier experiments that were the source of claims about ambiguity vanishing. The complexities involved in studying the impacts of description and experience on preferences and uncertainty attitudes are highlighted by the failure of any single model to win in all types of decision scenarios a model competition organised by [6].

Biologists' descriptions and models of species diversity also are instructive for our purposes because they motivate caution regarding universal models or methods for estimating the size and structure of Ω. Biological diversity models are strongly informed by Ω-specific knowledge, e.g., inter-species competition or predator-prey and parasite-host relations within a site versus between-site differences in resources or other conditions that enhance the survival of a species in one site but not in another.

A recent development in debates about estimating biodiversity is "dark diversity". Pärtel et al. [17] initiated a discussion about whether to include species that are unobserved in an ecosystem but nevertheless "belong" there, i.e., species that relevant experts believe ought to or could readily exist in that ecosystem. They argue that these additional absent species need to be considered in any serious attempts to estimate theoretical carrying capacities of environments or to make meaningful comparisons of biodiversity across similar habitats in which one or more may have suffered local species extinctions. This is a special case of considering propositions or events that are believed to belong in Ω but have yet to be observed, similar to the Coolen-Augustin [4] distinction between a "defined" unobserved event and a hypothetical unspecified unobserved event.

And then there are behavioural influences from the biologists themselves:

1. Unequal taxonomic effort: More effort often is devoted to capturing rare or elusive species than frequently-observed ones. More effort also is expended in accessible habitats.

2. Discovery capability and rate: Technological innovations can alter capture probabilities and habitat accessibility.

3. Amendments to species classification: Taxonomists can argue amongst one another indefinitely about which species an organism belongs to.

The combination of insights from biologists, psychological research on human intuitions about incomplete Ω, and the variety of capture-recapture and sample prediction models presents a rich set of competing testable hypotheses to guide research on how humans deal with sample space ignorance. To begin, the CAUB and the original claims from support theory and the partition dependence literature are now open questions to be asked afresh. For instance, under appropriate conditions can humans produce the Lincoln-Petersen estimate? When (and how) does K, the number of states known thus far, influence the subjective probability of discovery of new states? We have models (e.g., the CRP) suggesting no influence, models (e.g., Pitman-Yor) suggesting that greater K will increase this probability, and models (e.g., partition dependence) suggesting that greater K will decrease this probability. Moreover, according to [13] [3], and [9], the prevalence of rare species (e.g., singletons) rather than K will influence this probability.

Second, if judges are permitted to provide lower and upper estimates, the literature disagrees on the impact of K on the spread of those estimates. Walley's IDM and the imprecise CRP imply that K does not influence the difference between P and $\bar{P}$, whereas the Coolen-Augustin model implies that higher K will increase this difference.

A third set of issues concern the extent to which domain-specific prior information may influence judges' assessments of the size and structure of Ω, including how they react to samples from Ω. In addition to domain-specific knowledge or prior beliefs (such as "dark diversity"), there are potential influences from priming (e.g., making particular states more salient than others), anchoring (where an initial estimate strongly influences subsequent estimates), and beliefs about the underlying mechanisms generating states or events.

We now turn to the preliminary experiments to be reported here, which address the following questions:

1. Can "untutored" humans produce the Lincoln-Petersen estimator, and if so, under what conditions? When presented with capture-recapture sample information, what heuristics do people use to estimate the cardinality of Ω?

2. When are people's estimates of the likelihood of observing a novel state influenced by the diversity of states observed so far?

3.1. Capture-Recapture Estimator Experiments

Two experimental studies included investigations into whether people can produce the Lincoln-Petersen estimator when presented with the required capture-recapture information. Both studies were conducted online with samples from Prolific, a crowd-sourcing platform based in the UK.

Study 1 recruited 400 adult participants, with 207 females and mean age 31.7 (s.d. = 10.3). Participants were given a description of an estimation problem, as in the following example.

> A biologist is trying to estimate the population of carp in a small lake. The carp don't swim in groups but instead are evenly scattered throughout the lake. She drags a large net through the length of the lake and catches 100 carp. She tags them and releases them back into the lake. Shortly thereafter, she drags the net through the

> lake a second time and again catches 100 carp. She finds that 10 of these are carp she had tagged from the first catch. What should be her estimate of the total number of carp in the lake?

The experiment had a 2x2x2 between-subjects design. The factor was whether the biologist was estimating carp population or the number of fish species in the lake. The second factor was whether the resample information was a frequency (e.g., 10 carp from the first catch) or a percentage (e.g., 10% of the carp from the first catch). The third factor was whether the number (percentage) of recaptured or newly captured carp (species) was given.

The primary hypotheses were that the rate of correct responses would be higher if the recapture information was a percentage instead of a frequency, and higher if the sample information was about recaptured than newly-captured fish numbers. Although it is widely held that humans are better at computing with frequencies than with probabilities, the percentage-vs-frequency hypothesis is suggested by the requirement in the problem of having to convert frequencies to probabilities or percentages and then convert those to frequencies in order to arrive at an estimate. Being given the initial information as a percentage eliminates one of the steps and therefore should reduce the likelihood of computational errors.

Also of interest was whether correct response-rates would depend on whether the problem was estimating a fish population or estimating number of species. Based on a previous pilot study, it also was expected that the most popular response would be to add the numbers of fish (or species) unique to each sample together (i.e., $100+90=190$), instead of the Lincoln-Petersen estimate, $100*10=1000$; and that two other popular responses would be $100+10$ and $90*10$.

The numbers in the scenario were intended to make computation easy. Participants were given a multiple-choice response format, although they also could enter their own estimate. Both of these features were intended to provide ideal conditions, enhancing the chances of participants returning the correct answer. Logistic regressions found support for the percentage vs frequency hypotheses, but not for the recaptured vs newly-captured hypothesis (details are available from the author). It also turned out that the correct response rate was higher for the fish population problem than for the number-of-species problem. Table 1 displays the most popular response alternatives, along with the numbers and percentages of participants choosing each of them, for the fish vs species and frequency vs percentage comparisons. The strongest effects in both sub-tables are the opposing differences between conditions in percentages of participants opting for the 190 estimate versus the correct 1000 estimate. In the more favorable conditions, approximately 30% of the participants chose the correct estimate.

Table 1: Study 1 Population Estimation Task Responses

	species		fish	
estimate	freq.	%	freq.	%
110	22	11.9	14	6.7
190	87	47.0	62	29.8
900	24	13.0	42	20.2
1000	33	17.8	62	29.8
other	19	10.2	28	13.5
	freq.		percent	
estimate	freq.	%	freq.	%
110	16	8.5	20	9.8
190	83	43.9	66	32.4
900	34	18.0	32	15.7
1000	32	16.9	63	30.9
other	24	12.7	23	11.3

Study 2 obtained 324 participants with 147 females and mean age 33.4 (s.d. = 12.1). This experiment included a partial replication of the task used in Study 1. Participants were asked for estimates of a fish population with the same sample information as the Study 1 task, in a 2x2x2 between-subjects design whose experimental variables were (non)recapture rate as a frequency vs percentage, the recapture vs non-recapture rate, and multiple-choice vs free text-entry response format. The purpose of comparing the multiple-choice vs text-entry conditions was to evaluate the impact of providing the correct answer in the multiple-choice list. The main hypotheses were that the rate of correct responses would be lower in the text-entry than the multiple-choice condition, higher if the recapture information was a percentage instead of a frequency, and higher if the sample information was about recaptured than newly-captured fish numbers.

Logistic regressions revealed support for the recapture vs non-recapture hypothesis (details are available from the author), and partial support for the other two hypotheses via effects on the odds of participants choosing the 900 estimate (the $90*10$ alternative) and the 190 estimate. Table 2 displays the relevant frequencies and percentages. The frequency vs percentage effect from Study 1 is replicated for the 190 estimate, but there is little evidence of its impact on the choice of the 1000 estimate. The multiple-choice vs text-entry effects are mainly a tradeoff between "other" estimates and choosing the 110 and 900 estimates. It is noteworthy that providing respondents with a free text-entry format did not decrease the percentage producing the Lincoln-Petersen estimate.

3.2. Probability of Novel State Experiments

Studies 1 and 2 also contained tasks in which participants were asked to estimate the probability of heretofore unobserved states occurring in a sample subsequent to the

Table 2: Study 2 Population Estimation Task Responses

	text-ent.		multi-ch.	
estimate	freq.	%	freq.	%
110	3	1.8	13	8.3
190	36	21.4	38	24.4
900	8	4.8	29	18.6
1000	51	30.4	55	35.2
other	70	41.7	21	13.5
	freq.		percent	
estimate	freq.	%	freq.	%
110	8	4.8	8	5.1
190	44	26.3	30	19.1
900	15	9.0	22	14.0
1000	51	30.5	55	35.0
other	49	29.3	42	26.8
	capture		recapt.	
estimate	freq.	%	freq.	%
110	9	5.8	7	4.1
190	41	26.5	33	19.5
900	18	11.6	19	11.2
1000	40	25.8	66	39.1
other	47	30.3	44	26.0

sample provided to them. The main motivation for these tasks was to examine the effects of people's prior intuitions about Ω on their estimates of its cardinality when given sample information about Ω. A specific goal was to determine whether people's estimates of the cardinality of an unknown Ω covary with the number of states, K, they have observed thus far, depending on whether people have had prior familiarity with Ω in general. The "unfamiliar" Ω used in both studies was the proverbial large bag of marbles with unknown distribution of colors, and the "familiar" Ω was an unknown distribution of colors of automobiles in a large city.

In Studies 1 and 2, the marbles scenario consisted of two tasks.

> Imagine that you are a contestant participating in a game-show. The game-show's contest is about how well contestants can predict future outcomes when they're given only a small sample of information. The host shows you a large bag full of thousands of marbles, but doesn't reveal anything about the kinds of marbles in the bag. She then takes 20 marbles from the bag, sorts them into groups with the same colors, and shows these to you and the other contestants. The question she asks is: "If I take 100 more marbles from this bag, how many of them will be colors that are different from the colors we've seen so far?" The contestant whose estimate is closest to the outcome wins this part of the game. Please use the slider to make your best estimate.

In the first task the sample of 20 marbles either contained 4 colors or 15 colors. The second task was identical, but this time the 20 marbles contained either 15 colors or 4 colors, i.e., the 4 vs 15 colors conditions were counterbalanced.

> The host now shows you another large bag full of thousands of marbles. She then asks an audience member to take 20 marbles from this bag, sorts them into groups with the same colors, and shows these to you and the other contestants. The question she asks is: "If I take 100 more marbles from this bag, how many..."

In Studies 1 and 2 the automobile tasks also had two parts, with 4 vs 15 colors counterbalanced between them. The scenario in Study 1 was as follows.

> Imagine that you are a marketing researcher in a large city, studying the popularity of automobile colors. You are with a colleague, counting the colors of automobiles at a busy intersection. You've seen 20 automobiles, sorted them into groups with the same colors, and recorded them on a tablet in the graphic displayed here. Your co-worker asks: "As we observe 100 more automobiles going through this intersection, how many of them will be colors that are different from the colors we've seen so far?" The two of you decide to each estimate this number and bet 10 dollars that theirs is the most accurate. Whose estimate is closest to the outcome wins the bet. Please use the slider to make your best estimate.

In the second part, in one condition the participant was told that they had gone to a different intersection in the same city to sample automobiles. In another condition they were told that they had gone to another city. The rationale for this experimental manipulation was that it should seem more plausible to see, say, 15 colors out of 20 automobiles in one city and just 4 colors in a different city, than to see such different outcomes in the same city.

> You and your colleague have moved to another busy intersection, in the same city (in a different city). As before, you've seen 20 automobiles going through this intersection, sorted them into groups with the same colors, and recorded them on a tablet in the graphic displayed here. You and your coworker decide to bet 10 dollars on the same question as before: "As we observe 100 more automobiles going through this intersection, how many ..."

A summary of the results from this part of Study 1 is presented here. Details of the statistical analyses on which this summary is based are available in [26]. For the marbles

condition 59% of participants behaved as though they were PYP agents (qualitatively, in the sense that they gave higher estimates if they saw 15 colors than if they saw 4 colors), while 38% did the opposite. In the automobile conditions, 29% behaved like PYP agents and 67% did the opposite, so preconceptions about automobile colors appears to have reversed the majority tendency, thereby supporting both primary hypotheses. Interestingly, only small percentages (3% and 4%) gave equal estimates.

Table 3 displays evidence supporting both of the primary hypotheses. For the marbles task, regardless of whether the 15-color or the 4-color sample is presented first, the mean estimate of the proportion of a new sample that will consist of new colors is greater when the number of previously observed colors is 15 than when it is 4. Conversely, the automobile task produces a trend in the opposite direction: Seeing only 4 different colors in 20 automobiles prompts predictions of more new colors in a subsequent sample than 15 colors does. This effect occurs regardless of whether the scenarios take place in the same city or two separate cities.

Table 3: Study 1 Novel State Probability Mean Estimates

4 colors first		4 colors	15 colors
marbles		0.238	0.325
automobiles	same city	0.319	0.226
automobiles	diff. cities	0.264	0.171
15 colors first		4 colors	15 colors
marbles		0.270	0.364
automobiles	same city	0.358	0.258
automobiles	diff. cities	0.299	0.197

The automobile scenario in Study 2 differed from Study 1 by having the participant either go to another unspecified city, or to Hanoi. The rationale for this experimental manipulation was that Western participants might expect a city like Hanoi to have a smaller variety of automobile colors than a European city. Otherwise, this part of Study 2 was a replication of Study 1. A summary of the results from this part of Study 2 is presented next, and again, details of the statistical analyses on which this summary is based are available in [26].

For the marbles conditions 64% of participants behaved as though they were PYP agents, while 30% did the opposite and 6% returned equal estimates. These results are fairly similar to those from Study 1. However, in the automobiles conditions 45% behaved as PYP agents, while 49% did the opposite and 6% returned equal estimates. Thus, foreknowledge about automobile colors lowered the rate of PYP-like responses again, but not as strongly as the effect in Study 1.

Table 4 displays the mean estimates from Study 2. The marbles task replicates the effect found in Study 1, again suggesting that human judges predict a greater variety of colors in a subsequent sample if they have seen 15 rather than 4 distinct colors in the initial sample of 20 marbles.

Table 4: Study 2 Novel State Probability Mean Estimates

4 colors first		4 colors	15 colors
marbles		0.158	0.324
automobiles	city X	0.197	0.208
automobiles	Hanoi	0.174	0.230
15 colors first		4 colors	15 colors
marbles		0.221	0.304
automobiles	city X	0.270	0.193
automobiles	Hanoi	0.240	0.214

The automobiles task results are somewhat more complex than those in Study 1, and they only partially replicate the Study 1 findings. For the unspecified-city condition, the 15-colors-first condition replicates the effect found in Study 1, i.e., the reverse of the pattern seen in the marbles task. However, in the 4-colors-first condition the mean estimates do not significantly depend on how many automobile colors are seen in the initial sample of 20 automobiles. When the new city is identified as Hanoi, there is a mild trend toward the kind of pattern found in the marbles task in the 15-colors-first condition but this is reversed in the 4-colors-first condition.

Roughly speaking, as mentioned earlier the majority patterns in Studies 1 and 2 of the estimates in the marbles task are those we might expect if the judges were PYP agents. Indeed, it is possible to estimate the PYP parameters corresponding to PYP-like participants' new-colors estimates by dividing them by 100 (treating them as probabilities, say, $\pi_1 > \pi_2$). Then we have two equations in two unknowns:

$$\begin{aligned} \frac{\alpha+\beta K_1}{\alpha+n-1} &= \pi_1 \\ \frac{\alpha+\beta K_2}{\alpha+n-1} &= \pi_2 \end{aligned} \tag{3}$$

Given $K_1 = 15$, $K_2 = 4$, and $n = 20$, it can be shown that $\beta \geq 1$ if $\pi_1 \geq 11/15$ and $15\pi_1 - 4\pi_2 \geq 11$. Out of 149 potential PYP cases in Study 1, only 11 (7%) displayed these pathological characteristics, so for all others $0 < \beta < 1$, compatible with a PYP model. However, for 34 of the 149 cases (23%) $\alpha < -\beta$. Note, however, that these probability assignments in themselves are not pathological or irrational. In Study 2, of the 163 potential PYP cases, only 7 (4%) cases had $\beta \geq 1$ but 56 (34%) cases had $\alpha < -\beta$. Our results suggest that a substantial majority of PYP-like agents are returning estimates that are compatible with a PYP model. However, at present we lack a model that would account for the others' estimates. Including a "think-aloud" procedure in a future study should reveal more about how people arrive at their estimates.

4. Conclusions and Future Directions

Our investigations have revealed some promising avenues for advancing our understanding of human intuitions regarding the size of Ω, and how they compare with formal methods for estimating this cardinality. First, we found that some people are capable of producing the Lincoln-Petersen estimate when given capture-recapture sample information, and even without prompting. Reasonable questions to raise at this point for future research are what distinguishes these people from those who use a different heuristic (e.g., are they more numerate?), and when even the Lincoln-Petersen estimators may be persuaded to use different strategies for estimating the size of Ω (e.g., as in the "dark diversity" concept from biology). For instance, given that the most popular automobile colour is white, would people add "white" to their predicted count of automobile colours even if capture-recapture samples of automobile colours in, say, London did not include any white automobiles?

Likewise, we found that greater diversity in a sample from Ω induces either a Pitman-Yor-like heuristic or its opposite. A clear candidate explanation for which heuristic people use is their prior intuitions about the size of Ω. Those who believe its potential size could be much larger than K will tend to behave as PYP-like agents, while those who believe its size is close to K will behave in the opposite way. It is remarkable that none of the formal models referenced by this paper explicitly incorporate a prior on K, and this would seem to be a path for further development in this type of model. The experiments reported here manipulated cues for prior beliefs about size of Ω in a rather crude way (marble vs automobile colors). In the next round of studies this will need to be more systematically manipulated and people's prior estimates of Ω's cardinality will need to be elicited for comparison with their updated estimates. Finally, the experiments reported here did not permit participants to give imprecise estimates. The next round of studies will do so, and imprecise DP models can be deployed for comparison with human intuitions.

Several additional lines of research on this topic could be pursued. This paper concludes by briefly considering one of them, namely the effects of adding or subtracting states in Ω on existing probability assignments. Analogical examples in biology would be the appearance of new species in a habitat or the extinction of current species from the habitat. Standard probability frameworks, including Bayesian, are silent about how an agent should revise probabilities when new states are added to or existing states subtracted from Ω.

One suggestion regarding probability assignment revisions [15] is a "reverse-Bayes" principle that the ratios of probabilities of states, $P(X_k)/P(X_j)$, for X_k and X_j already in Ω should not be affected by the addition of new states to Ω or the removal of other states from it. This would be a very convenient principle to adopt, because it would enable modelers to make inferences about compositional aspects of an incompletely known Ω by making inferences based on odds and odds-ratios involving the states that are known. However, it is not difficult to find counter-arguments or examples against this idea, and a potentially interesting debate about when and how this principle should be applied has yet to take place. Moreover, some probability models dealing with the issue of probability reassignments under alterations of Ω do not adhere to this principle (e.g., [4]).

One consideration is that novel states in Ω could amount to an expansion (tacking another state, X_{K+1}, onto the K old states) or a refinement (splitting a previously unitary state, X_k, into two or more states, say, X_{k1} and X_{k2}). An obvious hypothesis regarding the effect of a refinement is that $P(X_k)$ remains as it was, and we simply have $P(X_{k1}) + P(X_{k2}) = P(X_k)$. But the refinement may yield a change in $P(X_k)$. Likewise, an expansion could introduce a state that is evidentially linked with a previously known X_k, thereby changing $P(X_k)$. In either case, the reverse-Bayes condition cannot hold. In any case, we have plenty of motivations for experimentally investigating the effects of adding (or subtracting) states from Ω on people's prior probability assignments, for both expansion (or contraction) and refinement (or coarsening).

Acknowledgments

The author is grateful for discussions with Alan Hájek and Katie Steele about criteria for constructing Ω, and for constructive and insightful comments from the reviewers of this paper.

References

[1] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pages 1026–1034, 2014.

[2] David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[3] Anne Chao. An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, 2001.

[4] Frank PA Coolen and Thomas Augustin. Learning from multinomial data: a nonparametric predictive alternative to the imprecise dirichlet model. In *ISIPTA*, volume 5, pages 125–134, 2005.

[5] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[6] Ido Erev, Eyal Ert, Alvin E Roth, Ernan Haruvy, Stefan M Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1):15–47, 2010.

[7] Eyal Ert and Stefan T Trautmann. Sampling experience reverses preferences for ambiguity. *Journal of Risk and Uncertainty*, 49(1):31–42, 2014.

[8] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[9] Klaus Fiedler, Christian Unkelbach, and Peter Freytag. On splitting and merging categories: A regression account of subadditivity. *Memory & Cognition*, 37 (4):383–393, 2009.

[10] Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein. Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2):330, 1978.

[11] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[12] Craig R Fox and Yuval Rottenstreich. Partition priming in judgment under uncertainty. *Psychological Science*, 14(3):195–200, 2003.

[13] IJ Good. The estimation of probabilities. cambridge. *Mass.. MlT*, 1965.

[14] Bruce M Hill. Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American Statistical Association*, 74(367):668–673, 1979.

[15] Edi Karni and Marie-Louise Vierø. " reverse bayesianism": A choice-based theory of growing awareness. *American Economic Review*, 103(7):2790–2810, 2013.

[16] Frederick Charles Lincoln et al. *Calculating waterfowl abundance on the basis of banding returns*. US Dept. of Agriculture, 1930.

[17] Meelis Pärtel, Robert Szava-Kovats, and Martin Zobel. Dark diversity: shedding light on absent species. *Trends in ecology & evolution*, 26(3):124–128, 2011.

[18] Carl Georg Johannes Petersen. The yearly immigration of young plaice in the limfjord from the german sea. *Rept. Danish Biol. Sta.*, 6:1–48, 1896.

[19] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[20] Tim Rakow and Ben R Newell. Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23(1):1–14, 2010.

[21] Yuval Rottenstreich and Amos Tversky. Unpacking, repacking, and anchoring: advances in support theory. *Psychological review*, 104(2):406, 1997.

[22] Donald B Rubin. The bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

[23] J Edward Russo and Karen J Kolzow. Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception and Performance*, 20(1):17, 1994.

[24] Steven Sloman, Yuval Rottenstreich, Edward Wisniewski, Constantinos Hadjichristidis, and Craig R Fox. Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30 (3):573, 2004.

[25] Michael Smithson and Carl Segale. Partition priming in judgments of imprecise probabilities. *Journal of Statistical Theory and Practice*, 3(1):169–181, 2009.

[26] Michael Smithson and Yiyun Shou. Ambiguity and conflict aversion when uncertainty is in the outcomes. The Australian National University, Canberra, Australia, 2019.

[27] Michael Smithson, Thomas Bartos, and Kazuhisa Takemura. Human judgment under sample space ignorance. *Risk, Decision and Policy*, 5(2):135–150, 2000.

[28] Michael Smithson, Daniel Priest, Yiyun Shou, and Ben R Newell. Ambiguity and conflict aversion when uncertainty is in the outcomes. *Frontiers in psychology*, 10, 2019.

[29] Amos Tversky and Derek J Koehler. Support theory: A nonextensional representation of subjective probability. *Psychological review*, 101(4):547, 1994.

[30] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[31] Peter Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–57, 1996.