Incompletely Known Sample Spaces: Models and Human Intuitions

Michael Smithson

The Australian National University

2-6 July 2019 /SIPTA 2019

Incomplete sample spaces

This is a preliminary investigation of models and human intuitions about incompletely known "sample spaces" (Ω).

- We bring in ideas and models from probability and statistics, biology, and psychology.
- These ideas enable experiments on how humans estimate the cardinality of an unknown Ω when given sample information from it.
- We conclude with some ideas for a research program on this topic.

State of the art

Very few guidelines for how best to form beliefs about incomplete Ω
Very little research on how humans do this
Where to start?

- How large do we think Ω is?
- Probability of observing a novel state in the next n samples?
- Probability of unobserved state(s)?

Where can we get ideas about human intuitions and/or models?

- Biologists and biostatisticians, regarding species diversity estimation
- Experiments on laypeople

Biostatistical and sample prediction models

Three well-established types of statistical models for estimating properties of an unknown Ω:

- Capture-recapture sampling models for estimating cardinality of Ω
- Sample prediction models such as the Pitman-Yor process for estimating the probability that a novel state will be observed in future samples
- Imprecise sample prediction models such as Walley's imprecise Dirichlet model

These models may differ on assumptions that can be compared with those that humans make, such as:

- Whether Ω is a closed population
- Whether the probability of a novel state is influenced by the number of states already observed
- Whether the probability of a novel state is influenced by the distribution of probabilities across observed states (e.g., the number of singletons)

From the empirical literature: Human intuitions

Five findings regarding human intuitions about unknown Ω:

- Humans 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).
- They tend to anchor on the number of salient known states, and their probability assignments are influenced by this (partition dependence, similar to the principle of indifference).
- Partition dependence occurs even when subjective lower-upper probabilities are elicited.
- The greater the number of ways they think an outcome could occur, the higher the probability they assign to it (support theory).
- Humans find sample-space ignorance aversive.

Capture-recapture estimator experiments

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?

> *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?*

Capture-recapture estimator experiments

Table: Study 1 Population Estimation Task Responses

estimate	species		fish	
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%
estimate	freq.		percent	
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%

Capture-recapture estimator experiments

But can they still get the right answer if it isn't in a multi-choice list?

Table: Study 2 Population Estimation Task Responses

estimate	text-ent.		multi-ch.	
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%
estimate	newly capt.		recapt.	
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%

Probability of Novel State Experiments

Do people's estimates of the cardinality of an unknown Ω covary with the number of states observed thus far? What influence does their prior belief about Ω have?

> *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.*

Probability of Novel State Experiments

Do people's estimates of the cardinality of an unknown Ω covary with the number of states observed thus far? What influence does their prior familiarity with Ω have?

> *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.*

Probability of Novel State Experiments

Participants were given the marbles and the automobile scenarios in counter-balanced order. They also were randomly assigned to either have seen 4 colors or 15 colors in the 20 observations.

- Marbles scenario: 59% gave higher estimates if they saw 15 colors than if they saw 4 colors, and 38% did the opposite.
- Automobile scenario: 29% gave higher estimates if they saw 15 colors than if they saw 4 colors, and 67% did the opposite.
- Order of scenario presentation had no effect.

Prior beliefs about the distribution of automobile colors appears to have had a strong impact on the effect of number of observed states. A second study partly replicated these findings (eliminating the positive effect of 15 vs 4 colors on estimates of new colors in a future sample).

Findings so far

Main findings:

- Some people are capable of producing the Lincoln-Petersen estimate even without prompting.
- Expressing capture-recapture information as recaptures and percentages enhance this ability.
- Greater diversity in a sample from Ω induces either a Pitman-Yor-like heuristic or its opposite, depending on prior beliefs about the size of Ω.
- To my awareness, there is no counterpart in the literature to the Pitman-Yor or Chinese-Restaurant process models that predicts *lower* estimates of novel states with greater numbers of observed states.

Future directions

Some future directions:

- Investigate conditions and individual-difference variables influencing ability to estimate Ω cardinality
- Elicit prior beliefs about Ω and imprecise cardinality estimates in future experiments
- Develop testable cognitive models (e.g., Bayes learning) that incorporate a prior on Ω cardinality
- Investigate the effect of newly-observed states on probability assignments to old states (e.g., test the hypothesis that the ratios of these probabilities should remain unchanged)

The End

Thanks!