Sorites

1 Introduction

- sorites paradox exists
  - sorites paradox is ...
  - These statements are clearly true on their own, but when taken together, they logically imply something false.
  - it's confusing with a binary, logical semantics
- dan's adjective model shows the qualitative behavior
  - model the premises as speech acts, which depend on context, rather than logical statements that are universally true or false
  - we get a *probability* of endorsement, which gives us graded endorsements
  - depending on the prior distribution over values, we might get different endorsements for the same value (e.g. different objects have difference price distributions, so inductive for coffee maker as some dollar amount epsilong will get different endorsement than the same statement about laptops)
- do humans actually show the behavior predicted by this model?
  - we collect endorsements for sorites-style utterancs for different prices for a range of everyday objects
  - we also elicit prior distributions over prices for each of thos object categories.
  - finally, we translate the sorites premises into speech acts where a speaker has a goal of communicating some information to the listener
- the endorsement expt shows the qualitative pattern of the sorites paradox, and also we can quantitatively predict the endorsements given the separate experiment eliciting the price distributions.

2 Experiment 1: Sorites Statements

2.1 Participants

We recruited 30 participants with US IP addresses over Amazon's Mechanical Turk. 2 participants were exluded from analysis for not being native English speakers, leaving 28 participants for analysis. The experiment took about 7 minutes and participants were paid $0.70.

2.2 Materials

The experiment consisted of 50 questions. There were 2 basic question types, *concrete* and *inductive* of the form:

- *Concrete*: An [OBJECT] that costs $[PRICE VALUE] is expensive.
- *Inductive*: An [OBJECT] that costs $[PRICE VALUE] is expensive.

There were 5 object categories (coffee maker, laptop, headphones, watch, and sweater) and 5 price values for each object category and premise type. The price values based on pilot experiments to constitute similar standard deviations of the price distribution for each object category and to capture a range of plausible values.

Table 1: Price values for concrete premise.

qtype	coffee maker	headphones	laptop	sweater	watch
concrete	24.00	24.00	350.00	18.00	100.00

qtype	coffee maker	headphones	laptop	sweater	watch
concrete	52.00	60.00	600.00	36.00	250.00
concrete	84.00	96.00	900.00	57.00	450.00
concrete	124.00	144.00	1250.00	87.00	900.00
concrete	188.00	234.00	1850.00	150.00	2000.00

Table 2: Price values for inductive premise.

qtype	coffee maker	headphones	laptop	sweater	watch
inductive	2.00	2.58	18.50	1.71	24.00
inductive	20.00	25.80	185.00	17.10	240.00
inductive	100.00	129.00	925.00	85.50	1200.00
inductive	140.00	180.60	1295.00	119.70	1680.00
inductive	200.00	258.00	1850.00	171.00	2400.00

2.3 Methods

At the start of the experiment, participants were told they would be asked questions about the prices of different household items. Each question started with a statment in bold, either the *concrete* statement or the *inductive* statement. For both types of questions, participants gave Likert responses for how much they agreed with the statement, on a scale from "Completely disagree" (1) to "Completely agree" (9).

Each participant then saw all 5 object categories with all 5 price values for each kind of sentence. Trials were presented in random order.

2.4 Results

Results of Experiment 1 are shown in Figure 1. We see a range of endorsements from very low (e.g. "A coffee maker that costs $24.00 is expensive" or "A laptop that costs $1850.00 less than an expensive laptop is expensive.") to very high (e.g. "A watch that costs $2000.00 is expensive" or "A sweater that costs $171.00 less than an expensive sweater is expensive."). We also see a gradual increase in endorsements for the concrete premises (the item is expensive) as the price of the item increases, and a gradual decrease in endorsements for the inductive premise (the less expensive item is expensive) as the *change* in price between the expensive and less expensive items increases.

Participants tended to be in agreement about these endorsements. The average correlation between participants's responses was 0.544 (CI: [0.517, 0.569]). Aggreement tended to be higher for the concrete premise (mean: 0.649, CI: [0.622, 0.677]) than for the inductive premise (mean: 0.46, CI: [0.42, 0.498]).

2.5 Discussion

In this first experiment, we qualitatively capture the three characteristic effects of vagueness:

- Judgements depend on context: Different object categories and price values yield different responses.
- Judgements are systematic: Within a particular context, different people give similar judgements.
- Judgements are graded: They range from very low to very high endorsement and include borderline cases.

In particular, we see strong endorsement of some inductive premises. Importantly this is also graded, with some inductive sentences judged as not good at all.

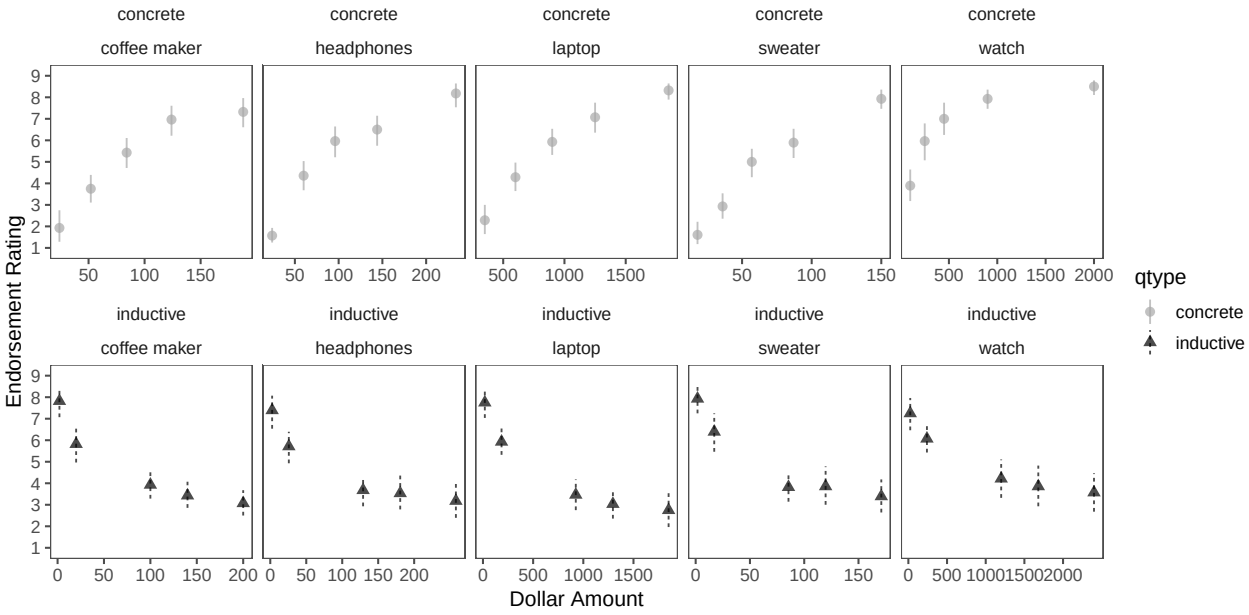


Figure 1: Results of Experiment 1. Error bars are 95% confidence intervals.

This suggests rich quantitative patterns to explain.

3 Experiment 2: Prior Distributions

Judgements to our sorites premises depended on the object categories. So, in order to model participants' endorsements, it was useful to have as input a measure of people's prior expectations about the prices of these object categories. In experiment 2, we elicit price distributions using a binned histogram approach (Franke et al. 2016).

3.1 Participants

We recruited 30 participants with US IP addresses over Amazon's Mechanical Turk. The experiment took about 16 minutes and participants were paid $1.60.

3.2 Materials

The experiment consisted of 5 question pages, on for each object category. Each page contained vertical slider bars corresponding to a range of prices (e.g. $0 - $50 or $450-$500). There were 50-80 sliders per page, depending on the object category (see Figure 2). The sliders were shown in rows of 10 sliders each.

There were 5 object categories (coffee maker, laptop, headphones, watch, and sweater). The price ranges for each object category were chosen based on pilot experiments. We wanted sufficient detail about the tails of the distributions, so we chose maximum values for each object category such that the average endorsement of the highest bin was very low. We also wanted sufficient granularity to address the sorites inductive premise, even for very small price values. We therefore chose the width of the bins so that, for every concrete price value $\$x$ and for every inductive price value $\$\varepsilon$ in our sorites premises experiment, we could confidently estimate the probability of an item $\$\epsilon$ less expensive than $\$x$. Our choise of maximum price and bin widths

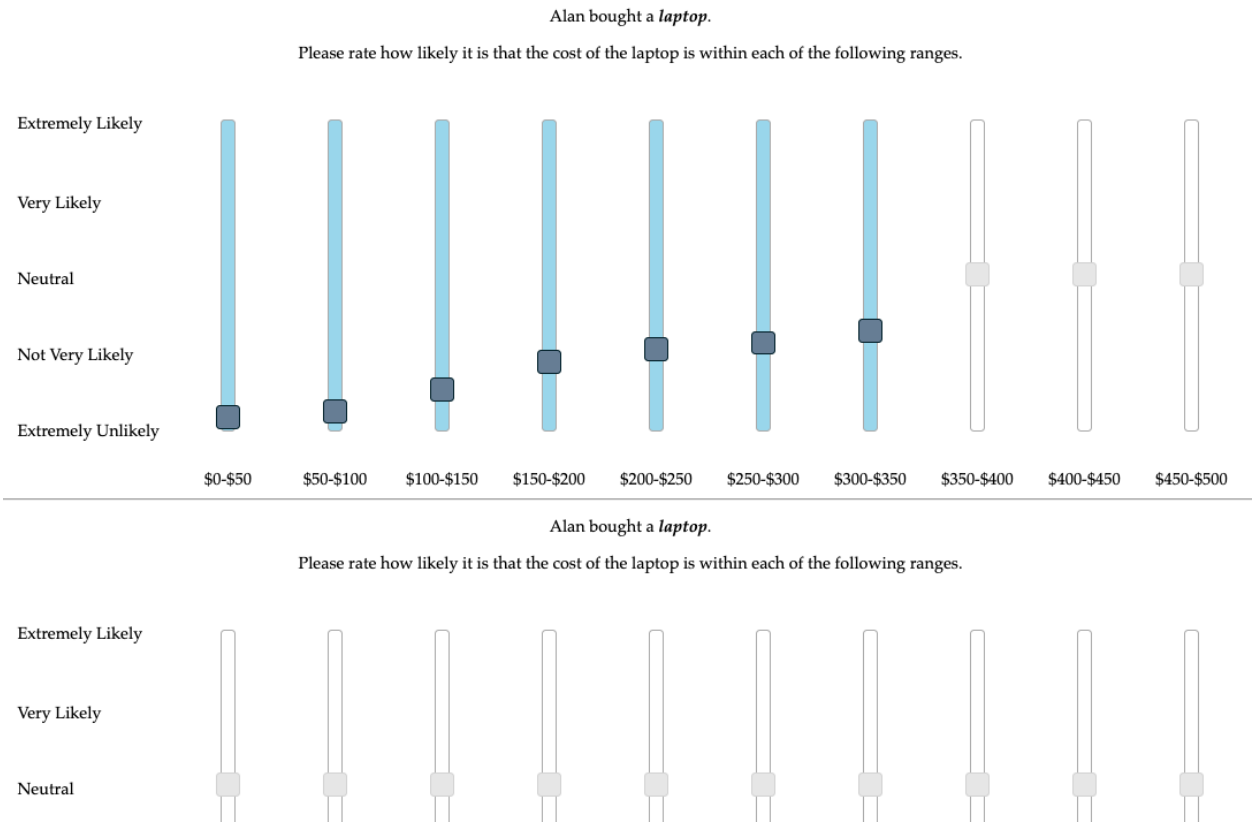


Figure 2: Screenshot from Experiment 2

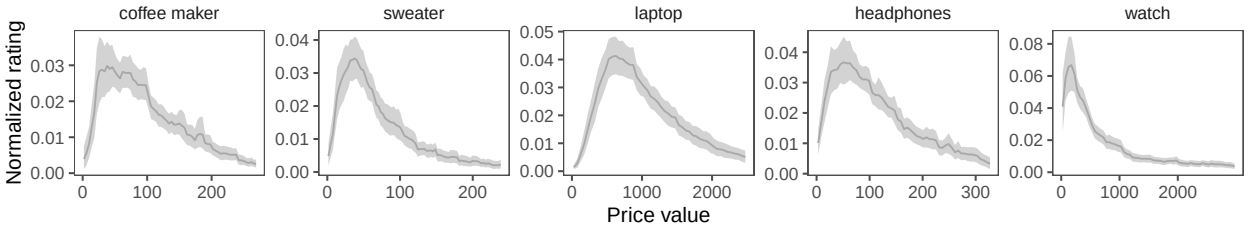


Figure 3: Experiment 2 results. Error bars are 95% confidence intervals.

are shown in Table 3. The resulting distributions are fairly smooth, allowing us to interpolate within the bins as needed. Our level of resolution also allowed us to capture detail in especially dense parts of the distributions (usually the smaller ranges).

Table 3: Price values for Experiment 2.

object category	max price	step size
watch	3000	50
laptop	2500	50
coffee maker	270	4
sweater	240	6
headphones	330	3

3.3 Methods

Each participant saw all 5 objects, which were presented in random order. For each object, participants saw the statement:

- [NAME] bought an [OBJECT]. Please rate how likely it is that the cost of the watch is within each of the following ranges.

Sliders corresponding to each price range were arranged in order from lowest price range to highest. Sliders were initialized in white, with a gray handle at the half-way mark. We required participants to drag the handle in order to register their response – from "Extremely unlikely" (0) to "Extremely Likely" (1). Once participants dragged the slider, the handle and slider changed color to blue. They submitted all of their responses for a given object at once.

3.4 Results

We normalize slider ratings within each participant and object category, since slider ratings likely reflect relative rather than absolute probabilities (Franke et al. 2016).

Resuts of Experiment 2 are shown in Figure 3. As stated earlier, we chose our price ranges so that we would get very low endorsements for the higher-price bins and relatively smooth curves to the distributions, which we do in fact see in participant responses.

Participants tended to be in agreement. The average correlation between participants' responses was 0.4209535 (CI: [0.400662, 0.4409868]), and correlations were similarly high within each object category.

Normalized slider ratings are very well-fit by log-normal distributions (see Appendix A for inference details). The empirical CDF is highly correlated with its best-fit log-normal CDF ($R^2 = 0.997$).

4 Model

4.1 Background

We base our model of the sorites premise endorsements on Lassiter (2015), which uses a scalar adjective model (Lassiter and Goodman 2013) to get a graded endorsement value for sorites. This model of scalar adjectives is a Rational Speech Act model (Frank and Goodman 2012; Goodman and Stuhlmüller 2013), which makes use of pragmatic principles (Clark 1996; Grice 1975; Levinson 1995) to resolve vague language in context. In this model, the literal semantics of a scalar adjective is relative to some unspecified threshold.

$$[[\text{X is expensive}]]_\theta = \text{price}(X) > \theta$$

The literal listener jointly infers the threshold θ and the value along the scalar adjective dimension, in our case, the price x of the "expensive" item X.

$$L_0(x, \theta|\text{X is expensive}) \propto P(x)P(\theta) \cdot \delta_{x>\theta}$$

Given a particular price value, the speaker chooses whether to endorse (utterance u="yes") or deny (utterance u="no") the statement "X is expensive" by soft-maximizing their utility of balancing informativity (the literal listener's ability to infer the correct price, marginalized over their interpretation of θ) and the cost of the utterance.

$$U_S(u; x) = \log\left(\int_\theta L_0(x, \theta|u)d\theta\right) - \text{cost}(u)$$

The main task in Experiment 1 was to choose how much to endorse a statement. We model this task as a speaker's choice between two alternative utterances: producing the given utterance to a naive listener, or staying silent (Degen and Goodman 2014; Franke 2014).

4.2 Concrete premise

The concrete premise in our experiment ("An [OBJECT] that costs $[PRICE VALUE] is expensive") contains a relative clause. However, we model the endorsement task as choosing between uttering the main clause "The [OBJECT] is expensive," or staying silent, given that the content of the relative clause ("The [OBJECT] costs $[PRICE VALUE]") is true. Hence, the concrete premise is exactly the scalar adjective utterance of Lassiter and Goodman (2013)'s model.

{{SHOW COR BETWEEN BOTH PHRASINGS?}}

4.3 Inductive premise

The meaning of the inductive premise is less straightforward. When someone states that "A watch that costs $3 less than an expensive watch is expensive," what information is assumed to be in common ground, and what information is the speaker trying to communicate? Since the price of the "expensive watch" is unknown to participants, how would they know whether or not to say that a less expensive watch is "still expensive"?

Rather than model the sorites inductive premise as a speaker endorsement task, Lassiter (2015) look at the *listener*'s joint posterior distribution over the price of the object x and the threshold θ given the utterance "an expensive [OBJECT]" and directly compute the marginal probability that $x - \varepsilon > \theta$ is true.

We extend Lassiter (2015)'s approach slightly and compute instead a speaker's probability of communicating that a less expensive item is "still expensive" given a listener's joint posterior over x and θ. That is, we assume participants take the following statements as given:

- An [OBJECT] is expensive.
- Another [OBJECT] costs $[PRICE VALUE] less.

Similar to Lassiter (2015), we use a listener model to simultaneously compute the threshold θ and the price x_1 of the "expensive" object (and consequently the price x_2 of the "less expensive" object).

$$L_0(x_1, \theta | X_1 \text{ is expensive})$$
$$x_2 = x_1 - \varepsilon$$

However, in our model of the inductive premise, the listener then becomes a speaker. The speaker has the goal to communicate the less expensive price x_2 to a naive listener (who knows the threshold but not the price), and must choose whether to make the following statement or stay silent:

- [THE LESS EXPENSIVE OBJECT] is expensive.

We therefore model this naive listener's inferred price distribution with vs. without the statement that the less expensive item would be "still expensive":

$$\begin{aligned} L_0(x_2 | X_2 \text{ is expensive}, \theta) &= \frac{Pr(x_2)\delta_{x_2 > \theta}}{\int_\theta Pr(x)\delta_{x>\theta} d\theta} \\ &= \left\{ \begin{array}{ll} \frac{Pr(x_2)}{1-CDF(\theta)} & \text{if } \; x_2 > \theta \\ 0 & \text{otherwise} \end{array} \right\} \end{aligned}$$

$$L_0(x_2 | \text{silent}, \theta) = Pr(x_2)$$

If the cost of the "still expensive" utternace is $-\frac{1}{\lambda}\log(C_X)$ and the cost of the staying silent is $-\frac{1}{\lambda}\log(C_S)$, the speaker endorsements are:

$$S_1(u | x_2, \theta) = \frac{\exp(\lambda\left(\ln(L_0(x_2|u,\theta)) - c(u)\right))}{\sum_{u'} \exp(\lambda\left(\ln(L_0(x_2|u',\theta)) - c(u')\right))} = \frac{C_u \left(L_0(x_2|u,\theta)\right)^\lambda}{\sum_{u'} C_{u'} \left(L_0(x_2|u',\theta)\right)^\lambda}$$

So for the "still expensive" utterance:

$$\begin{aligned} S_1(X_2 \text{ is expensive} | x_2, \theta) &= \frac{C_X \left(L_0(x_2 | X_2 \text{ is expensive}, \theta)\right)^\lambda}{C_X \left(L_0(x_2 | X_2 \text{ is expensive}, \theta)\right)^\lambda + C_S \left(L_0(x_2 | \text{silent}, \theta)\right)^\lambda} \\ &= \left\{ \begin{array}{ll} \frac{C_X \left(\frac{Pr(x_2)}{1-CDF(\theta)}\right)^\lambda}{C_X \left(\frac{Pr(x_2)}{1-CDF(\theta)}\right)^\lambda + C_S (Pr(x_2))^\lambda} & \text{if } \; x_2 > \theta \\ 0 & \text{otherwise} \end{array} \right\} \\ &= \left\{ \begin{array}{ll} \frac{C_X}{C_X + C_S (1-CDF(\theta))^\lambda} & \text{if } \; x_2 > \theta \\ 0 & \text{otherwise} \end{array} \right\} \end{aligned}$$

Finally, we average over participant's initial guess of x_1 and θ to get the expected endorsement.

$$S_1(\text{still expensive}) = \int_{x,\theta} L_1(x_1, \theta | \text{x is expensive}) \left\{ \begin{array}{ll} \frac{C_X}{C_X + C_S (1-CDF(\theta+\varepsilon))^\lambda} & \text{if } \; x_1 - \varepsilon > \theta \\ 0 & \text{otherwise.} \end{array} \right\} dx d\theta$$

If the costs of the two utterances are equal, then the $S_1(\text{still expensive} | x, \theta)$ value is always between 1/2 and 1 whenever $x_1 - \varepsilon > \theta$. In this case, the expected endorsement is very similar to simply oberving the joint distribution $L_0(x_1, \theta | X_1 \text{ is expensive})$ and taking the expectation of $x_1 - \varepsilon > \theta$. But the smaller θ is, the less informative the "still expensive" utterance would be, and so the less likely it is that a speaker would actually endorse it over staying silent.

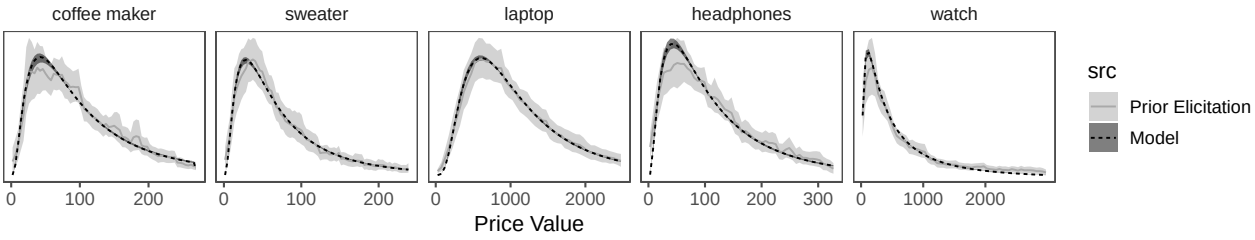


Figure 4: Posterior price distributions fit to Experiments 1 and 2. Error bars are 95% confidence intervals.

4.4 Analysis

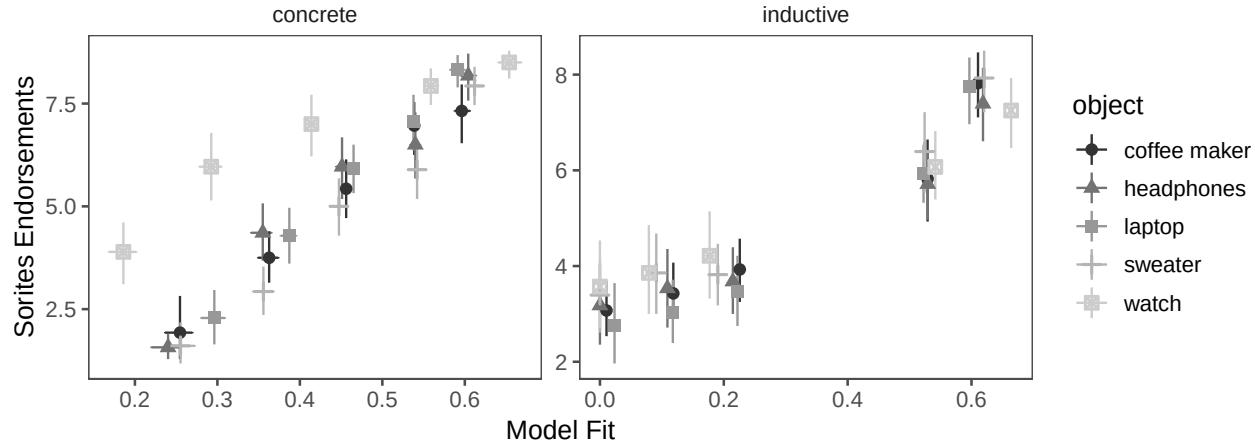
Given the prior distributions over prices for object categories (elicited in Experiment 2), to model responses in Experiment 1, we need only 2 parameters: speaker optimality λ and the cost of the utterances. For simplicity, we assume that both utterances have equal cost in this task (similar to giving a simple "yes" or "no" response), leaving the single parameter for speaker optimality. We use Bayesian methods to jointly model Experiments 1 and 2, simultaneously inferring latent parameters of the object price distributions and speaker optimality. Averaging over the posterior distribution of latent parameters, we show how much variance in the endorsement data we can explain by the pragmatic model.

We model the prior distribution over prices as log-normal, with each object category o having its own mean μ_o and standard deviation σ_o. We model slider responses for each bin in Experiment 2 as Gaussian, centered at the probability of that bin with standard deviation σ_{bin}.[1] We put uninformative priors over the price distribution parameters $\mu_o \sim \text{Uniform}(0, 10), \sigma_o \sim \text{Uniform}(0, 10)$, the response parameter $\sigma_{\text{bin}} \sim \text{Uniform}(0, 0.5)$, and the speaker model parameter $\lambda \sim \text{Uniform}(0, 10)$.

Details of how we carried out this Bayesian inference are described in Appendix A.

4.5 Results

The posterior price distributions inferred by fitting to both Experiments 1 and 2 (shown in Figure 4) are very close to the normalized ratings from Experiment 1 (comparing empirical vs. model CDF, $R^2 = 0.995$).



The posterior endorsement levels (expected probability of choosing the utterance) for Experiment 2 as well as participants actual responses are shown in Figures 5 and 6. The correlation between participants' mean responses and the model's mean endorsement is very high ($R^2 = 0.906$ for the inductive premise and

[1] This is a deviation from Franke et al. (2016)'s analysis of binned histogram data which used a logit transform in the linking function from probability to slider ratings.

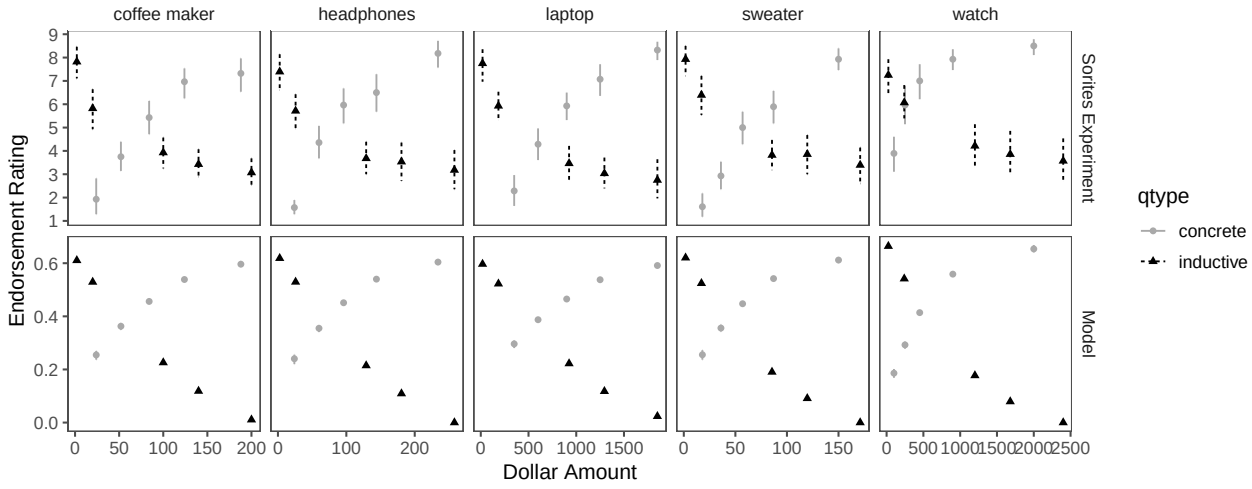


Figure 5: Results of Experiment 1. Error bars are 95% confidence intervals.

$R^2 = 0.788$ for the concrete premise). These values are close to the split-half correlation (for inductive, mean: 0.897, CI: [0.826, 0.95]; for concrete, mean: 0.939, CI: [0.886, 0.972]), which represents a ceiling on explainability. For the "watches" object category, we find that participants' endorsements of the concrete premise "A watch that costs $[PRICE VALUE] is expensive," are noticeably higher than the model's.

There are 11 parameters necessary for modeling Experiment 2 (5 mean and 5 variance parameters for the object categories and one variance parameter for the slider responses) and only one additional parameter necessary for modeling Experiment 1 (the speaker optimality parameter). Since the fit for Experiment 1 is our primary interest, we want to check how much the Experiment 2 parameters are being changed to fit Experiment 1's data. We therefore set the optimality parameter to the the maximum a posteriori estimate under the full Bayesian data analysis ($\lambda = 0.8$), and directly compute responses to Experiment 1, fitting the object category price prior distribution parameters to Experiment 2's data only. This gives similar correlations with the empirical responses in Experiment 1 ($R^2 = 0.914$ for inductive and $R^2 = 0.778$ for concrete). This indicates that the model is not overfitting.

5 Discussion

We show a close fit to the empirical responses, except in the case of watches, where participants' endorsements are higher than the model's endorsements for the statement that a watch of a particular price is expensive. This may have to do with the long tail in the distribution of watch prices. Luxury watches can be extremely expensive relative to regular watches, and might be thought of as categorically different from typical watches. When primed with a range of prices that are very high, as in our priors experiment, participants might think more about high-end watches, but when considering whether a watch of a particular price is expensive, the contrast class might be more typical watches.

Given the within-subjecs design of each experiment, there could be a concern about task demands within each experiment. The fact that people had to answer many similar questions with different prices may have encouraged participants to use the full range of the scale even if their natural responses tended more towards the endpoints, resulting in the qualitative finding of graded judgements for the sorites premises. However, Experiments 1 and 2 involved a completely separate set of participants, but using the prior data from Experiment 2 and a single optimality parameter was sufficient to quantiatively model the data from Experiment 1. The fact that this model captures the relationship between the prior distributions and the sorites judgements is not explained by task demands.

This model of the sorites premises as speech acts predicts the paradox that participants endorse both premises at certain values of the price (in the concrete premise) and price difference (in the inductive premise). We show that the model also quantiatively predicts the graded judgements that people actually give for the premises as we vary those values.

Using a sensible choice for the speech act for each sorites premise and uninformative priors over hyperparameters, we are able to use the data from Experiment 2 to fit the prior price distributions.

A Inference details

The code for our model and procedures and data from our experiments can be found at {{URL}}. We implemented our model in the probabilistic programming language WebPPL (Goodman and Stuhlmüller 2014). We used an incrementalized version of the Metropolis- Hastings algorithm (Ritchie, Stuhlmüller, and Goodman 2016). We discarded the first 5000 samples for burn in. After burn in, we kept every 10th sample. In total, we kept 10000 samples to represent the posterior distribution.

We set uninformative priors over the hyperparameters for the price distribution parameters $\mu_o \sim \text{Uniform}(0, 10), \sigma_o \sim \text{Uniform}(0, 10)$, the response parameter $\sigma_{\text{bin}} \sim \text{Uniform}(0, 0.5)$, and the speaker model parameter $\lambda \sim \text{Uniform}(0, 10)$.

A.0.1 Discretization

Since inference over the speaker model has a nested inference of the literal listener distribution, we used a simplification. For each continuous sample of the set of prior parameters μ_o and σ_o, we discretized the domain of these priors and computed exact probability distributions for the prior on prices.

We chose a discretization that would have few bins, but would also closely resemble the behavior of the true distributions. Since the range of prices is so wide, and since some prices are very close together, it would be difficult to create a discretization that works for every dollar amount used across all experiments. Because of this, we implemented a separate discretization for each experiment.

For discretizing the prior, we created a set of bins (not necessarily of equal width) such that each price in the experiment falls into a unique bin. We also want there to be bins below the lowest price's bin and above the highest price's bin. The probability of a price being sampled from each bin is computed from the CDF: `p_bin = pnorm(upper_boundary) - pnorm(lower_boundary)`.

The distribution of the threshold θ was also discretized and represented relative to the prior bins. We assumed the true distribution of θ to be independent of x and uniform, with maximum value at 10% higher than the largest dollar amount used for that item in the experiment. Since the bins of of different widths, the probability of θ falling between prices from adjacent bins varied, depending on the bins. We used a simple scaling based on the width of the bins. This is an approximation to the true joint distribution over prices x and thresholds θ.

The discretization used for the full set of sorites experiments is shown in Figure 7. The height of the bars represent the probability that theta will fall between each pair of midpoints. The midpoints of the bins appear in black. Bins are shown divided by vertical lines. The representative theta values, which appear at the boundaries of the bins, are shown in grey.

References

Clark, Herbert H. 1996. *Using Language*. Cambridge University Press.

Degen, Judith, and Noah D. Goodman. 2014. "Lost Your Marbles? The Puzzle of Dependent Measures in Experimental Pragmatics." In *CogSci*.

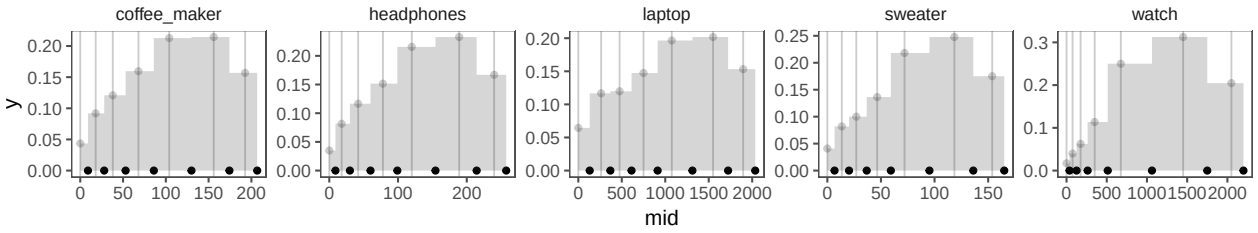


Figure 6: Discretization and marginal distribution over the threshold variable $ heta$ for each object category.

Frank, Michael C., and Noah D. Goodman. 2012. "Predicting Pragmatic Reasoning in Language Games." *Science* 336 (6084): 998–98. https://doi.org/10.1126/science.1218633.

Franke, Michael. 2014. "Typical Use of Quantifiers: A Probabilistic Speaker Model." In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36.

Franke, Michael, Fabian Dablander, Anthea Schöller, Erin Bennett, Judith Degen, Michael Henry Tessler, Justine T. Kao, and Noah D. Goodman. 2016. "What Does the Crowd Believe? A Hierarchical Approach to Estimating Subjective Beliefs from Empirical Data." In *CogSci*.

Goodman, Noah D., and Andreas Stuhlmüller. 2013. "Knowledge and Implicature: Modeling Language Understanding as Social Cognition." *Topics in Cognitive Science* 5 (1): 173–84. https://doi.org/10.1111/tops.12007.

Goodman, Noah D, and Andreas Stuhlmüller. 2014. "The Design and Implementation of Probabilistic Programming Languages." http://dippl.org/.

Grice, H. Paul. 1975. "Logic and Conversation." In *Syntax & Semantics*, edited by P. Cole and J. Morgan. Vol. 3.

Lassiter, Daniel. 2015. "Adjectival Modification and Gradation." *Handbook of Contemporary Semantic Theory*, 143–67.

Lassiter, Daniel, and Noah D. Goodman. 2013. "Context, Scale Structure, and Statistics in the Interpretation of Positive-Form Adjectives." In *Semantics and Linguistic Theory*, 23:587–610.

Levinson, Stephen C. 1995. "Interactional Biases in Human Thinking." In *Social Intelligence and Interaction*, 221–60. Cambridge University Press.

Ritchie, Daniel, Andreas Stuhlmüller, and Noah D. Goodman. 2016. "C3: Lightweight Incrementalized MCMC for Probabilistic Programs Using Continuations and Callsite Caching." In *AISTATS 2016*.