Sorites

Contents

Description of experiments

Prior elicitations

After many pilot experiments, we found a set of lower and upper bounds for the bins in a binned histogram elicitation experiment. It took a lot of tries, because we needed detailed, accurated information about the shape of the distributions in the *tails* of the distribution. The sorites premises condition on expensive items, and so we need to know about the upper tails of the distributions. The tails of distributions are low probability, though, and so it's difficult to get the information from people.

For each item, we choose a maximum price to ask about, and a step-size – the width of the bins.

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

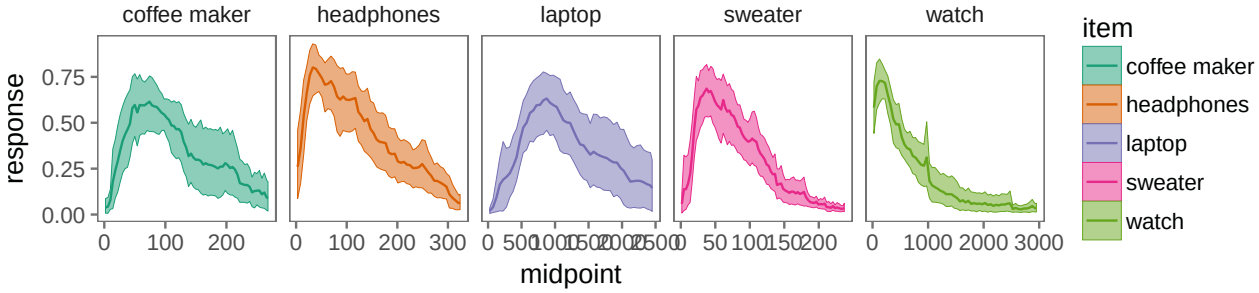
For each of these prior elcitation experiments, we asked participants to move sliders to represent their estimate of the likelihood that the prices of various items would fall within the given ranges.

"Experiment 6": 5 items, 10 Ss

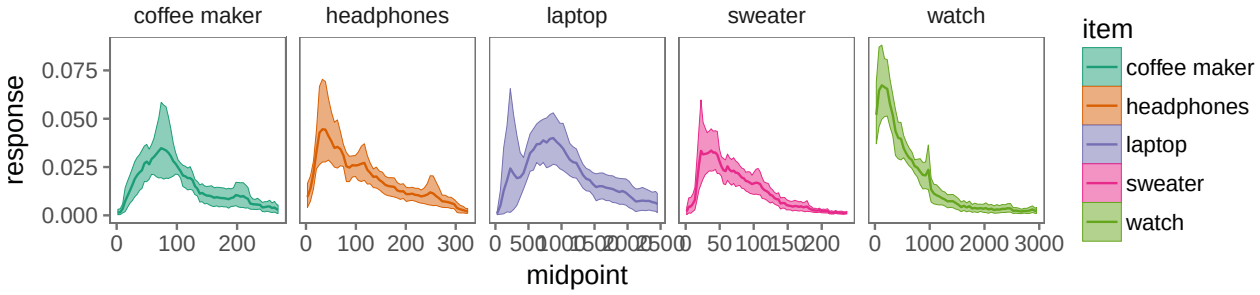
A copy of the experiment can be found at: `experiments/experiment6-conditionA/morebins.html`.

There were 10 participants.

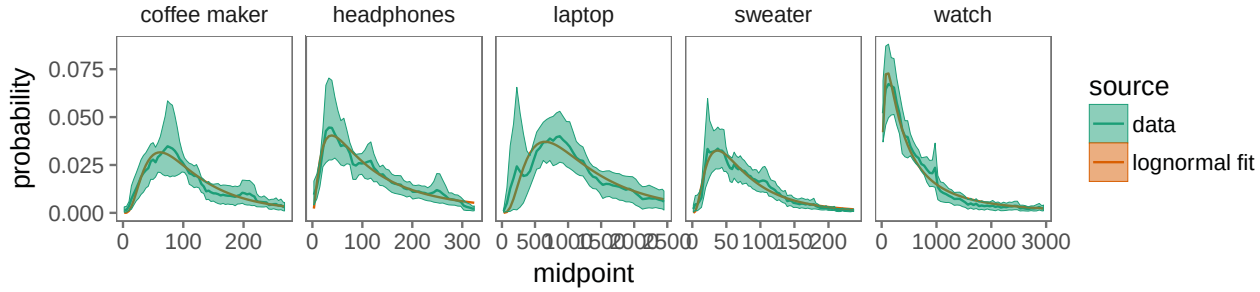
Here are the average responses, without rescaling each participants' responses:



Because we are trying to get a probability distribution from each participant, we rescale so that each participant's responses for a particular item sum to one.



These rescaled and then averaged values, along with the midpoint of the bins, is what we use to fit the parameters of a lognormal distribution.

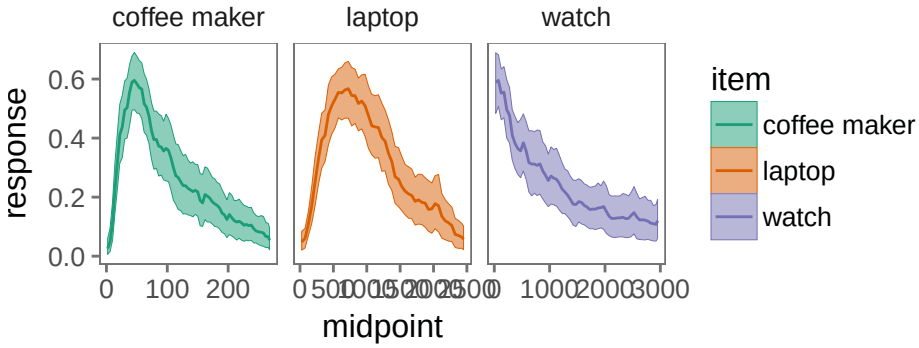


"Experiment 9": 3 items, 36 Ss

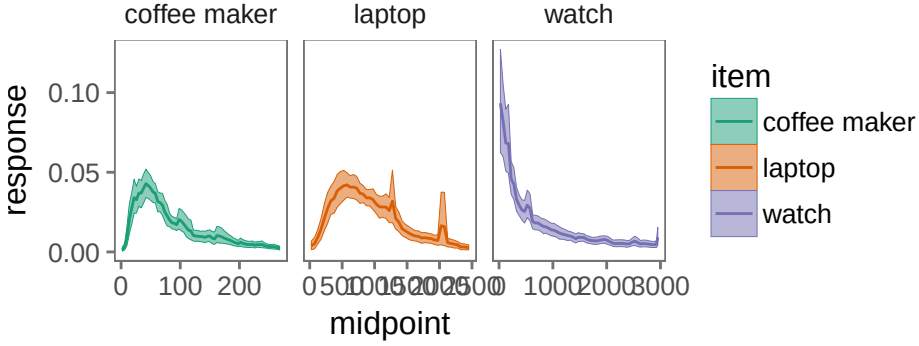
A copy of the experiment can be found at: `experiments/experiment9/morebins.html`.

There were 36 participants.

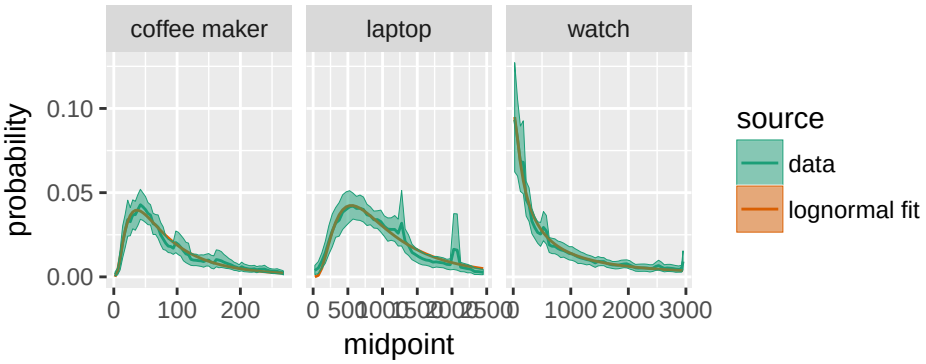
Here are the average responses, without rescaling each participants' responses:



Because we are trying to get a probability distribution from each participant, we rescale so that each participant's responses for a particular item sum to one.

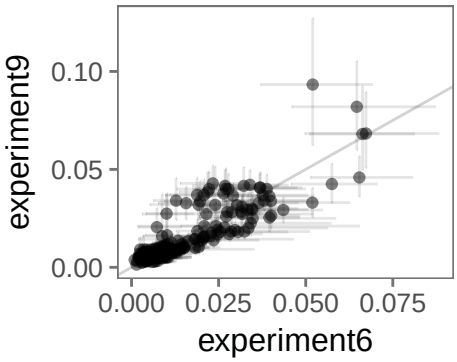


These rescaled and then averaged values, along with the midpoint of the bins, is what we use to fit the parameters of a lognormal distribution.



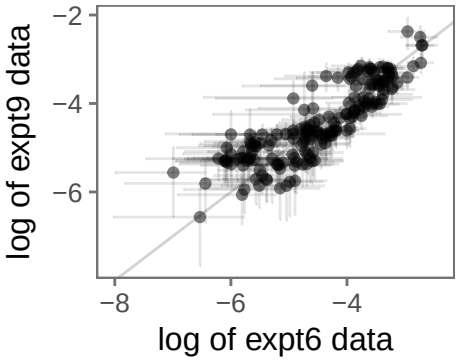
Comparison

Looking at just the items that were in both experiments, we can check if we got similar results.



The correlation of the probabilities derived from these experiments is r=0.863572.

Converting to log-space (so we can see things at a smaller scale a little more clearly), we get the following plot:



The correlation of the scores (log(probability)) derived from these experiments is r=0.8617967.

I think these are close enough, so I'm presenting an aggregate of the data from both experiments. Note that the three items that were in Experiment 9 have *much* more data than the other two.

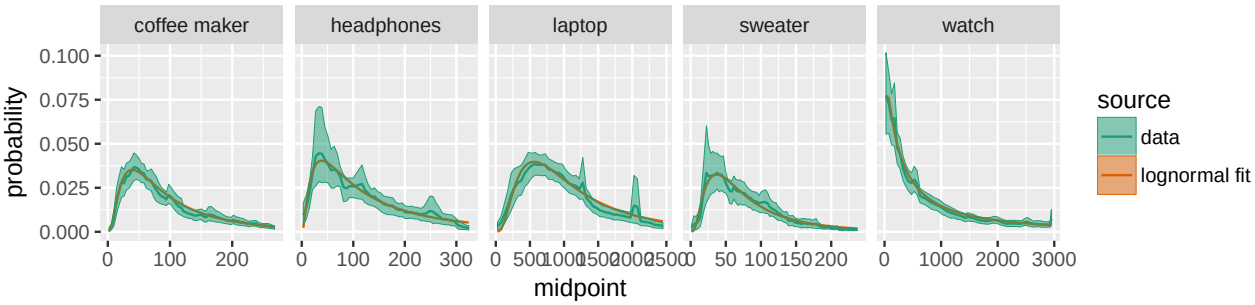
Aggregate

Here's the aggregate data and the fit lognormal parameters:

```
## <environment: namespace:tidyr>

##           item  meanlog     sdlog
## 1 coffee maker 3.672489 0.8664941
## 2   headphones 3.643866 1.0579738
## 3       laptop 6.327010 0.7574419
## 4      sweater 3.607647 0.7777513
## 5        watch 3.705547 1.7344699
```

And here are the fit lognormal curves plotted against the aggregated normed responses from Experiments 6 and 9.



These aggregate priors are used for the model predictions below.

Sorites experiments

In the sorites experiment, we had two types of utterances that we asked participants to endorse:

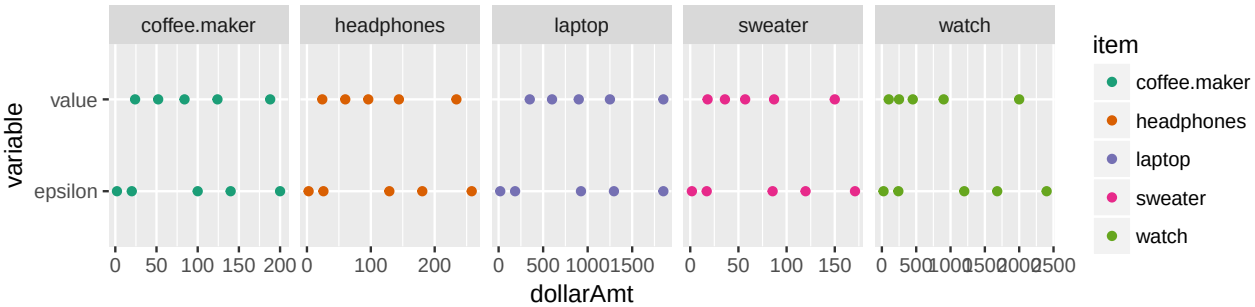
- Concrete premise: "A watch that costs $V is expensive" where V (value) can a range of values.
- Inductive premise: "A watch [that costs $E less than an expensive watch] is expensive" where E (epsilon) can take a range of values. (The phrasing of the inductive premise varied across two conditions.)

Similar to the priors experiments, it took a lot of pilot experiments to figure out which ranges resulted in varied judgements for these two premises. We wanted some values of E to result in low endorsement of the inductive premise and some to result in high endorsement.

In choosing the values of V and E to use in the sorites experiment, we used the means and standard deviations for each item from Experiment 6.

- **Values V** We chose the dollar amount values of the items to be 0, 1, 2, 3, and 4 standard deviations above the mean prices of the given item category.
- **Epsilons E** We chose the dollar amount for the "difference" E in the inductive premise to be 0.01, 0.1, 0.5, 1, 2, and 3 multiples of a standard deviation for the price of the given item category.

Given this algorithm, we settled on the following ranges (though I made a slight calculation error, and so this algorithm only approximately produces the values we actual used in the sorites experiment):



For each of the 5 items, there were 5 possible concrete premise sentences and 5 possible inductive premise sentences. So each participant rating 50 sentences.

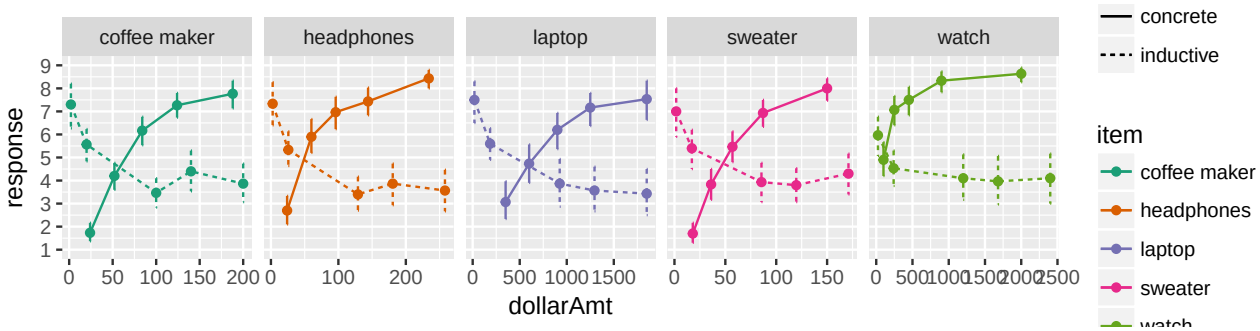
For each of these experiments (really the same experiment with 2 conditions), the concrete premise was of the form "A laptop that costs $V is expensive,", where V could take any of the values shown above.

The inductive premise varied between "If a laptop is expensive, then another laptop that costs $E less is also expensive," in the "Conditional" version (Experiment 10), and "A laptop that costs $E less than an expensive laptop is also expensive," in the "Relative clause" version (Experiment 11), where E could be any of the epsilons shown above.

Conditional

A copy of the experiment can be found at: `experiments/experiment10/exp3-sorites.html`.

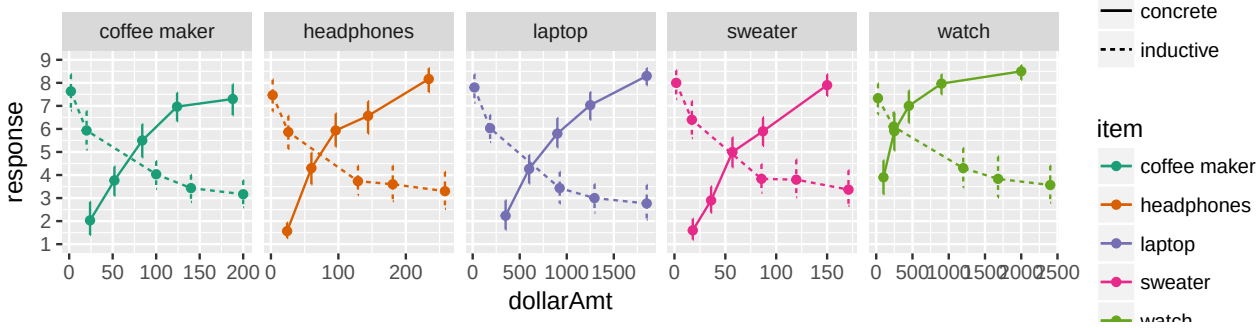
The inductive premise was of the form: "If a laptop is expensive, then another laptop that costs $E less is also expensive."



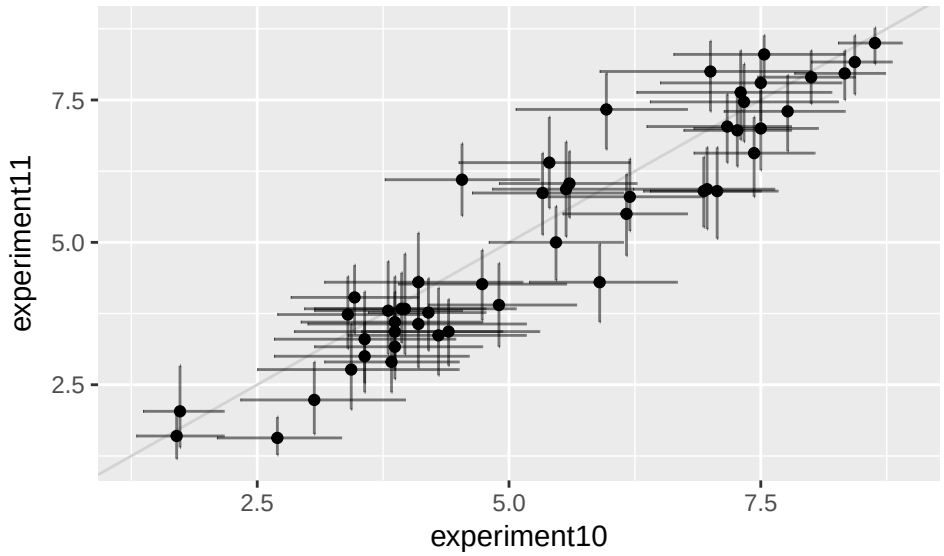
Relative clause

A copy of the experiment can be found at: `experiments/experiment11/exp4-sorites.html`.

The inductive premise was of the form "A laptop that costs $E less than an expensive laptop is also expensive."

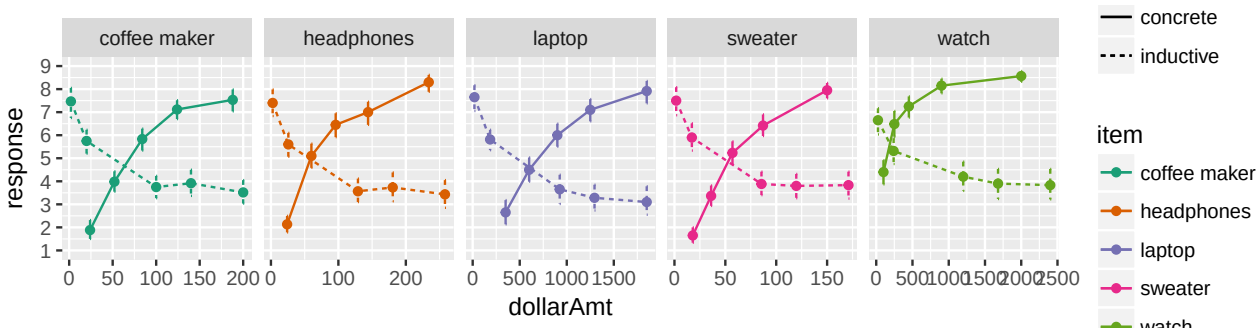


Comparison



The correlation between the endorsements in the two experiments was r=0.941109. I aggregate the data for comparing to the model comparison. This resulted in data from 60 total participants.

Here's the aggregate data plotted by itself:



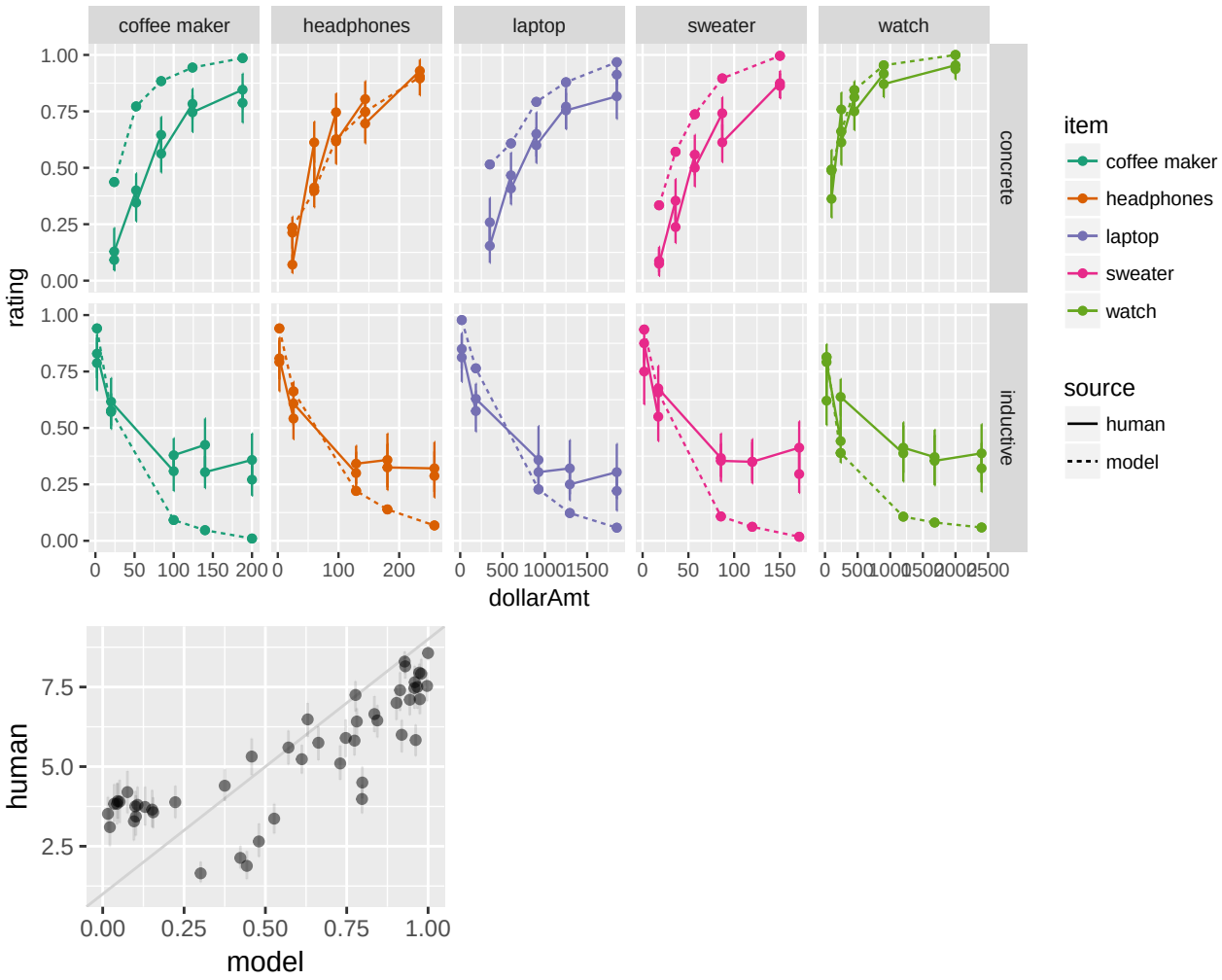
Model

For the model, we just used RSA to compute the inferred value of an item if a speaker said it was "expensive" (rather than staying silent) and the jointly inferred value for the lifted threshold variable θ. The adjective utterance "expensive" was more costly than staying silent, and we soft-minimized cost.

Once we have the cost of the expensive item and the threshold for the adjective "expensive", we compute the following.

- For the concrete premise, we simple check if the value V is greater than the inferred threshold θ.
- For the inductive premise, we check if the inferred value X of an expensive item is more than epsilon (E) greater than the inferred threshold θ, i.e. X-E>θ.

There were two parameters to this model: the cost of the expensive utterance (staying silent had cost=0) and the rationality of the speaker. We set both parameters to 1 as a baseine, but we get similar results when alpha=5 and cost=6 (which is a good fit in other experiments). Technically, we should fit these parameters to the data and infer their values, but I haven't done that yet.


coffee maker
headphones
laptop
sweater
watch
concrete
inductive
rating
dollarAmt
item
coffee maker
headphones
laptop
sweater
watch
source
human
model

human
model