# What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data

**Michael Franke (mchfranke@gmail.com)**
Department of Linguistics, Wilhelmstraße 19
72074 Tübingen, Germany

**Judith Degen (judith.degen@gmail.com)**
ladida
blubb, USA

**Michael Henry Tessler (XXX@XXX)**
ladida
blubb, USA

**Noah D. Goodman (XXX@XXX)**
ladida
blubb, USA

**Who else? (XXX@XXX)**
ladida
blubb, XXX

Each author's name should appear on a separate line, 11 point bold, and centered, with the author's email address in parentheses. Under each author's name list the author's affiliation and postal address in ordinary 10 point type.

**Abstract**

Subjects' beliefs about everyday events are of theoretical interest on their own and also an important ingredient in especially Bayesian models of such diverse phenomena as logical reasoning, future predictions or language use and interpretation. Here, we scrutinize one recently popular method for measuring subjective beliefs experimentally. We present a hierarchical Bayesian model for inferring likely "population-level beliefs" as the central tendency of subjects' individual-level beliefs. Individual-level beliefs define likelihood functions for three types of task measures. Our results suggest that the previous practice of using averaged normalized slider ratings for binned quantities is a practical and fairly good approximator of latently inferred population-level beliefs.

**Keywords:** subjective beliefs, hierarchical modeling, Bayesian data analysis, Bayesian cognitive models

## Motivation

We cannot look into a person's head. But in making sense of observed behavior we readily ascribe beliefs and desires to fellow agents. This happens intuitively, in folk psychology, but also in science. Ascriptions of latent mental states play a big role in many explanations of especially higher-order cognition, like decision making, planning, reasoning, or language use. Naturally, then, it becomes important how to validate any putative ascription of mental states for explanatory purposes.

A family of models where this is particularly pressing are Bayesian models of cognition which seek to explain task behavior in a variety of domains as partially informed by what subjects believe about mundane events. Take interpretation of language. Empirical data on whether a statement like "That watch cost $n$ dollars" is understood to convey speaker affect, rather than literal meaning, can be explained well by a Bayesian model of utterance interpretation (Kao, Wu, Bergen, & Goodman, 2014) in which a crucial role is played, as is quite intuitive, by an empirical measure of subjects' expectations about the likely or normal prices of a watch. Other examples of domains in which empirically successful models have included some measure of subjects' belief include making future predictions (Griffiths & Tenenbaum, 2006), the strength of pragmatic enrichments (Degen, Tessler, & Goodman, 2015), the interpretation of vague quantifiers (Schöller & Franke, 2015), . . . FILL ME . . .

Many methods of assessing subjective beliefs exist. One approach is to take actual frequencies as an approximation (Griffiths & Tenenbaum, 2006). This, however, does not work for beliefs about one-shot or imaginary events. Moreover, even where available, real-world frequencies may deviate systematically from subjects' beliefs, and this can be problematic for the predictions of a model that relies on subjective beliefs (Marcus & Davis, 2013).

Another approach is to try to experimentally measure relevant beliefs. Many techniques for this exist, especially in the economics literature (Morgan & Henrion, 1990; Manski, 2004; Schlag, Tremewan, & van der Weele, online first; Andersen, Fountain, Harrison, & Rutström, 2014). The perhaps most prominent method uses so-called *scoring rules* (Savage, 1971; Schlag et al., online first). Roughly speaking, scoring rules are particular schemes of rewarding participants' answers in such a way as to make sure that responses are what we want them to be: the actual subjective probability of an event, the mean of a distribution, a credible interval etc. Scoring rules are highly faithful but require participants to become sufficiently familiar with the (usually probabilistic) payoff scheme. Scoring rules also must make assumptions about how much participants value certain probabilistic gambles (Andersen et al., 2014).

For common applications, many of these procedures may be excessively complex or too demanding on both researcher and participant. It would be beneficial if there was an easily applicable technique of eliciting subjective beliefs that is good enough for testing the general predictions of Bayesian models. A further complication is that it may be impractical to obtain, from the same participant, data on subjective beliefs and data on the task of interest, especially if exposure to one of these measures is likely to affect the other. This suggests that, for practical reasons, there is a demand of reliable-enough measures of what we will sloppily call "population-level beliefs": estimates of the central tendency or average of subjective beliefs in a given population.

A very simple technique that has been used with apparent success is to have subjects adjust sliders in order to express their (relative) levels of support for values of uncertain contiguous quantities.

- explain slider rating task
- our goal here is to validate whether "population-level beliefs" are approximated well enough by normalized average slider ratings
- use hierarchical Bayesian modeling
  - calculate likelihoods from subjective beliefs
  - each subjective belief is drawn from a population-level central tendency, which is a "belief" as well

## Experiments

- why within-participant?
- what was the exact randomization / procedure?
- include all items?
- include bins as well?
- what were the lighting round bin comparisons?
  - 1 vs 2
  - 2 vs 6
  - 6 vs 11
  - 11 vs 14
  - 14 vs 15

## Model

The data we would like to explain are: (i) the normalized slider ratings $s_{ijk} \in [0;1]$ of subject $i \in \{1,\dots,20\}$ for item $j \in \{1,\dots,8\}$ and bin $k \in \{1,\dots,15\}$; (ii) the bins $n_{ij} \in \{1,\dots,15\}$ in which subject $i$'s number choice for item $i$ was;

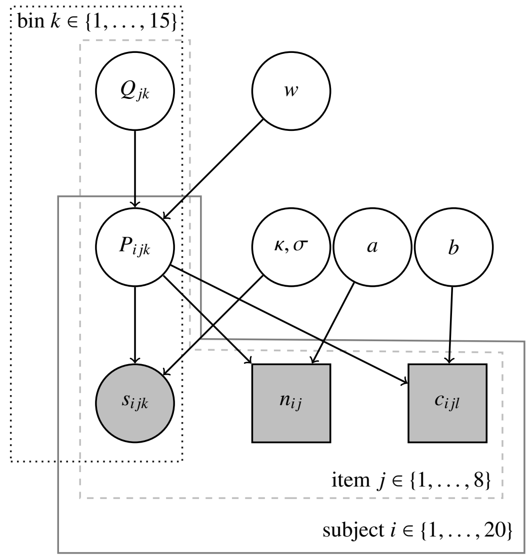



Figure 1: The data-generating model as a probabilistic graphical model, following conventions of Lee and Wagenmakers (2015). Shaded nodes are observed, white nodes are latent variables. Square nodes represent categorical, round nodes continuous variables. Boxes indicate scope of indices.

and (iii) the binary choices $c_{ijl} \in \{0,1\}$ of whether subject $i$ selected the higher bin for item $j$ in the bin comparison condition $l$. There two simplifications in need of commenting. In (i), we focus on slider ratings after normalizing for each subject, because we assume that slider adjustments reflect relative, not absolute estimates of subjective beliefs. In (ii), we focus on bin choices, not actual number choices, in order to avoid, as much as possible, considerations of salience of particular numbers, and also because otherwise data from items with smaller domains of plausible number would get more weight than data from items which allows for a wider set of number choices.

All three pieces of data are to be explained as functions of subjective beliefs $P_{ij}$, where $P_{ij}$ is a probability vector of length 15 with $P_{ijk}$ being of subject $i$'s belief about the relative likelihood of bin $k$ for item $j$. So, each $P_{ij}$ defines a likelihood for our data, via appropriate link functions (to be spelled out presently). Variance in subjective beliefs is harnessed by a population-level hyper-prior with central tendency $Q_j$, i.e., a stochastic vector of length 15, which we call a "population-level belief" about item $j$. The structure of this model is pictured in Figure 1.

To fill the structure in Figure 1 with life, we need to spell out three parameterized link functions, one for each task type, and the relation between population-level belief $Q_j$ to individual beliefs $P_{ij}$. Let's start with the latter. For fixed $Q_j$, the idea is that $P_{ij}$ are noise-perturbed variants scattered around $Q_j$, with some parameter $w$ to determine how much perturbation we should expect. To realize this, the model assumes that $P_{ij}$ are distributed according to a Dirichlet distribution with weights given by $wQ_j$:

$$Q_j \sim \text{Dirichlet}(1,\ldots,1) \qquad w \sim \text{Gamma}(2,0.1)$$
$$P_{ij} \sim \text{Dirichlet}(wQ_j)$$

The higher $w$, the more likely it is that $P_{ij}$ is "close" to $Q_j$.

The link function for the slider rating data uses a logit transformation to project observed slider ratings $s_{ijk}$ and latent probabilities $P_{ijk}$, which are bound to lie between 0 and 1, to the reals. The likelihood of logit-transformed observation $s_{ijk}$ is given by a Gaussian with standard deviation $\sigma$ around the logit-transformed predictor $P_{ijk}$. On top of that, there is a parameter $\kappa$, the steepness of the logit transform of $P_{ijk}$, that allows response likelihoods to capture end-point affinity for $\kappa > 1$ (values of $P_{ijk}$ close to 0 or 1 are likely mapped to 0 or 1) or end-point aversion for $\kappa < 0$ (values of $P_{ijk}$ are likely to be realized as more median), with a prior that expects $\kappa = 1$.

$$\text{logit}(s_{ijk}) \sim \text{Norm}(\text{logit}(P_{ijk},\kappa),\sigma)$$
$$\sigma \sim \text{Gamma}(0.0001,0.0001) \quad \kappa \sim \text{Gamma}(5,5)$$

The link function for number choice data treats each bin $n_{ij}$ as the draw from a categorical distribution where the probability of bin $k$ is proportional to $\exp(aP_{ij})$, i.e., a soft-max choice from $P_{ij}$. The higher parameter $a$, the more likely the mode of $P_{ij}$ is. For $a \to 0$, all bins become equiprobable.

$$n_{ij} \sim \text{Categorical}(\exp(aP_{ij})) \qquad a \sim \text{Gamma}(2,1)$$

Finally, consider the link function for bin comparisons. We are after the likelihood with which subject $i$ selects the higher bin for item $j$ in bin comparison condition $l$. Suppose $l$ is about comparing the lower bin $b_l$ to the higher $b_h$. The perhaps most natural approach would be to link the likelihood of choosing $b_h$ over $b_l$ to the difference between $P_{ijb_h}$ and $P_{ijb_l}$. This, however, does not appear to be what subjects were doing. For example, take the "marbles" case where we see almost unanimous choices of bin 11 over bin 6, despite seeing any pronounced difference in slider ratings for these bins. Another possible link function that accommodates for this is to assume that what matters in bin comparisons is the distance to the mode of $P_{ij}$: soft-max prefer the bin that is closer to the mode of $P_{ij}$; select random if both bins are equally far from the prototype.[1]

$$c_{ijl} \sim \text{Bern}((1+\exp(2b(1-p^{\text{high}}_{ijl})))^{-1}) \quad b \sim \text{Gamma}(2,1)$$

$$p^{\text{high}}_{ijl} = \begin{cases} 2 & \text{if } mode(P_{ij}) \text{ is closer to higher} \\ & \text{ bin of } l \text{ than to lower bin} \\ 1 & \text{if equal distance} \\ 0 & \text{otherwise} \end{cases}$$

[1]Notice that $(1+\exp(2b(1-p^{\text{high}}_{ijl})))^{-1} = \frac{\exp(bp^{\text{high}})}{\exp(bp^{\text{high}})+\exp(bp^{\text{low}})}$ with $p^{\text{low}}_{ijl} = 2 - p^{\text{high}}_{ijl}$.

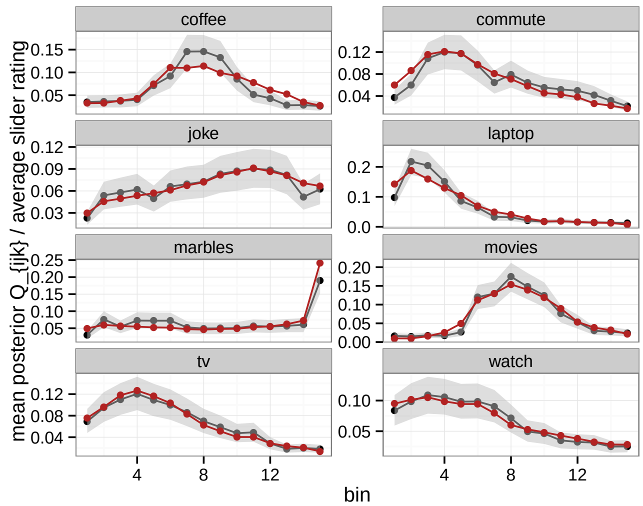



Figure 2: Means of posteriors over $Q_j$ in black with gray area indicating 95% HDIs. Red lines give the average normalized slider ratings for comparison.

## Inference

The model was implemented in JAGS (Plummer, 2003). 50,000 samples were obtained after a burn-in of 100,000 that ensured convergence according to $\hat{R}$ (Gelman & Rubin, 1992).

Of most theoretical interest are posterior estimates of $Q_j$. Figure 2 shows means of the posterior $Q_j$, with their 95% HDIs, alongside the averaged normalized slider ratings. The latter provide a very good approximation of the latent, inferred population-level beliefs. Inspection of posteriors of individual $P_{ij}$ shows that there is ample variation between subjects. Nonetheless, the way the model suggest we should think about harnessing the individual $P_{ij}$s together in a population-level central tendency is closely approximated by simply averaging over slider ratings. Although the match is clearly not perfect, it is good enough to maintain that the latter are a reasonable (enough) and practical way of assessing what the crowd believes despite individual differences.

## Model criticism

Inferences based on the model are only as reliable as the model itself is plausible. Model criticism is therefore important. Figure 3 shows posterior predictive checks at the population-aggregate level for all of our three task types. For the slider rating task, posterior predictions are spot-on. Some of the number estimation data is surprising even for the model trained on this very data. This could have various reasons: (i) the number data does not have a huge influence on the posterior likelihood, (ii) number choices may be influenced by saliency and/or roundness of numbers after all. Finally, there is one condition in the bin comparison task that the model definitely got wrong. This is the choice of what is more likely: that one or that none of 14 marbles thrown into a pool would sink. The model would predict that, given the answers in other conditions, should consider it more likely that one marble sank, but in reality subjects may revise beliefs about "normality" of the marbles, while holding on to an assumption that all marbles behave the same (Degen et al., 2015).

Posterior predictive checks indicate that the trained model captures patterns of answers at the aggregate population level well. To have a more fine-grained measure of model fit, we also looked at posterior predictive $p$-values at the level of subjects and items. Unfortunately, only the slider rating task, provides enough data for such a fine-grained analysis. For the slider rating task, fixing a subject and an item, observations and replicates are probability vectors of length 15. In a first analysis, we used the mean of these probability vectors as a test statistic. The minimum posterior predictive $p$-value over all 20 (subjects) times 8 (items) cases was 0.13, suggesting that the means of observed $s_{ij}$ are non-surprising to the trained model. In a second analysis, we used entropy as a test statistic. In this case, two cases gave significant posterior predictive $p$-values. These were from the two participants who gave a very extreme slider rating for the "marbles" item, basically assigning all "mass" to the last bin. What this suggests is that the model can cope reasonably well also with individual-level data, but has problems accounting for "extreme" choices, given that the population-level hyper-prior on $P_{ij}$ will lead to shrinkage.

## Conclusions

## Acknowledgments

MF's work was supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) and the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727).

## References

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. E. (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty*, *48*, 207–229. doi: 10.1007/s11166-014-9194-z

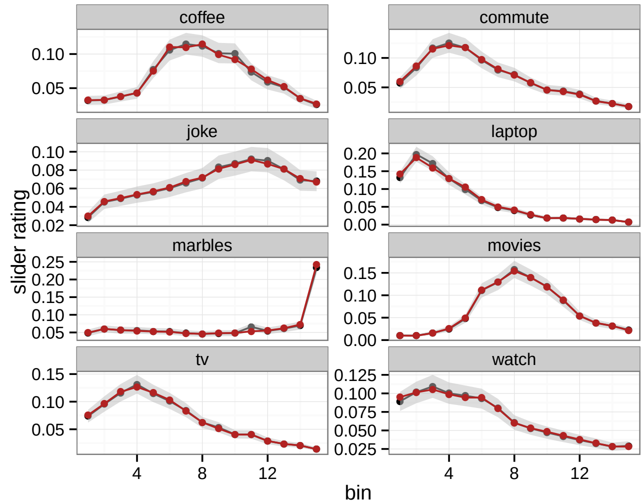
Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of cogsci.*

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

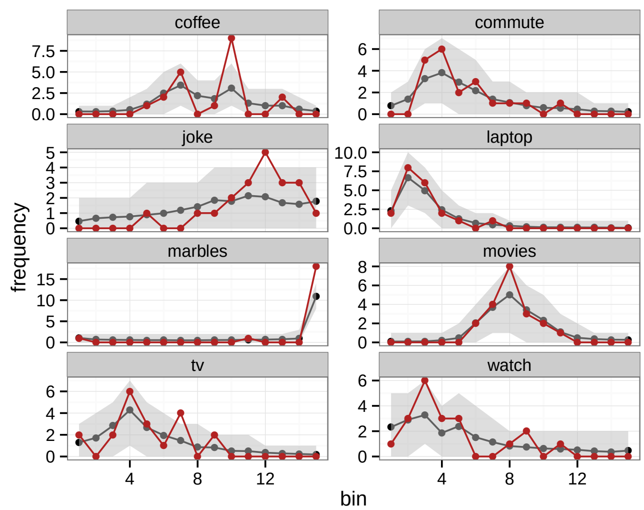
Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *PNAS*, *111*(33), 12002–12007. doi: 10.1073/pnas.1407479111

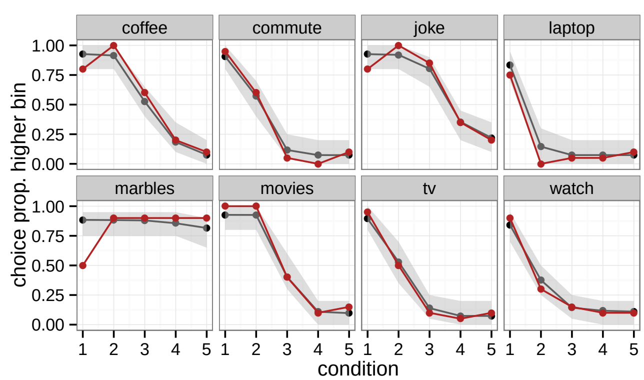
Lee, M. D., & Wagenmakers, E.-J. (2015). *Bayesian cognitive modeling: A practical course*. Cambridge, MA: Cambridge University Press.




(a) slider rating




(b) number estimation




(c) bin comparison

Figure 3: Posterior predictive checks for aggregate data. Red lines give empirical observations. Black lines are means of posterior predictive samples, gray areas are 95% HDIs.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998. doi: 10.1111/j.1551-6709.2009.01045.x

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376. doi: 10.1111/j.1468-0262.2004.00537.x

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychol Science*, *24*(12), 2351–2360. doi: 10.1177/0956797613495418

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge, UK: Cambridge University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.*

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801. Retrieved from `http://www.jstor.org/stable/2284229`

Schlag, K. H., Tremewan, J., & van der Weele, J. J. (online first). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*. doi: 10.1007/s10683-014-9416-x

Schöller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising *few* & *many*. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of SALT 45* (pp. 143–162). doi: 10.3765/salt.v25i0.3058