### Title
Illusory bimodality in repeated reconstructions of probability distributions




### Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

### ISSN
1069-7977



### Authors
Nie, Shaozhi
Wang, Muzhi
Zhang, Hang



### Publication Date
2021

Peer reviewed

# Illusory bimodality in repeated reconstructions of probability distributions

**Shaozhi Nie (nieshaozhi@pku.edu.cn)**
School of Psychological and Cognitive Sciences,
Peking University, 100871, China

**Muzhi Wang (wangmuzhi@pku.edu.cn)**
School of Psychological and Cognitive Sciences,
Peking University, 100871, China

**Hang Zhang[†] (hang.zhang@pku.edu.cn)**
School of Psychological and Cognitive Sciences,
Peking University, 100871, China

## Abstract

Probability density estimation is widely-known as an ill-posed statistical problem whose solving depends on extra constraints. We investigated what prior beliefs people might have in their learning of an arbitrary probability distribution, especially whether the distribution is believed to be unimodal or multimodal. In each block of our experiments, participants repeatedly reconstructed a one-dimensional spatial distribution after observing every 60 new samples from the distribution. The probability distribution function (PDF) they reported on each trial was submitted to a spectral analysis, where the powers for 1-cycle, 2-cycle, ..., n-cycle components respectively indicate participants' tendency of reporting unimodal, bimodal, ..., n-modal distributions. In two experiments, the reported PDFs showed significant bimodality—that the 2-cycle power was above chance and even larger than the 1-cycle power—not only for bimodal distributions, but also for uniform distribution. Such illusory bimodality for uniform distribution was first found when we used an adaptive procedure analogous to "human MCMC", updating the generative distribution of samples from trial to trial to reinforce potential biases in PDFs (Experiment 1). However, even when we fixed the generative distribution across trials (Experiment 2), the illusory bimodality did not vanish. The illusory bimodality was even observed before participants experienced any bimodal distributions in the experiment . We considered a few kernel density models and discuss further computational explanations (e.g. prior beliefs following Chinese Restaurant Process) for this new phenomenon.

**Keywords:** Probability density estimation, human MCMC

## Introduction

Learning a probabilistic model of the world lies at the heart of categorization (Flannagan et al., 1986), inference (Griffiths & Tenenbaum, 2006), optimal foraging (Stephens & Krebs, 1986), and more generally, the human ability of generalizing from a few instances of experience (Tenenbaum et al., 2011). However, the learning of such model as probability density estimation is itself an ill-posed statistical problem that cannot be solved without extra constraints (Vapnik, 1995). What prior beliefs do people have when they learn arbitrary probability distributions?

According to some early studies, people are slower in learning U-shaped than learning Gaussian distributions (Flannagan et al., 1986) and may mistakenly recall multimodal distributions to be unimodal (Nisbett & Kunda, 1985), which seems to suggest that people expect the to-be-learned probability distribution to be unimodal or Gaussian. However, later studies show that people may not have a general difficulty in learning multimodal distributions. For example, people can learn a distribution with as many as four modes from as few as 70 samples (Sun et al., 2019) or use multimodal prior distributions for inference (Acerbi et al., 2014; Sanborn & Beierholm, 2016). Bimodal or U-shaped distributions are not slower to learn or reduced to Gaussian distributions unless samples are associated with high perceptual or mnemonic noises (Yeung & Whalen, 2015). As Tran et al. (2017) pointed out, mistakenly reporting multimodal distributions as Gaussian may simply reflect a failure to learn the distribution except for its mean.

In studies where probability distributions are reasonably learned, evidence even points to the reverse direction: In a motor decision task, people are found to represent their Gaussian-like, unimodal motor error distributions as multimodal distributions (Zhang et al., 2015). Multimodal instead of unimodal representations are also found for inference based on very few (four) samples from a Gaussian distribution (Schustek & Moreno-Bote, 2018). In one study of temporal distributions (Acerbi et al., 2012), hints of multimodal representations for unimodal distributions, though not being described in the text, are visible in their figures. These pieces of evidence together suggest that people's prior beliefs for probability distributions may be (counterintuitively) multimodal instead of unimodal. But this is far from a conclusion, partly because the assertion is too counterintuitive to be built on only a few studies, and partly because the estimation of distribution representations in all these studies is indirect, relying on sophisticated modeling analysis whose assumptions are not necessarily true.

Here we developed a new distribution report task and data analysis method to more directly and sensitively detect the potential multimodality in participants' prior beliefs for probability distributions. In each block of our experiments, we asked participants to reconstruct a distribution on a 20-bin distribution reporter, after observing every 60 new samples generated from the distribution. An adaptive procedure was applied to the generative distribution after each trial so that it

[†]Corresponding author

changed slightly towards participants' report on the trial. Inspired by the idea of human Markov chain Monte Carlo (human MCMC, Sanborn and Griffiths (2008)) and more closely by Lew and Vul's (2015) cascading method, this adaptive procedure was designed to reinforce the biases in participants' behavioral patterns and thus to more efficiently capture the influences of prior beliefs that might otherwise be transient.

A common practice of previous distribution report studies (Nisbett & Kunda, 1985; Yeung & Whalen, 2015) was to average the reported probability distribution functions across trials and participants, which might have cancelled out different multimodal distributions in different trials or participants. To avoid so, we applied spectral analysis to the reported probability distribution functions in single trials and averaged the power separately for different frequency components. The 1-cycle component corresponds to unimodal, 2-cycle component corresponds to bimodal, and so on.

In two experiments, we found participants not only were able to reconstruct real bimodal distributions, but also showed bimodality in their reconstruction of uniform distributions—a phenomenon we termed "illusory bimodality". Further statistical and modeling analysis have allowed us to exclude a few trivial explanations for illusory bimodality. We discuss its implications for human prior beliefs in representing probability distributions.

## Methods

### Participants

Forty-eight students from Peking University participated in our experiments (Experiment 1: 24 participants, 2 male, aged 18–28; Experiment 2: 24 participants, 9 male, aged 19–25). The study was approved by the ethics committee of School of Psychological and Cognitive Sciences at Peking University. All participants provided written informed consent. Participants received a base payment of 60 RMB plus a bonus that could range from 0 to 60 RMB according to their performance.

### Experiments

Participants were required to reconstruct a visually-presented one-dimensional probability distribution based on the samples they saw, in a cover story of "collecting meteors on planets" (Figure 1A). On each planet, participants watched animated meteors falling down to the ground and were told that the horizontal locations of the fallen meteors followed a specific distribution. They were encouraged to learn this distribution and allocate "meteor-collecting machines" among the possible locations proportional to each location's likelihood of welcoming meteors. Participants would receive bonus points for the collected meteors. The scoring rule of our task was designed in such a way that matching the allocation of meteor-collecting machines to the generative distribution of meteor locations would maximize expected reward and participants were explicitly instructed so. That is, participants were motivated to reconstruct the generative distribution of locations as accurately as they could.

Each block was accompanied by a unique background image of a planet. Participants were told that the meteor locations on each planet (block) had its own distribution. Each planet (block) started with an initial 60 meteor samples and proceeded through 30 trials, on each of which participants reported the distribution of meteor locations on the planet and then received 60 more samples as feedback.

**Samples**. During the presentation of samples, each meteor appeared at one of 20 possible horizontal locations, falling to the ground from the same height and at the same speed. Every 100 ms a new meteor appeared and lasted for 500 ms, resulting in no more than 4 meteors on the screen at a time and a total presentation time of 6400 ms for 60 samples.

**Distribution report**. The distribution report was self-paced, where participants could either click the mouse on a bar or drag the mouse through bars to set the heights of 20 bars to indicate the relative probability of each location. Participants could change their report as many times as they needed before clicking on the "OK" button to confirm. In the subsequent feedback, the relative probabilities participants reported were scaled to adding up to 1 and shown as the background for the next 60 samples.

**Adaptive procedure**. In Experiment 1 ("adaptive-change experiment"), to amplify the potential biases in participants' report, after each trial the reported distribution was used to update the generative distribution of next trial. According to the adaptive procedure (dashed-line box at the bottom of Figure 1A), the generative distribution $P_{t+1}$ on trial $t+1$ is a weighted summation of the reported distribution $R_t$ and the generative distribution $P_t$:

$$P_{l,t+1} = \alpha \cdot R_t + (1-\alpha) \cdot P_{l,t}. \quad (1)$$

The updating factor $\alpha$ was set to 0.2, which is small enough to prevent participants from discovering the adaptive procedure and meanwhile large enough for the generative distribution and participants' report to converge. In Experiment 2 ("fixed-distribution experiment"), the generative distribution was fixed in each block and no adaptive procedure was used between trials, equivalent to setting the $\alpha$ in Eq. 1 to be 0.

**Design**. Except for the presence or absence of the adaptive procedure, the two experiments were identical in stimuli, procedure, and design. In both experiments, each participant completed a practice block of 5 trials and six experimental blocks of 30 trials, which took approximately 80 minutes. A block might start with one of three different generative distributions (see Figure 1B): a uniform distribution and two bimodal distributions that are mixtures of two (discretized) left- or right-skewed beta-distribution components ("skewed bimodal distributions"). The practice block used the uniform distribution. The six experimental blocks used an ABCABC design for the three different distribution conditions, with their orders counterbalanced across participants.

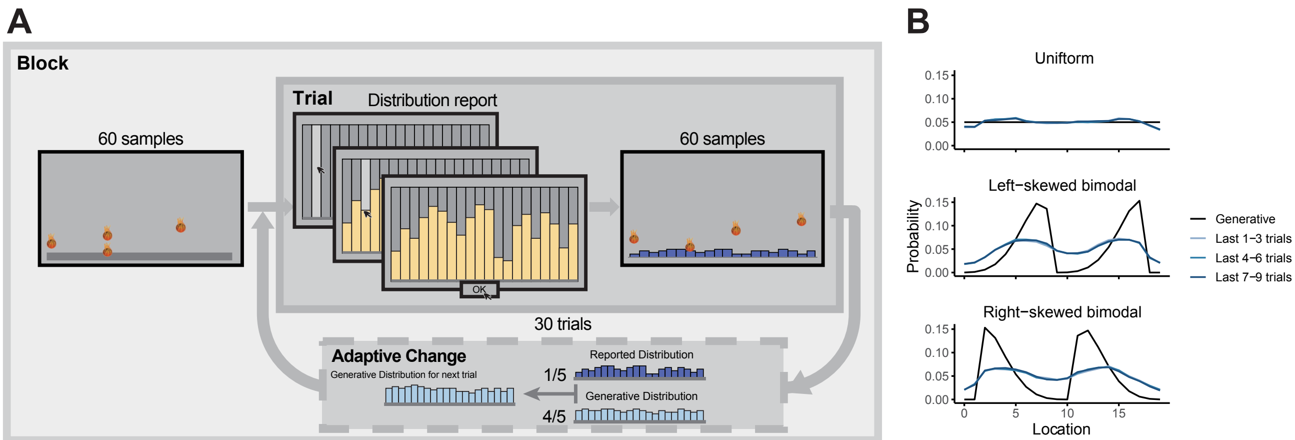


**Figure 1. Repeated reconstruction of probability distributions**. **A**, Task. Each planet (block) started with an initial 60 meteor samples (visualized as falling firing balls) and proceeded through 30 trials. On each trial, participants were required to report the distribution of the horizontal locations of the meteors on the planet. To do so, they used the mouse to adjust the heights of the 20 bars and clicked on OK to confirm. They then received 60 more samples as feedback. In each block, the generative distribution for samples was initially one of three distributions (one uniform distribution and two bimodal distributions, as shown in **B**). In Experiment 1 ("adaptive-change experiment"), the generative distribution was updated after each trial according to an adaptive procedure (weighted summation of reported distribution and the previous generative distribution, see dashed-line box). In Experiment 2 ("fixed-distribution experiment"), the generative distribution was fixed throughout each block. **B**, Response convergence in Experiment 1. Each panel is for one of the three different initial generative distributions (black curve). The reported probability density functions (PDFs) of the last 1–3 trials, 4–6 trials, and 7–9 trials, averaged across participants, overlapped with each other, indicating that participants' responses well converged at the end of each block.

In other words, participants would not experience the same distribution condition in two consecutive blocks. Neither were they informed that two different blocks might have the same distribution.

## Statistical analysis

**Across-block consistency**. We first tested whether participants' reports in two blocks of the same distribution had a higher similarity than those of different distributions, using the procedures as follows. Distribution reports in each block were averaged across 30 trials, which resulted in a 20-dimensional vector (i.e., the averaged PDF) for the block. The similarity of two blocks were defined as the cosine value of the angle between their vectors. For each participant, 6 blocks were divided into the first 3 blocks and the last 3 blocks. Between one block from the first group and another block from the last group, there were 9 similarities, among which, 3 were between two blocks with same initial distribution and 6 were not. We averaged the similarities within each participant and obtained the similarity between blocks of same distribution and the similarity between blocks of different distributions for each participant. Last, a paired t-test was applied to see whether the two similarities were different.

**Spectral analysis**. For each participant and trial, we applied Fast Fourier Transform (implemented by the fft function of the Python 3 library Numpy) to the PDF (i.e., a series of 20 probabilities) participants reported.

**Permutation test**. In both Experiment 1 and Experiment 2, we generated surrogate data by randomly shuffling the series of 20 probabilities separately for each trial and each participant, and applied the Fourier transformation procedures described above for real data to the surrogate data. This permutation procedure was repeated for 1000 times to obtain the distribution of chance-level powers for each frequency, based on which we tested whether the powers at particular frequencies in real data were significantly above the chance level.

## Modeling

**Histogram model**. The histogram model assumes an observer who has perfect memory and who separately counts the number of samples fallen into each bin of the distribution report task. Up to trial t of a block, the observer has observed a total of $60 \cdot t$ samples. The reported probability for bin location $l (l \in [1, 20])$ is thus the proportion of samples cumulated at the location:

$$R_{l,t} = \frac{\Sigma_{\tau=1}^{t} N_{l,t}}{60 \cdot t}. \quad (2)$$

where $N_{l,t}$ is the cumulated number of samples at location $l$ up to trial $t$.

**Gaussian model**. The Gaussian kernel model takes the limited capacity of working memory as well as perceptual and mnemonic noises into consideration. It assumes that the observer can only use the 60 samples observed in the last trial. Each sample influences not only the hit location but also its nearby locations, with the strength of influence following a Gaussian kernel function centered at the hit location. The reported probability for bin location $l (l \in [1,20])$ is thus

$$R_{l,t} = \frac{\Sigma_{\xi=1}^{20} N_{l,t} \cdot \varphi(\frac{l-\xi}{\sigma})}{\Sigma_{l'=1}^{20} \Sigma_{\xi=1}^{20} N_{l',t} \cdot \varphi(\frac{l'-\xi}{\sigma})}. \tag{3}$$

where $\varphi$ denotes the density function of standard Gaussian distribution. The Gaussian kernel model has one parameter, the kernel width $\sigma$.

**Mexican-hat model**. We adopted the Mexican-hat curve widely used in modeling neuronal receptive fields (Marr & Hildreth, 1980) as another kernel function to model the potential influence of a sample on locations other than the hit location. It assumes that each sample has positive effects for nearby locations but negative effects for more distant locations. Otherwise the model is similar to the Gaussian kernel model described above. The reported probability for bin location $l (l \in [1,20])$ is thus

$$R_{l,t} = \frac{\mathrm{ReLU}(\Sigma_{\xi=1}^{20} N_{l,t} \cdot \psi(l-\xi,\sigma))}{\Sigma_{l'=1}^{20} \mathrm{ReLU}(\Sigma_{\xi=1}^{20} N_{l',t} \cdot \psi(l'-\xi,\sigma))}. \tag{4}$$

where $\psi$ is the Mexican hat wavelet with width $\sigma$. The rectification function ReLU(.) sets negative values to 0 and does not change positive values. The Mexican-hat model has one free parameter $\sigma$.

## Results

We first verified that participants in our task did learn from samples and make reliable report of probability distribution functions (PDFs). Figure 1B shows the PDFs reported in the last a few trials of a block, separately for the three distribution conditions (uniform, left-skewed bimodal, and right-skewed bimodal) and averaged across all 24 participants in Experiment 1. There were little differences between the last 1–3, 4–6, and 7–9 trials, all of which resembled the starting generative distribution in the block. Moreover, each participant completed two separate blocks for each distribution condition, which allowed us to evaluate their consistency across blocks, in other words, to test whether the PDFs reported for the same condition were more similar to each other than those of two different conditions (see Methods). As expected, the between-block cosine similarity was significantly higher for the same distribution condition than for two different distribution conditions ($t_{23} = 3.60, p = 0.002$).

Meanwhile, participants' report had patterned deviations from the initial generative distribution (Figure 1B), among which one pattern was counterintuitive: The reported PDFs for the uniform distribution seemed to have two modes. As we will see, this "illusory bimodality" is more pronounced in single trials, because averaging multimodal PDFs with varying mode positions across trials or participants might cancel out the modes. Next we look into single trials and, to circumvent the cancelling-out issue, we use a spectral analysis to provide group-level measures of illusory bimodality.

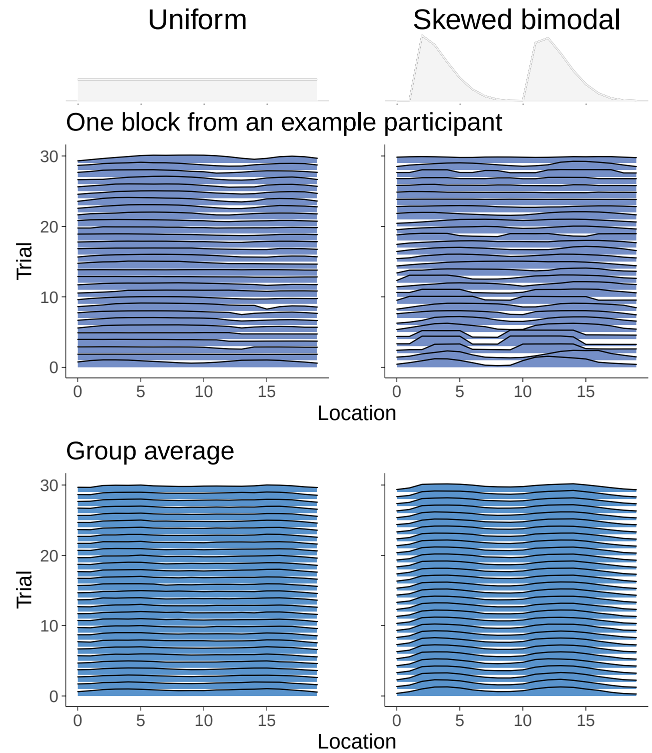


**Figure 2. Ridgeline plots for reported PDFs of the 30 trials across the block**. Top and bottom rows are respectively for one block from an example participant and the group average. Left and right columns are respectively for the uniform and skewed-bimodal conditions (left- and right-skewed collapsed). Closely-located modes in adjacent trials would overlap with each other to form vertical ridges, that is, darker bands along the vertical axis. Two separate ridges are visible not only for the skewed-bimodal condition whose generative distributions were really bimodal, but also for the uniform condition, especially in the plot for one block of one participant, which demonstrates the existence of illusory bimodality in single trials.

### Illusory bimodality in uniform distributions

The reported PDFs for each of the 30 trials in a block are shown in Figure 2 as ridgeline plots, where closely-located modes in adjacent trials would overlap with each other to form vertical ridges (i.e., darker bands). Two separate ridges are visible not only for the skewed-bimodal condition (left- and right-skewed collapsed) whose generative distributions were really bimodal, but also for the uniform condition, especially in the plot for one block of one participant (upper-left panel of Figure 2), which demonstrates the existence of illusory bimodality in single trials.

To measure bimodality and higher-order multimodality signals, we performed a spectral analysis on the reported PDF, which decomposes the PDF of each trial (i.e., a series of 20 probabilities) into frequency components of 1-cycle, 2-cycle, ..., 9-cycle. The resulting power spectrum was then averaged across participants and blocks, shown separately for trials 1–30 in a block (Figure 3A). The power spectrum of the skewed-bimodal condition was evidently dominated by the 2-cycle component. Our further statistical analysis thus focused on the power spectrum of the uniform distribution.

Compared with random permutations (see Methods), the reported PDFs for the uniform distribution had significantly higher power in 1-cycle ($p < 0.001$), 2-cycle ($p < 0.001$) and 3-cycle ($p < 0.001$).We further tested the power of these three frequencies separately for the 30 trials in the uniform blocks and found significant 1-cycle and 2-cycle power across almost all trials (Figure 3B).

Could this illusory bimodality arise naturally from samples? We considered three kernel density models, which differ in the assumed functional form for the kernel, or in cognitive terms, the way each sample influences the reported PDF. Among them, the histogram model assumes that the reported PDF is a histogram of all samples, with each sample contributing to the bin it belongs to. In contrast, the Gaussian model applies a Gaussian kernel to each sample, echoing visuo-motor Gaussian noises. The Mexican-hat model assumes a difference-of-Gaussian kernel (mimicking the shape of a Mexican hat), so that each sample increases the density of its neighboring regions but decreases the density in more distant regions, which is motivated by psychophysical findings (Marr & Hildreth, 1980). The histogram model has no free parameters and for the other two models, we estimated the kernel width parameters that minimize the summed square errors between the PDFs participants reported and those predicted by the model (see Methods).

We compared the power spectrum predicted by these models to the data in Experiment 1 (Figure 3A), focusing on the qualitative differences in the pattern. The prediction of the histogram model bore little similarity to the data. The Gaussian model captured the higher power in lower frequencies but contradicted the data in predicting a lower power in 2-cycle than in 1-cycle for the uniform condition. The prediction of the Mexican-hat model was close to the Gaussian model.

The over-simplified models we considered above were used to provide computational insights for the cognitive processes behind illusory bimodality. Together, they suggest that illusory bimodality can hardly arise from simple kernel density models, where each sample is treated as equal. Later we will discuss prior beliefs and more structured representations of probability distributions.

## Illusory bimodality does not vanish under fixed generative distributions

Though the adaptive procedure in Experiment 1 provides a sensitive measurement for prior beliefs that might otherwise be too elusive to grasp, its results are potentially open to some trivial explanations. For example, participants might have realized their response could influence future stimuli and tried to take advantage of the adaptive procedure. If so, however, they should have placed all collecting machines into one bin, instead of spreading them out over two modes, let alone shifting the modes from trial to trial (Figure 2). As a second concern, early random patterns in samples, after amplified by the adaptive procedure, might have led to illusory bimodality. But this possibility is also largely ruled out, according to the modeling analysis we reported above.

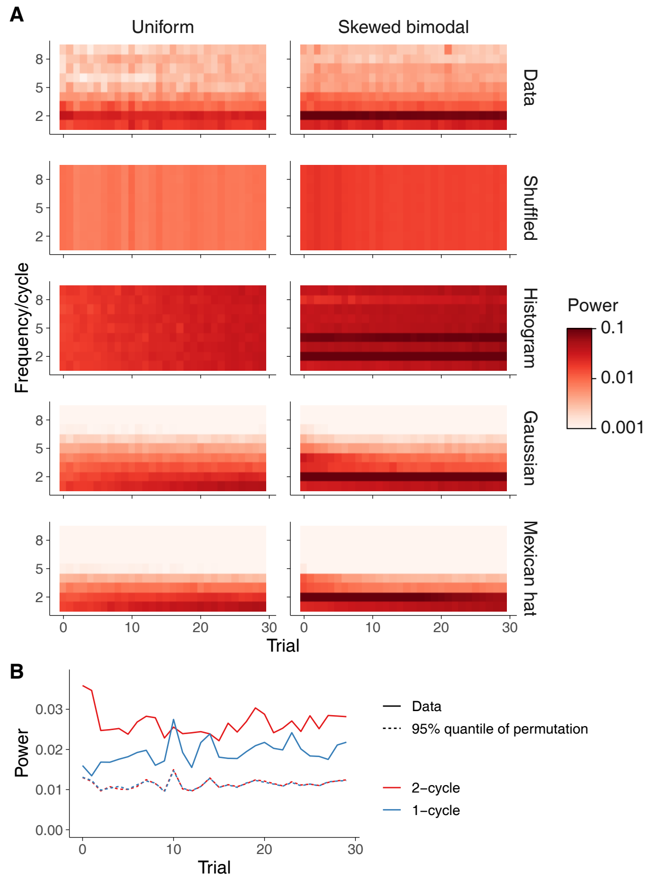


**Figure 3. Powers of frequencies for Experiment 1. A**, Power spectrum for all frequencies (from 1-cycle to 9-cycle) averaged across participants for each of the 30 trials in the block. Left and right columns are for the uniform and skewed-bimodal distributions. Colors code power spectrum on the logarithmic scale. Darker colors denote higher values. Rows from top to bottom: power spectra of the reported PDFs from real data, shuffled data and three different model simulations. The powers of lower-frequency components, especially the 1-cycle and 2-cycle components, were visually higher in real data than those of shuffled data (see **B** for results of statistical tests). See text for the implications of model simulations. **B**, Powers for the 1-cycle and 2-cycle components in the uniform condition. Both the 1-cycle and 2-cycle powers in real data (blue and red solid curves) were higher than the 95% quantile of those of shuffled data (blue and red dashed curves, overlapped).

Still, one may have some other reasons to doubt that the bimodal illusion observed in Experiment 1 was an artefact from the adaptive procedure. To exclude this possibility and to further test whether illusory bimodality can be washed out by continuing exposure to uniform distribution, we performed Experiment 2 (with 24 new participants), which was identical to Experiment 1 except that there was no adaptive procedure and the generative distribution was fixed throughout each block.

Under such fixed generative distributions, we still observed illusory bimodality (Figure 4A), with the 2-cycle power in the uniform condition significantly above the chance level ($p <$

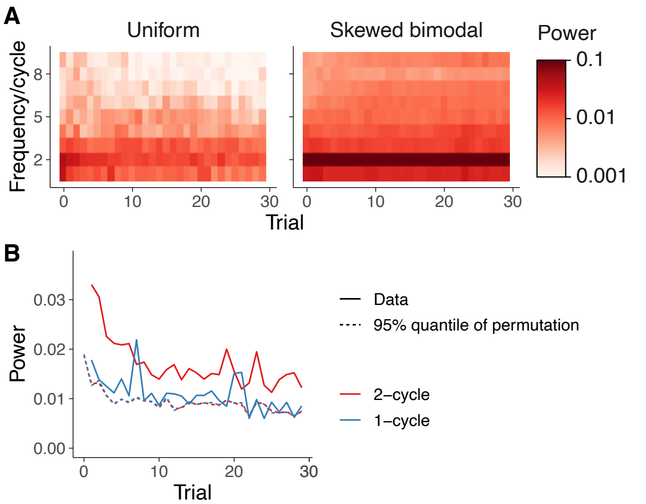


**Figure 4. Power spectrum of the reported PDFs for Experiment 2**. Conventions follow Figure 3. **A**, Power spectrum for all frequencies (from 1-cycle to 9-cycle) averaged across participants for each of 30 trials in the block, separately for the uniform and skewed-bimodal condition but also in the uniform condition. **B**, Powers for the 1-cycle and 2-cycle components in the uniform condition. Similar to Experiment 1, the above-chance-level 2-cycle power in the uniform condition ("illusory bimodality") persisted throughout the block in Experiment 2, while the 1-cycle power was indistinguishable from chance.

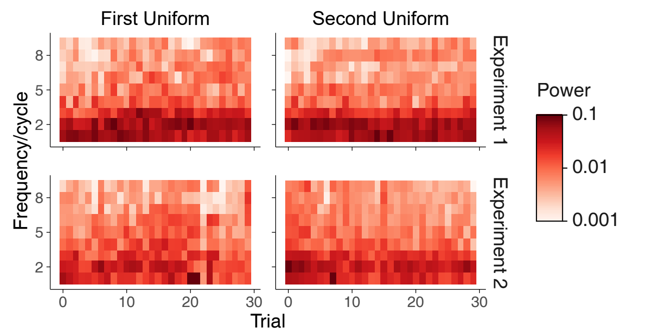


**Figure 5. Power spectra of the reported PDFs in the two uniform blocks for the participants whose experiment started from the uniform condition**. Conventions follow Figure 3A. Left and right columns are respectively for the first and second uniform blocks in the experiment. Top and bottom rows are respectively for Experiments 1 and 2, each of which had 8 participants whose experiment started from the uniform condition. Even in these participants' first uniform block, the 2-cycle power was the 2-cycle power was still significantly above chance ($p < 0.001$ for both Experiments 1 and 2).

0.001). Moreover, the significance in the 2-cycle power was throughout all 30 trials in the uniform blocks (Figure 4B).

## Illusory bimodality emerges even before real experience of bimodal distributions

In our experiments, each participant completed two blocks of uniform condition and four blocks of bimodal conditions. The illusory bimodality in the uniform condition might be explained by an influence from earlier bimodal blocks. For example, participants might use the distribution learned in a previous block as their prior for the present block, though this would not explain why illusory bimodality did not vanish even after participants had observed up to 1800 samples from the uniform distribution.

We further extracted the first block of the participants whose experiment started from the uniform condition (Figure 5). We found that even in these blocks (8 participants for each experiment), the 2-cycle power was still significantly above chance ($p < 0.001$ for both experiments).

## Discussion

In the present study, we asked human participants to reconstruct probability distributions from samples. Despite their fast learning and overall accurate reconstructions, we found a counterintuitive bias—illusory bimodality—in participants' reconstructions of uniform distributions, that is, participants reported a bimodal distribution after observing samples from a uniform distribution. We first found the bias in Experiment 1, where an adaptive procedure was used to reinforce participants' biases. In Experiment 2, we excluded the possibility that illusory bimodality could be an artefact of the adaptive procedure: when the adaptive procedure was removed, we still observed illusory bimodality. Remarkably, participants exhibited illusory bimodality even before they were exposed to any other distributions in our experiments and even after they had continuously observed 1800 samples from the uniform distribution.

Some biases we found in representing real bimodal distributions (Figure 1B) had been reported in earlier studies, such as the loss of local details in variance (Tran et al., 2017) or skewness (Sun et al., 2019). But why was not illusory bimodality found in earlier studies? An important reason, we think, is our use of more sensitive experimental paradigm and data analysis methods, including the adaptive procedure, multiple times of reconstructions and spectral analysis. Had we used a one-time test for each distribution or directly averaged the reported probability distributions across trials or participants, as many previous studies did, we would probably have lacked the statistical power to detect illusory bimodality.

Our finding of illusory bimodality provides further evidence for the hypothesis that people may have multimodal representations for unimodal distributions (Zhang et al., 2015). Unlike previous evidence for the hypothesis (Schustek & Moreno-Bote, 2018; Zhang et al., 2015), our results do not rely on complicated modeling methods.

What is the origin of illusory bimodality? Illusory bimodality was found even in the first block, that is, before exposure to any non-uniform distributions in the experiment, which suggests that bimodality is already part of participants' prior beliefs before our laboratory experiments.

Though seeming counterintuitive at first sight, a prior belief of multimodality is consistent with the fact that the world is often "bumpy", with clustered distributions of resources (Orhan & Jacobs, 2013) and co-existence of multiple causes

(Gershman & Blei, 2012). Meanwhile, multimodal representations may also reflect a discretized representation, probably of basis functions (Anderson & Van Essen, 1994; Poggio, 1990), which serves as a useful approximation under limited cognitive resources (Tran et al., 2017; Zhang et al., 2015).

How are prior beliefs of multimodality used to constrain the learning of arbitrary probability distributions from samples? From the perspective of Bayesian inference, distributions resulting from a Chinese Restaurant Process can serve as a flexible prior to accommodate distributions with indefinite number of modes (Gershman & Blei, 2012). Predictions from these perspectives need to be tested in future research.

## Acknowledgments

This work was partly supported by grant 31871101 from National Natural Science Foundation of China and funding from Peking-Tsinghua Center for Life Sciences to H.Z.

## References

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLOS Computational Biology*, *10*(6), e1003661. https://doi.org/10.1371/journal.pcbi.1003661

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLOS Computational Biology*, *8*(11), e1002771. https://doi.org/10.1371/journal.pcbi.1002771

Anderson, C. H., & Van Essen, D. C. (1994). Neurobiological computational systems. In J. M. Zurada, II Marks, & C. J. Robinson (Eds.), *Computational intelligence imitating life* (pp. 213–222). IEEE Press.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(2), 241–256. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Gershman, S. J., & Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12. https://doi.org/10.1016/j.jmp.2011.08.004

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *207*(1167), 187–217.

Nisbett, R. E., & Kunda, Z. (1985). Perception of social distributions. *Journal of Personality and Social Psychology*, *48*(2), 297–311.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological review*, *120*(2), 297–328. https://doi.org/10.1037/a0031541

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor symposia on quantitative biology*, *55*, 899–910.

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLOS Computational Biology*, *12*(4), e1004859. https://doi.org/10.1371/journal.pcbi.1004859

Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain monte carlo with people. *Advances in neural information processing systems*, *20*, 1265–1272.

Schustek, P., & Moreno-Bote, R. (2018). Instance-based generalization for human judgments about uncertainty. *PLOS Computational Biology*, *14*(6), e1006205. https://doi.org/10.1371/journal.pcbi.1006205

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton University Press.

Sun, J., Li, J., & Zhang, H. (2019). Human representation of multimodal distributions as clusters of samples. *PLOS Computational Biology*, *15*(5), e1007047. https://doi.org/10.1371/journal.pcbi.1007047

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tran, R., Vul, E., & Pashler, H. (2017). How effective is incidental learning of the shape of probability distributions? *Royal Society Open Science*, *4*(8), 170270. https://doi.org/10.1098/rsos.170270

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Yeung, S., & Whalen, A. (2015). Learning of bimodally distributed quantities. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, J. C. D., & M. P. P. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2745–2750).

Zhang, H., Daw, N. D., & Maloney, L. T. (2015). Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nature Neuroscience*, *18*(8), 1152–1158. https://doi.org/10.1038/nn.4055