

# A Bayesian Truth Serum

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Bayesian Truth Serum</b>	<b>1</b>
2.1	Model . . . . .	2
2.2	Score . . . . .	3
<b>3</b>	<b>Properties</b>	<b>3</b>
<b>4</b>	<b>Practical Aspects</b>	<b>6</b>
<b>5</b>	<b>Summary</b>	<b>6</b>

## 1 Introduction

Many fields of science build on subjective data, e.g. surveys of behaviour in psychology. The value of such surveys, unfortunately, is limited by dishonesty of the respondents. Quality could therefore potentially be improved if respondents were incentivised to answer as honestly as possible, for example via a scoring system. But scoring answers to questions that have no objective truth is no easy task.

Prelec [Pre04] presents a scoring system, called the Bayesian truth serum, that in expectation scores honest answers higher than dishonest answers. Use of the Bayesian truth system, e.g. as the basis of a monetary reward, could incentivise respondents to answer honestly, thereby potentially improving the quality of surveys.


In this document, we examine the Bayesian truth serum, and attempt to lay out its argument as clearly as possible.

## 2 The Bayesian Truth Serum

Suppose that some population of people take a survey with a subjective question. You are asked for your best guess of the distribution of answers to this question. As an individual, your guess is not entirely uninformed: you know your own opinion on the survey question. Although this one

sample might not say much about the population’s distribution, it does constitute a valid sample and hence should say *something*; that is, in a way one’s own opinion is an informative “sample of one” [Pre04].

Denote your best guess by  $f$ , which consists of frequencies for every answer. Compared to the population’s average best guess  $\langle f \rangle$ , the *common prediction*, it thus can be expected that  $f - \langle f \rangle$  is highest at your opinion, because unlike the common prediction your best guess is informed by your opinion, albeit ever so slightly; this phenomenon is indeed observed in practice [Pre04]. Therefore, if you are a meta-rational<sup>1</sup> Bayesian agent and realise that this is the case, you should believe that your opinion has the highest probability of being *more common than commonly predicted*.

The Bayesian truth serum (BTS) is based upon exactly this observation. In the BTS, answers are scored by a metric  that measures how common an answer is compared to the common prediction. Indeed, as argued above, from the respondents’ point of view, the probability of scoring highest is maximised by answering honestly.

## 2.1 Model

Besides an answer  $x^n$  to every question, the BTS requires every respondent  $n$  to also give their best guess  $f^n$  of the distribution of answers to this question. These random variables are modelled according to a particular graphical model, depicted in Figure 1 and described next.

Every individual  $n$  holds an opinion  $t^n$ , where we denote  $t^n = i$  simply by  $t_i^n$ . The opinions of all individuals are conditionally independent given some latent variable  $\omega$ , and the conditional distribution  $p(t_i^n | \omega) = \omega_i$  is the *same* for every individual. Furthermore, given some opinion  $t^n$  held by individual  $n$ , they answer the survey’s question with  $x^n$  according to some *answering strategy*  $p(x^n | t^n)$ , where we again denote  $x^n = i$  simply by  $x_i^n$ , and they predict that *others* answer  $i$  to the question with frequency  $f_i^n$ . Note that assuming that a latent variable  $\omega$  such that  $p(t_i^n | \omega) = \omega_i$  exists is equivalent to assuming that the opinions  $(t^n)$  form an exchangeable sequence.

It often does not matter which particular individual we’re talking about. In such cases, to simplify notation, we suppress the index of the individual. For example,  $t_i$  means that a particular individual holds opinion  $i$ ;  $x_j$  means that a particular individual answers  $j$ ; and  $f_k$  means that a particular individual predicts answer  $k$  with frequency  $f_k$ . It should be clear from the context whether these particular individuals are the same or different.

---

<sup>1</sup> The individual needs to reason about their own reasoning.

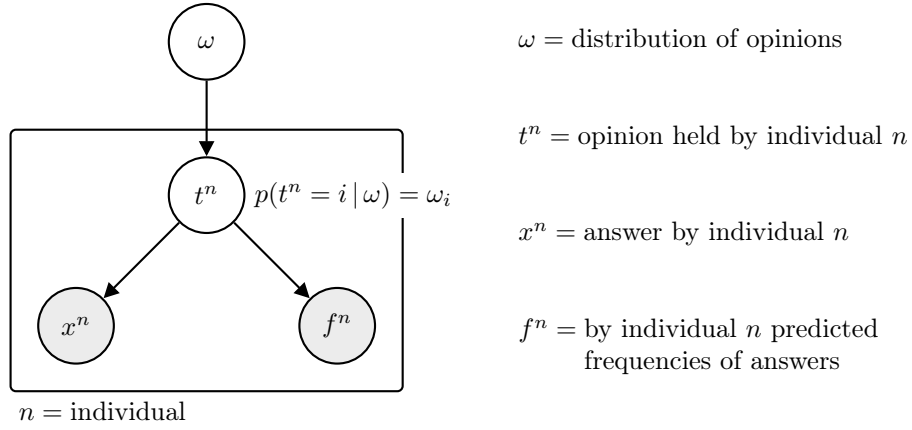




Figure 1: Model of the world assumed by the Bayesian truth serum

## 2.2 Score

Given answers ( $x^n$ ) and predicted frequencies ( $f^n$ ), the Bayesian truth serum assigns a response  $(x_i, f)$  score  ( $x_i, f$ ):

$$\text{score}_{\text{flask}}(x_i, f) = \overbrace{\log \frac{\langle x \rangle_i}{\langle f \rangle_i}}^{\text{information score}} - \overbrace{\sum_j \langle x \rangle_j \log \frac{f_j}{\langle x \rangle_j}}^{\text{prediction penalty}},$$

$\langle x \rangle_i$  = average of  $(\mathbb{1}(x^n = i))$ ,  
 $\langle f \rangle$  = geometric average of  $(f^n)$ .

The score  consists of two contributions:

- (1) the *information score* measures how common answer  $i$  is compared to the common prediction, and
- (2) the *prediction penalty* forms a Kullback–Leibler (KL) divergence between the true and predicted frequency of answers.

## 3 Properties

The Bayesian truth serum enjoys a number of encouraging theoretical results that motivate its use. In this section, we explicitly list the assumptions made by the BTS (Assumptions 3.1 to 3.3) and present three propositions (Propositions 3.1 to 3.3). All results in this section have originally been presented by Prelec [Pre04].

**Assumption 3.1** (Exchangeability). The opinions of the individuals ( $t^n$ ) form an exchangeable sequence. This assumption justifies the graphical model assumed by the BTS (Figure 1). ◀

**Assumption 3.2** (Stochastic Relevance). Different opinions imply different posterior distributions over  $\omega$ : if  $i \neq j$ , then  $p(\omega | t_i) \neq p(\omega | t_j)$ . This assumption is a technical convenience that will be used to conclude uniqueness of maximisers. ◀

**Assumption 3.3** (Sufficiently Large Sample Size). The variances of  $\langle x \rangle$  and  $\langle f \rangle$  are sufficiently low so that they can be approximated by their limits. ◀

**Proposition 3.1** (Truth Telling is an Equilibrium). Suppose that a respondent holds opinion  $k$ , answers  $i$ , and predicts  $f$ ; and everyone else answers and predicts honestly. Then the respondent does best also by answering and predicting honestly:

$$\max_{(i,f)} \mathbb{E}(\text{🧪}(x_i, f) | t_k) = (k, p(t | t_k)).$$

Furthermore,

$$\mathbb{E}(\text{information score}(x_i) | t_i) = \sum_j p(t_j | t_i) D_{\text{KL}}(p(\omega | t_i, t_j) \| p(\omega | t_j)). \quad \blacktriangleleft$$

*Proof.* If everyone else answers and predicts honestly, then

$$\langle x \rangle_i \approx p(t_i | \omega) = \omega_i, \quad \log \langle f \rangle_i \approx \sum_j p(t_j | \omega) \log p(t_i | t_j) = \sum_j \omega_j \log p(t_i | t_j).$$

The information score depends only on  $i$  and the prediction penalty only on  $f$ , so we may consider them separately. First,

$$\begin{aligned} \mathbb{E}(\text{information score}(x_i) | t_k) &= \int p(\omega | t_k) \sum_j p(t_j | \omega) \log \frac{p(t_i | \omega)}{p(t_i | t_j)} \\ &= \sum_j p(t_j | t_k) \int p(\omega | t_k, t_j) \log \frac{p(\omega | t_i, t_j)}{p(\omega | t_j)} d\omega \\ &\leq \sum_j p(t_j | t_i) D_{\text{KL}}(p(\omega | t_i, t_j) \| p(\omega | t_j)) \end{aligned}$$

with equality if and only if  $k = i$ . Second,

$$\begin{aligned} \mathbb{E}(\text{prediction penalty}(f) | t_k) &= \int p(\omega | t_k) \sum_j p(t_j | \omega) \log \frac{f_j}{p(t_j | \omega)} \frac{p(t_j | t_k)}{p(t_j | t_k)} d\omega \\ &= \int p(\omega | t_k, t_j) \sum_j p(t_j | t_k) \log \frac{f_j}{p(t_j | t_k)} \frac{p(\omega | t_k)}{p(\omega | t_k, t_j)} d\omega \\ &\leq -D_{\text{KL}}(f_j \| p(t_j | t_k)) - \sum_j p(t_j | t_k) D_{\text{KL}}(p(\omega | t_k, t_j) \| p(\omega | t_k)) \end{aligned}$$

with equality if and only if  $f_j = p(t_j | t_k)$ . ◻

Q.E.D.

Proposition 3.1 shows that truth telling is a Bayesian Nash equilibrium. It also shows that the truth-telling information score, which also is the optimal information score, measures how much on

average another's posterior distribution over  $\omega$  changes upon learning your opinion; this suggests that experts might enjoy higher expected information scores. The following proposition shows that truth telling is also the best Bayesian Nash equilibrium.

**Proposition 3.2** (Truth Telling is the Best Equilibrium). The truth-telling equilibrium is the equilibrium that maximises the expected information score. ◀

*Proof.* In an arbitrary equilibrium,

$$\langle x \rangle_i \approx p(x_i | \omega), \quad \log \langle f \rangle_i \approx \sum_j p(t_j | \omega) \log p(x_i | t_j).$$

Then

$$\begin{aligned} & \mathbb{E}(\text{information score}(x_i) | t_k) \\ &= \sum_j p(t_j | t_k) \int p(\omega | t_k, t_j) \log \frac{p(\omega | x_i, t_j)}{p(\omega | t_j)} d\omega \\ &= \underbrace{\sum_j p(t_j | t_k) \int p(\omega | t_k, t_j) \log \frac{p(\omega | t_k, t_j)}{p(\omega | t_j)} d\omega}_{\text{truth-telling equilibrium}} + \underbrace{\sum_j p(t_j | t_k) \int p(\omega | t_k, t_j) \log \frac{p(\omega | x_i, t_j)}{p(\omega | t_k, t_j)} d\omega}_{-D_{\text{KL}}(p(\omega | t_k, t_j) \| p(\omega | x_i, t_j))} \leq 0 \end{aligned}$$

Q.E.D.

Finally, in the truth-telling equilibrium, the BTS is a zero-sum game, and the score measures how common an answer is compared to the prior expected frequency.

**Proposition 3.3** (Zero-Sum Game). In the truth-telling equilibrium, the scores add up to zero in the limit:

$$\sum_i p(t_i | \omega) \text{flask}(x_i, p(t | t_i)) = 0.$$

Furthermore,

$$\text{flask}(x_i, p(t | t_i)) = \log \frac{\omega_i}{\mathbb{E}_{p(\omega)}(\omega_i)} + \text{constant}(\omega). \quad \blacktriangleleft$$

*Proof.* This follows from a simple calculation:

$$\text{flask}(x_i, p(t | t_i)) = \sum_j p(t_j | \omega) \log \frac{p(\omega | t_i)}{p(\omega | t_j)} = \log p(\omega | t_i) - \sum_j p(t_j | \omega) \log p(\omega | t_j).$$

Q.E.D.

## 4 Practical Aspects

Compared to other scoring techniques, the Bayesian truth serum has a number of benefits [Pre04]. First, the respondents can honestly be instructed be that answering honestly is their best strategy. The BTS score, for example, does not encourage respondents to answer the most common answer, because the most common answer will also be predicted to occur most commonly. Second, there is no need to limit questions to those where empirically estimated prior and conditional probabilities are available, because the BTS requires no such probabilities. Third, the BTS makes no assumptions about the population, which means that the same survey can be applied to different populations.

There are two generic ways in which the Bayesian truth serum might fail [Pre04]. First, the BTS works poorly if the prior  $p(\omega)$  is sharp. In this case, the “sample of one” that is one’s own opinion will only minimally update their belief about the population. The difference in score between honest and dishonest answers will then be minimal. For example, a person’s gender will negligibly impact their belief about the proportion of men and women in the population. Second, respondents might honestly answer similarly, but form different posteriors because they differ in another characteristic:  $t^n = t^m$ , but  $p(\omega | t^n) \neq p(\omega | t^m)$ . The assumption of exchangeability (Assumption 3.1) then does not hold, and the BTS score might not correlate with honesty anymore. For example, a person with nonstandard political views might interpret their liking of a candidate as evidence that others will not [Pre04]. The remedy here is to expand the survey to reveal this characteristic, e.g. by including a question about the respondent’s political view.

## 5 Summary

The Bayesian truth serum assigns high scores to answers that are more common than commonly predicted. A meta-rational Bayesian agent should conclude that their opinion has the highest probability of being more common than commonly predicted, because unlike the common prediction their best guess about the distribution of answers is informed by their own opinion. Therefore, if respondents are rewarded according to their BTS scores, they should believe that answering honestly has the highest probability of earning a high reward.

## References

- [Pre04] Dražen Prelec. “A Bayesian Truth Serum for Subjective Data”. In: *Science* 306.5695 (2004), pp. 462–466. DOI: 10.1126/science.1102081. URL: <http://science.sciencemag.org/content/306/5695/462> (cit. on pp. 1–3, 6).