Bayesian Plinko

Study notes

A. L. S. Filipowicz
<***@***>

P.G.L. Porta Mana
<pgl@portamana.org>

Draft of 17 April 2018 (first drafted 14 April 2018)

How does a Bayesian robot do at Plinko, using an infinitely exchangeable model and borrowing a human participant's prior?

We are human, after all
Much in common, after all
(Daft Punk 2005a)

If people always felt obliged to back their opinions
when challenged, we would be spared a few of the
"certain" predictions that are so freely made.
(Good 1950)

1 Introductory remarks on the methods

The experimental setup is open to a huge number of analyses and interpretations on the participants' part, inspired by past experience. As a participant, we can surmise that there's a connection between different trials, some sort of 'constant mechanism' at some level. Or we can surmise that there's no such connection, hence observation of past trials doesn't say anything about the next one. Or we can surmise that the next trial is influenced by the participant's own bar-height assignments. And many other hypotheses. We can also entertain all these hypotheses at the same time, and shift from one to another during the experiment. For example, if we suddenly wondered whether the computer program is actually using our bar distribution, we could suddenly move all probability to a slot at the edge and check if this seems to influence the next outcome (cf. participant 31).

The same goes for the choice of initial distribution. As a participant, we can say 'alright, there are 40 slots', and just give a uniform distributions to the 40 possibilities. Or we can consider the pyramidal mechanism of the game, which leads to a binomial distribution. Or we can consider that this is a computerized version of the game. The computer could simulate the physics of the actual game; but the image of the mechanism could also be just for show, the computer being programmed to distribute the outcomes according to a predetermined, completely arbitrary distribution. From this point of view we could again decide to assign a uniform distribution.

✠ L's opinion Should a participant's 'performance' be assessed against a probability distribution that supposedly lies behind a computer program? This is a tricky question. Let's assume an operational point of view. We have a computer program in which we specify two parameters: 'mean' and 'standard deviation'. The program then generates a sequence of, say, 100 numbers between 1 and N. Finally we ask the participant to guess a distribution that supposedly lies behind these 100 outcomes; suppose we also ask him to choose a normal distribution. This is equivalent to ask him to guess the two parameters we gave to the computer program. But here's the catch: there are two parameter values under which the data are more likely than under the values we chose: the *empirical* mean and standard deviation. From this point of view it's unfair to assess the participant's performance against the values we chose, given that there are other values under which the data are more likely. It's as if we made a carpenter – who turns out to be a good cook – prepare a delicious dish, then asked a person to guess who prepared it, and then assessed their answer, 'a chef', as bad because they didn't guess that it was a carpenter ✠ L: alright, I must find a better example.

The assessment problem just described is worse if the data are few compared with the number of possible outcomes, because the empirical mean and standard deviation will then usually be very different from our parameter choices. It is even worse if we ask the participant to choose not among normal distributions but among all possible distributions. There will always be a distribution under which the data are be more likely than under our normal.

After all, the computer program is generating the data deterministically. Shouldn't we assess the participant according to the prediction of the *exact* outcome sequence, then?

To avoid this paradox we can find other assessment measures. For example, we can set up several inference robots, constructed according to different assumptions, and compare a participant's response to the robots'.

De Finetti introduced the notion of exchangeability, to be presented in the next section, in order to avoid thinking in terms of 'mechanisms' or of 'true unknown distributions' (cf. Good 1965 ch. 3). He characterized a statistical model in terms of *features of the predictive distribution* – for example, that it be invariant under exchanges of the outcomes. This point of view was taken up by many other statisticians and has been used to characterize many other statistical models, like Markov ones for example (Freedman 1962; Diaconis et al. 1980; Zaman 1984; Fortini et al. 2002; 2014).

2 The Bayesian robot

In the context of these notes and of the Plinko experiments (Filipowicz et al. 2014; 2016) we call 'model' any set of assumptions that allows us to assign a probability to a new observation, given a number of observations of a similar kind. Denote such assumptions by a proposition M – a proposition surely very difficult to express in writing. Denote the proposition 'The outcome of the ith observation is d' by D^i_d, with $d \in \{1, \dots N\}$. Then M allows us to give a numeric value to

$$\mathrm{P}(D^{m+1}_{d_{m+1}} | D^1_{d_1} \wedge D^2_{d_2} \wedge \cdots \wedge D^m_{d_m} \wedge M), \tag{1}$$

We will abbreviate logical conjunction '$\wedge$' with a comma, for simplicity. Our statistical terminology and notation follow ISO standards (ISO 2009; 2006) otherwise.

We shall consider a robot who uses an *infinitely exchangeable* model. This kind of models, introduced by de Finetti (1930; 1937; Heath et al. 1976) and described in detail in Bernardo et al. (2000 § 4.2), is determined by the following assumption of *infinite exchangeability*: the joint distribution for any number of observations is symmetric with respect to their order; that is, the order of the observations is irrelevant for inferential purposes. Distributions for different number of observations must of course be consistent with one another through marginalization. Infinite exchangeability may in turn be motivated by other specific assumptions, but the details of these are irrelevant for the mathematical form of this model.

Infinite exchangeability determines this form of the probability above:

$$\mathrm{P}(D^1_{d_1}, D^2_{d_2}, \dots, D^m_{d_m} | M) = \int_\Delta \left(\prod_{i=1}^m q_{d_i} \right) \mathrm{p}(\boldsymbol{q} | M)\, \mathrm{d}\boldsymbol{q}, \tag{2}$$

where $\boldsymbol{q}$ is a normalized N-tuple of positive numbers: $\Delta \coloneqq \{\boldsymbol{q} \in \mathbf{R}^N \mid q_i \geqslant 0, \sum_{i=1}^N q_i = 1\}$. This N-tuple can be thought of the relative, long-run frequencies of the possible outcomes[1], and $\mathrm{p}(\boldsymbol{q}\,|\,M)\,\mathrm{d}\boldsymbol{q}$ as their probability density. From this point of view it is as if the robot first assumes to know the long-run frequencies of the different outcomes and, not knowing their particular order in the observation, assigns to the occurrence of each a probability proportional to its frequency: this is the term $\prod_{i=1}^{m} q_{d_i}$ in the integral. Then, not being sure about the long-run frequencies, the robot assign to them the density $\mathrm{p}(\boldsymbol{q}\,|\,M)\,\mathrm{d}\boldsymbol{q}$ – which is determined by additional assumptions besides exchangeability.

As an explicit example, say with $N = 40$,

$$\mathrm{P}(D^1_{37}, D^2_{6}, D^3_{25}, D^4_{37}\,|\,M) = \int_{\Delta} q_6\, q_{25}\, {q_{37}}^2\, \mathrm{p}(\boldsymbol{q}\,|\,M)\,\mathrm{d}\boldsymbol{q}. \tag{3}$$

In the following we omit the integration domain Δ.

From Bayes's theorem we obtain the expression for the predictive probability (1) of an infinite exchangeable model:

$$\mathrm{P}(D^{m+1}_{d_{m+1}}\,|\,D^1_{d_1}, \dots, D^m_{d_m}, M) = \int q_{d_{m+1}} \mathrm{p}(\boldsymbol{q}\,|\,D^1_{d_1}, \dots, D^m_{d_m}, M)\,\mathrm{d}\boldsymbol{q}, \tag{4a}$$

$$\mathrm{p}(\boldsymbol{q}\,|\,D^1_{d_1}, \dots, D^m_{d_m}, M) = \frac{\bigl(\prod_{i=1}^{m} q_{d_i}\bigr)\,\mathrm{p}(\boldsymbol{q}\,|\,M)}{\int \bigl(\prod_{i=1}^{m} q'_{d_i}\bigr)\,\mathrm{p}(\boldsymbol{q}'\,|\,M)\,\mathrm{d}\boldsymbol{q}'}. \tag{4b}$$

Continuing our numeric example (3) this could be

$$\mathrm{P}(D^5_{6}\,|\,D^1_{37}, D^2_{6}, D^3_{25}, D^4_{3}, M) = \int q_6\, \mathrm{p}(\boldsymbol{q}\,|\,D^1_{37}, D^2_{6}, D^3_{25}, D^4_{3}, M)\,\mathrm{d}\boldsymbol{q}, \tag{5a}$$

$$\mathrm{p}(\boldsymbol{q}\,|\,D^1_{37}, D^2_{6}, D^3_{25}, D^4_{3}, M) = \frac{q_6\, q_{25}\, {q_{37}}^2\, \mathrm{p}(\boldsymbol{q}\,|\,M)}{\int q_6\, q_{25}\, {q_{37}}^2\, \mathrm{p}(\boldsymbol{q}'\,|\,M)\,\mathrm{d}\boldsymbol{q}'}. \tag{5b}$$

Formula (4) tell us how our robot would update its predictive probabilities at each new observation of a Plinko outcome.

3 General remarks on the robot's behaviour

The exchangeable-model formula (4) leads to some characteristic features of the robot's beliefs and of their evolution:

[1] 'But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.' (Keynes 2013 § 3.I, p. 65)

- The robot's predictions can be interpreted in several ways. One is this: the robot believes that there's 'something' constant in all trials; loosely speaking, a 'constant mechanism'. Another, maybe preferable, is this: the robot has memory of past observations, but not of their order; any trends are therefore invisible to it.
- As data accumulate, the robot's probabilities for the next outcome approach the observed frequencies. Such approach happens independently of the form of the prior $\mathrm{p}(\boldsymbol{q}|M)\,\mathrm{d}\boldsymbol{q}$ – unless the latter is zero in peculiar regions of the integration domain – but the prior determines the celerity of the approach. A prior heavily peaked on a frequency $\boldsymbol{q}'$ will require a lot of data to move the predictions to a very different frequency $\boldsymbol{q}$.
- As data D accumulate, the updated density $\mathrm{p}(\boldsymbol{q}|D,M)\,\mathrm{d}\boldsymbol{q}$ will become more and more peaked at the N-tuple of observed frequencies.
- Suppose that we first have a long sequence of observations concentrating around frequencies $\boldsymbol{q}$ – say, a very long sequence of 1s in a row – and then a shift to other frequencies $\boldsymbol{q}'$ – say, suddenly 2s only appear. After the shift, the predictive probabilities will eventually become peaked around the new frequencies, but the shift in the peaks will take a larger number of observations around the new frequencies than the number around the old frequencies.

4 Initial prior

The shape of the initial prior heavily determines the predictions in the first observations, so it must be chosen with care. The Plinko data tell us the initial predictive probabilities of the participants,

$$\mathrm{p}(D^1_k|M) \equiv \int q_k\ \mathrm{p}(\boldsymbol{q}|M)\,\mathrm{d}\boldsymbol{q}, \tag{6}$$

but not their prior $\mathrm{p}(\boldsymbol{q}|M)\,\mathrm{d}\boldsymbol{q}$.

As a first exploration we consider a *Johnson-Dirichlet* prior, proportional to a monomial $\prod_i \boldsymbol{q}_i^{\,x_i}$ for some values of x_i:

$$\mathrm{p}(\boldsymbol{q}|M_\mathrm{J}) = \frac{\Gamma(\Lambda)}{\prod_i \Gamma(\Lambda\,\nu_i)} \prod_{i=1}^{N} \boldsymbol{q}_i^{\,\Lambda\,\nu_i - 1}, \qquad \Lambda > 0, \nu \in \Delta. \tag{7}$$

This prior is determined by the additional assumption – call it M_J – that that the frequencies of other outcomes are irrelevant for predicting a particular one:

$$\mathrm{P}(D_k^{m+1} | N\boldsymbol{f}, M_\mathrm{J}) = \mathrm{P}(D_k^{m+1} | N f_k, M_\mathrm{J}) \qquad k \in \{1, \dots, N\}, \tag{8}$$

where $\boldsymbol{f}$ is the N-tuple of observed relative frequencies. This assumption is called 'sufficientness' (Johnson 1924; 1932; Good 1965 ch. 4; Zabell 1982; Jaynes 1996). This is a conjugate prior (DeGroot 2004 ch. 9; Diaconis et al. 1979) and it has two convenient properties: it updates to a density of the same mathematical form, and its corresponding predictive distribution can be calculated analytically using the formula

$$\int_\Delta \prod_{i=1}^N {q_i}^{x_i - 1}\,\mathrm{d}\boldsymbol{q} = \frac{\prod_i \Gamma(x_i)}{\Gamma\big(\sum_i x_i\big)}. \tag{9}$$

We obtain for the initial distribution:

$$\mathrm{P}(D_k^1 | M_\mathrm{J}) = \int q_k \;\mathrm{p}(\boldsymbol{q} | M_\mathrm{J})\,\mathrm{d}\boldsymbol{q} = \nu_k, \tag{10}$$

and for the updated density:

$$\mathrm{p}(\boldsymbol{q} | D_{d_1}^1, \dots, D_{d_m}^m, M_\mathrm{J}) = \frac{\Gamma(\varLambda')}{\prod_i \Gamma(\varLambda' \nu_i')} \prod_{i=1}^N \boldsymbol{q}_i^{\varLambda' \nu_i' - 1}$$
$$\text{with} \quad \varLambda' = \varLambda + N, \quad \boldsymbol{\nu}' = \frac{\varLambda\boldsymbol{\nu} + N\boldsymbol{f}}{\varLambda + N}. \tag{11}$$

Formula (10) says that the Johnson-Dirichlet prior can produce any initial probabilities assigned by the participants, just by equalling the parameters $\boldsymbol{\nu}$ to them. The parameter $\varLambda$ is left arbitrary. We can call it the *stubbornness* of the robot, Here's the reason.

Suppose that after some observations the predictive distribution is $\boldsymbol{\nu}$, and that the next outcome is k. Then the probability for slot k is updated to $\frac{\varLambda\nu_k + 1}{\varLambda + 1}$, and that for all other slots j to $\frac{\varLambda\nu_j}{\varLambda+1}$. This update corresponds to a participant's raising the bar assignment under slot k, leaving the others untouched, and/or lowering the bar assignments for *all* other slots by the same proportion. The parameter $\varLambda$ is increased by 1. The larger $\varLambda$, the more reluctant the robot is in revising its guesses in the light of new observations. The update formula (11) says that the robot behaves as if it had already made $\varLambda$ observations with outcome frequencies $\boldsymbol{\nu}$.

5 Examples

Let's choose a participant, and use formula (10) to choose the $\boldsymbol{\nu}$ parameters of the robot's prior, equating it to the initial predictive distribution of the participant. Let's set a value for the robot's stubbornness $\varLambda$, and check how the robot updates its predictive distribution, using formula (11), while observing the same outcomes as the participant.

Figure 1 shows the means and standard deviations of the sequence such predictive distributions, for participant 12 and a robot with stubbornness $\varLambda = 0.1$. This low value makes the robot give great consideration to the first outcomes, as the initial variability in the figure shows. The program generating the outcomes had a change in standard deviation, shifting to a narrower distribution at trial 101. The robot adapted to this change very slowly.

Figure 2 is analogous to fig. 1 but for a robot with stubbornness $\varLambda = 50$. This robot is even more slow to adapt to the narrowing in the standard deviation of the generated outcomes.

Figures 3 and 4 show the same for participant 30. The change in standard deviation was from narrow to large in this case.

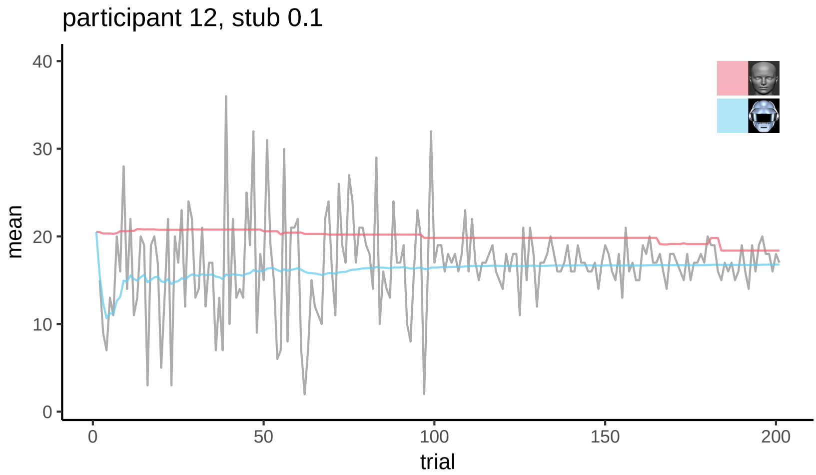
The robot with low stubbornness seems to adapt to the widening of the outcome outputs faster than it had for the narrowing of the previous case: the change in the slope of the robot's standard-deviation curve seems steeper in fig. 3 than in 1.

If we look at the sequence of outcomes of figs 1 or 3, we perceive that something changed around trial 100. If we could plot these outcomes while they are generated, we would likely notice the change by around trial 25. Our robot, however, can't detect this change for the reasons explained in § 3; any outcomes from narrow or wide generating processes are mingled in the robot's memory.

Only non-exchangeable or hierarchic models can exhibit a short evidence memory and be capable of believing that the underlying 'mechanism' has changed.

Some conclusions can be drawn from the properties of our model and from the examples:

- Participants who have great inertia against updating their predictions in view of the observations are *not* necessarily behaving at variance with the probability calculus. The latter says that they can be as stubborn as they please: larger $\varLambda$. If we judge such inertia as irrational, our judgement cannot be based on such a simple model; possibly it's based



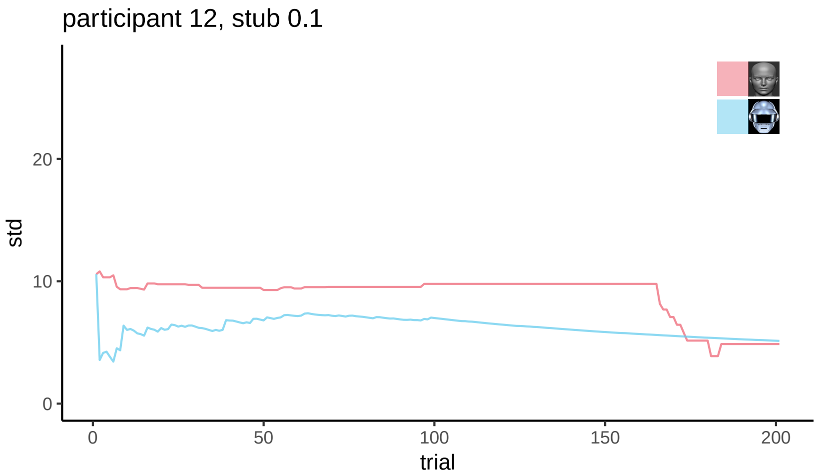


Figure 1 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubborness $\Lambda = 0.1$

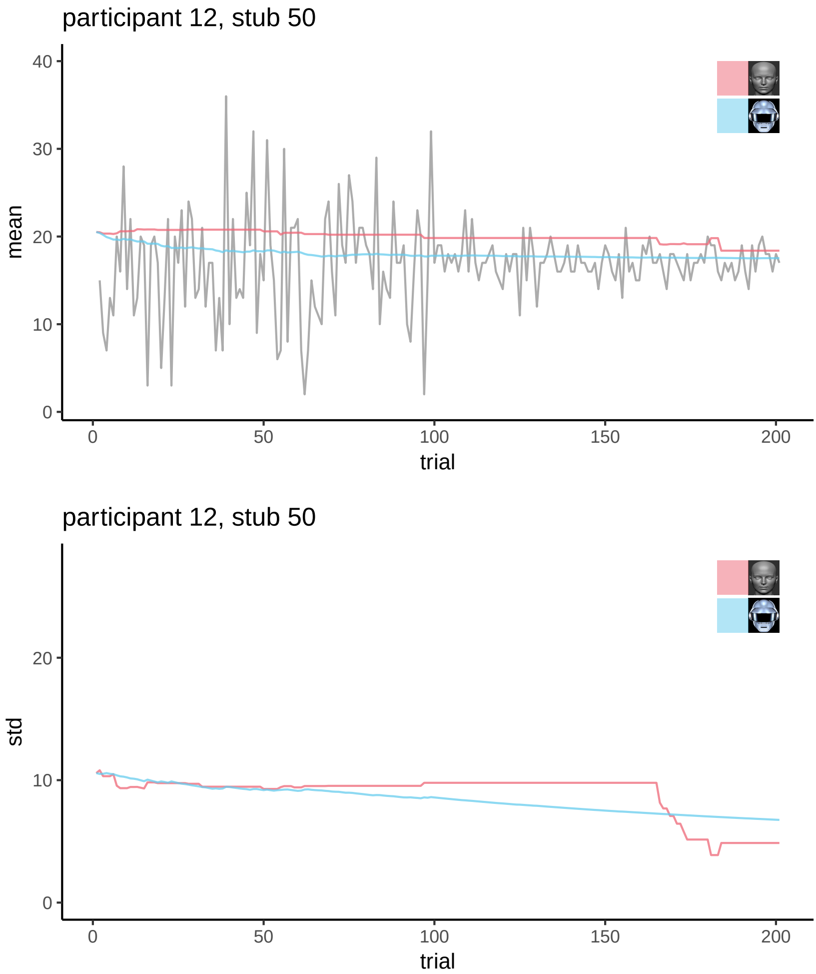
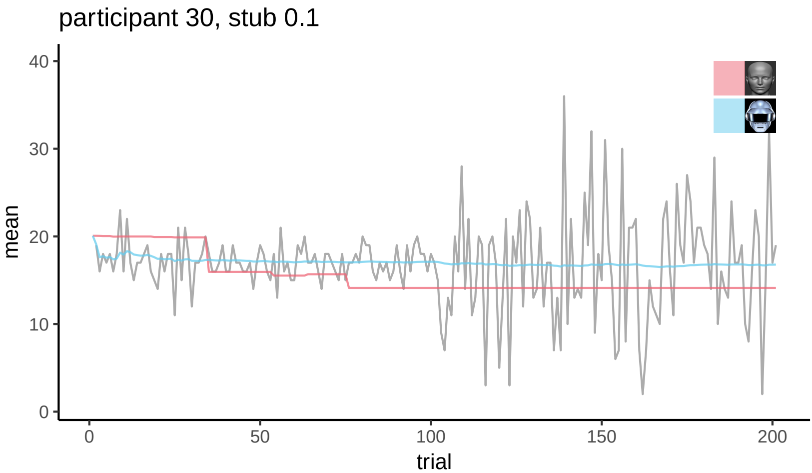


Figure 2 Comparison of the means and standard deviations of the predictive distributions of participant 12 and of a robot with stubborness $\Lambda = 50$



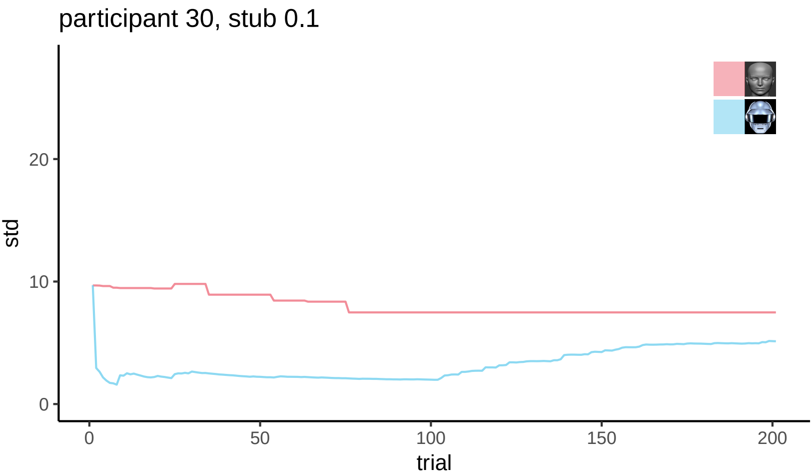
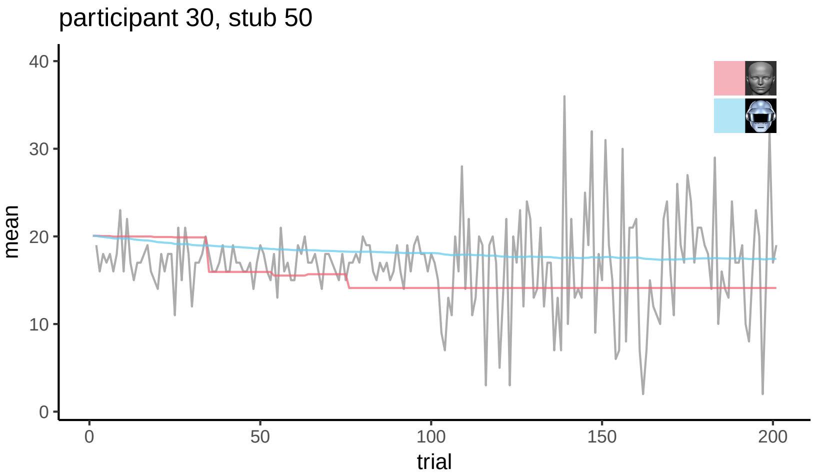


Figure 3 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubborness $\Lambda = 0.1$



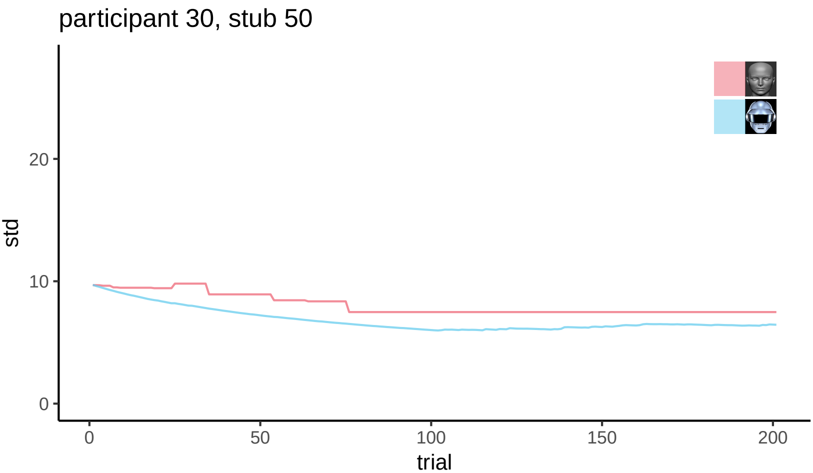


Figure 4 Comparison of the means and standard deviations of the predictive distributions of participant 30 and of a robot with stubborness $\Lambda = 50$

on a hierarchic model where Λ is given a probability that depends on past experiences.

- The slots have a specific physical order, and from the way the ball falls into them it seems reasonable to assume that updates to the probability for one slot should affect those for nearby slots. The Johnson-Dirichlet model does not take this into account.
- A participant who, after observing outcome k, raises the bar under that slot *and nearby bars* is therefore not acting according to a Johnson-Dirichlet exchangeable model.
- Infinitely exchangeable priors are incapable of quickly adapting to changes in the empirical statistics of the outcomes.

Bibliography

('de X' is listed under D, 'van X' under V, and so on, regardless of national conventions.)

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

Daft Punk (2005a): *Human after all*. In: Daft Punk (2005b).

— (2005b): *Human After All*. (Virgin, worldwide).

de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**^{5}, 86–133. http://www.brunodefinetti.it/Opere.htm.

— (1937): *La prévision : ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**^{1}, 1–68. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

DeGroot, M. H. (2004): *optimal statistical decisions*, reprint. (Wiley, New York).

Diaconis, P., Freedman, D. (1980): *De Finetti's theorem for Markov chains*. Ann. Prob. **8**^{1}, 115–130.

Diaconis, P., Ylvisaker, D. (1979): *Conjugate priors for exponential families*. Ann. Stat. **7**^{2}, 269–281.

Filipowicz, A., Valadao, D., Anderson, B., Danckert, J. (2014): *Measuring the influence of prior beliefs on probabilistic estimations*. Proc. Annu. Meet. Cogn. Sci. Soc. **36**, 2198–2203.

— (2016): *Rejecting outliers: surprising changes do not always improve belief updating*. Decision ****, **.

Fortini, S., Ladelli, L., Petris, G., Regazzini, E. (2002): *On mixtures of distributions of Markov chains*. Stochastic Processes Appl. **100**$^{1-2}$, 147–165.

Fortini, S., Petrone, S. (2014): *Predictive characterization of mixtures of Markov chains*. arXiv: 1406.5421.

Freedman, D. A. (1962): *Mixtures of Markov processes*. Ann. Math. Stat. **33**^{1}, 114–118.

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

— (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. (MIT Press, Cambridge, USA).

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. American Statistician **30**^{4}, 188–189.

ISO (2006): *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization.

ISO (2009): *ISO 80000:2009: Quantities and units*. International Organization for Standardization. First publ. 1993.

Jaynes, E. T. (1996): *Monkeys, kangaroos, and N*. http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03john, https://archive.org/details/johnsonslogic03johnuoft.

— (1932): *Probability: the deductive and inductive problems*. Mind **41**^{164}, 409–423. With some notes and an appendix by R. B. Braithwaite.

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of second ed. (Cambridge University Press, Cambridge). First publ. 1923.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.

Zabell, S. L. (1982): *W. E. Johnson's "sufficientness" postulate*. Ann. Stat. **10**^{4}, 1090–1099. Repr. in Zabell (2005 pp. 84–95).

— (2005): *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. (Cambridge University Press, Cambridge).

Zaman, A. (1984): *Urn models for Markov exchangeability*. Ann. Prob. **12**^{1}, 223–229.