TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 10 (2018) 192–208
Copyright © 2018 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12319

This article is part of the topic "Best of Papers from the 2017 International Conference on Cognitive Modeling," William G. Kennedy, Marieke K. Van Vugt, Adrian P. Banks (Topic Editor). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview.

Probability Theory Plus Noise: Descriptive Estimation and Inferential Judgment

Fintan Costello,[a] Paul Watts[b]

[a]*School of Computer Science and Informatics, University College Dublin*
[b]*Department of Theoretical Physics, National University of Ireland*

Received 15 September 2017; received in revised form 3 November 2017; accepted 24 November 2017

Abstract

We describe a computational model of two central aspects of people's probabilistic reasoning: descriptive probability estimation and inferential probability judgment. This model assumes that people's reasoning follows standard frequentist probability theory, but it is subject to random noise. This random noise has a regressive effect in descriptive probability estimation, moving probability estimates away from normative probabilities and toward the center of the probability scale. This random noise has an anti-regressive effect in inferential judgement, however. These regressive and anti-regressive effects explain various reliable and systematic biases seen in people's descriptive probability estimation and inferential probability judgment. This model predicts that these contrary effects will tend to cancel out in tasks that involve both descriptive estimation and inferential judgement, leading to unbiased responses in those tasks. We test this model by applying it to one such task, described by Gallistel et al. (2014). Participants' median responses in this task were unbiased, agreeing with normative probability theory over the full range of responses. Our model captures the pattern of unbiased responses in this task, while simultaneously explaining systematic biases away from normatively correct probabilities seen in other tasks.

Keywords: Probability estimation; Inferential judgment; Biases in reasoning

Correspondence should be sent to Fintan Costello, School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland. E-mail: fintan.costello@ucd.ie

1. Introduction

We live in a world of nonstationary stochastic processes, where events occur with some associated probability, and this probability itself changes unpredictably over time. To make successful predictions about event occurrence in such a world, we must use two distinct types of probabilistic reasoning: descriptive probability estimation (given the sample of events we have seen recently, what is the current underlying population probability of *A*?) and inferential probability judgment (given our current estimate for the probability of *A*, is the current sample of events consistent with that probability? Or should we infer that the underlying probability of *A* has changed?). Our aim in this paper is to present a computational model of these two interacting components of probabilistic reasoning.

One revealing aspect of human probabilistic reasoning is the reliable occurrence of a number of systematic biases in people's judgement of probability: biases such as conservatism (Erev, Wallsten, & Budescu, 1994), subadditivity (Tversky & Koehler, 1994), and the conjunction fallacy (Tversky & Kahneman, 1983). The model we present was originally developed to explain these biases in terms of the effect of random noise in reasoning (see Costello & Watts, 2014). Here we extend this model to inferential probability judgment and show that this model explains patterns of bias seen in such judgment. This model predicts that, in situations involving both forms of reasoning, these effects will cancel out, leaving subjective probability estimates that tend to agree with the normatively correct values with no detectable systematic bias. Such agreement is seen in recent studies of probability estimation for nonstationary stochastic processes by Gallistel, Krishan, Liu, Miller, and Latham (2014). We demonstrate the model by applying it to Gallistel et al.'s study.

2. The probability theory plus noise model

Our model assumes that people's probability judgments are produced by a mechanism that is fundamentally rational but perturbed by purely random noise or error, which causes systematic biasing effects. We take $P(A)$ to represent the "true" probability of event *A* (i.e., the proportion of items in memory that represent *A*). We take $p_*(A)$ to represent an individual estimate of the probability of event *A*, and take $P_*(A) = \langle p_*(A) \rangle$ to represent the expectation value or mean of these estimates for *A*: This is the value we would expect to get if we averaged an infinite number of individual estimates for $p_*(A)$. In standard probability theory, the probability of some event *A* is estimated by drawing a random sample of events, counting the number of those events that are instances of *A*, and dividing by the sample size. The expected value of these estimates is $P(A)$, the probability of *A*. We assume that people estimate the probability of some event *A* in exactly this way: randomly sampling events from memory, counting the number of instances of *A*, and dividing by the sample size. If this counting process was error-free, people's estimates would have an expected value of $P(A)$. Human memory, however, is subject to various forms of random error or noise. To reflect this, we assume events have some chance

$d < 0.5$ of randomly being counted incorrectly: There is a chance d that a $\neg A$ (*not A*) event will be incorrectly counted as A, and the same chance d that an A event will be incorrectly counted as $\neg A$. Note that this single noise term d is intended to cover the influence of many different potential sources of error: noise in processing, noise in item recall, noise due to contextual effects, and so on. For mathematical tractability we collapse all these sources into a single noise term, that is, by assumption, symmetrical and not in itself subject to any bias: There is no more noise associated with A than with $\neg A$.

Given this form of noise, a randomly sampled event will be counted as A if the event truly is A and is counted correctly (with a probability $(1-d)P(A)$, since $P(A)$ events are truly A and events have a $1-d$ chance of being counted correctly), or if the event is truly $\neg A$ and is counted incorrectly as A (with a probability $(1-P(A))d$, since $1-P(A)$ events are truly $\neg A$, and events have a d chance of being counted incorrectly). Summing the probabilities of these two mutually exclusive situations, we see that the chance of a randomly sampled event being counted as A is

$$P(counted\, as\, A) = (1-d)P(A) + (1-P(A))d = (1-2d)P(A) + d$$

and so the expected average value for a noisy probability estimate of $P(A)$ is

$$P_*(A) = \langle p_*(A) \rangle = P(counted\, as\, A) = (1-2d)P(A) + d \tag{1}$$

with individual estimates varying independently around this expected value (see Appendix A for a detailed derivation of this result). This average is systematically biased away from the "true" probability $P(A)$, such that estimates will tend to be greater than $P(A)$ when $P(A) < 0.5$, and they will tend to be less than $P(A)$ when $P(A) > 0.5$: a pattern of systematic regression toward 0.5, the center of the probability scale.

Since this model assumes that the probability $P(A)$ is estimated by retrieving a random sample of episodes from memory and counting the number of A's, it may seem that the model is only able to give probability estimates for events that have already been seen. This view depends on a conception of memory as being nothing but a store of recorded events. We can, however, take an alternative conception of memory as a constructive process that can generate representations of events even if those specific events have not previously been seen (events that might occur in the future, for example). Support for this view comes from evidence that remembering past events and imagining future events are very similar cognitive processes (see, e.g., Schacter, 2012).

If we take this "constructive" or "simulation" view of memory, then our model can apply to probability estimates for all forms of event, whether previously seen or completely novel. In this view, estimating the probability of some event A happening in the future, for example, would involve generating or imagining a number of possible future outcomes and counting the proportion that contained event A (subject to random error in counting). We follow this constructive view of memory, and so assume that this model applies to all forms of event, both previously seen and completely novel.

2.1. Conservatism in probability estimation

Regression, in this model, explains a number of observed patterns of bias in people's descriptive probability estimates. One such pattern is a bias we refer to as "conservatism in estimation." This is the finding that people's estimates for the probability of an event tend to be systematically biased away from the true probability in a characteristic way: The closer $P(A)$ is to 0, the more likely it is that a person's estimate is greater than $P(A)$, while the closer $P(A)$ is to 1, the more likely it is that the person's estimate is less than $P(A)$. Differences between true and estimated probabilities are low when $P(A)$ is close to 0.5 and increase as $P(A)$ approaches 0 or 1. Erev et al. (1994), for example, found this pattern in a study where participants played a video game and then estimated the probability of different events in that game: Participants reliably overestimated the probability of low-probability events and underestimated that of high-probability events. Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) found this pattern in a series of studies where participants estimated the probability of different causes of death: Participants reliably overestimated low–frequency causes and underestimated high-frequency causes. Teigen (1973) found this pattern in a study where participants estimated the frequency of occurrence of a given symbol in a presented sequence: Participants reliably overestimated the occurrence of rare symbols and underestimated the occurrence of frequent symbols (for similar results, see, e.g., Erlick, 1964; Poulton, 1973).

This pattern occurs as a straightforward consequence of random variation in our model. As we saw in Eq. 1, $P_*(A)$ deviates from $P(A)$ in a way that systematically depends on $P(A)$. If $P(A) = 0.5$, this deviation will be 0. If $P(A) < 0.5$, then since d cannot be negative we have $P_*(A) > P(A)$, with the difference increasing as $P(A)$ approaches 0. Since estimates $p_*(A)$ are distributed around $P_*(A)$, this means that $p_*(A)$ will tend to be greater than $P(A)$, with the tendency increasing as $P(A)$ approaches 0. Similarly if $P(A) > 0.5$, then $P_*(A) < P(A)$ and estimates $p_*(A)$ will tend to be less than $P(A)$, with the tendency increasing as $P(A)$ approaches 1.[1]

2.2. Cancellation of bias

This model also makes a number of novel predictions about patterns of agreement with probability theory in people's judgment. Probability theory requires that certain identities must hold for probability estimates involving any pair of events A and B. One such identity is the addition law, which requires that

$$P(A) + P(B) - P(A \wedge B) - P(A \vee B) = 0 \tag{2}$$

If we substitute the expected values from Eq. 1 into the addition law identity, for example, we get an expected value of

$$\begin{aligned}&P_*(A)+P_*(B)-P_*(A\wedge B)-P_*(A\vee B)\\&=(1-2d)P(A)+d+(1-2d)P(B)+d\\&\quad-(1-2d)P(A\wedge B)-d-(1-2d)P(A\vee B)-d=0\end{aligned}$$

with the regressive effects of noise in estimates for each term being, on average, cancelled out. Our model thus predicts that this expression will have a value of 0, on average, in people's probability judgments just as required by standard probability theory. Exactly this pattern of agreement is seen in experimental results (Costello & Mathison, 2014; Costello & Watts, 2014, 2016, 2017; Fisher & Wolfe, 2014).

2.3. Inferential probability judgment

Equation 1 gives the expected value of a descriptive probability $p_*(A)$ in this model, produced when a reasoner sees a sample containing some instances of event A and then estimates the underlying probability parameter $P(A)$ describing the population from which that sample was drawn. We now consider the estimation of an inverse or inferential probability $P(x,n|p)$ in this model: the probability of seeing a sample of n items that contains exactly x A's, given that the sample was drawn from a population where $P(A) = p$. Frequentist probability theory provides a normative mechanism for estimating such inferential probabilities: to estimate $P(x,n|p)$, draw a series of random samples, each of size n, from a population where $P(A) = p$ and count the proportion of samples that contain exactly x instances of A. This proportion gives an estimate of $P(x, n|p)$, the probability of the observed sample occurring in a population with $P(A) = p$: the lower this estimate, the less likely it is that the observed sample came from such a population. The expected value of this estimate is given by the binomial probability function

$$P(x,n|p)=\binom{n}{x}p^x(1-p)^{n-x} \tag{3}$$

In our model we assume that people estimate inferential probabilities just as in frequentist probability theory: by drawing a series of random samples of size n from a (simulated) population where $P(A) = p$ and counting the proportion of samples that contain exactly x instances of A. We assume that this counting process is subject to random error; that the count of occurrences of A in a sample is subject to random noise at a rate d (there is d chance that an instance of A in a given sample will be counted as $\neg A$, and d chance that an instance of $\neg A$ in a given sample will be counted as A). From Eq. 1, with a noise rate d, the chance of an instance in a sample being counted as A is equal to P (*counted as* A) = $(1-2d)p$,+ d. This means that the probability of getting a sample which is counted as containing x A's, given a population probability of p and a noise rate of d is given by

$$\begin{aligned}\langle p_*(x,n|p)\rangle &= \binom{n}{x} P(counted\,as\,A)^x (1-P(counted\,as\,A))^{n-x}\\ &= \binom{n}{x}((1-2d)p+d)^x((1-2d)(1-p)+d)^{n-x}\end{aligned} \tag{4}$$

This is simply the normative binomial probability function (Eq. 3), but with the probability that a randomly sampled item *is A* being replaced by the probability that a randomly sampled item is *counted as A*. This expression $\langle p_*(x,n|p)\rangle$ gives the average noisy inferential probability that a sample of n items containing x A's came from a population where $P(A) = p$ (given a noise rate d).

Note that the inferential probabilities given in Eqs. 3 and 4 are both binomially distributed with common terms x and n. If we take p_e to be our current estimate of the probability of A in the population in question, this means that, for any given values of x and n, the associated noisy inferential probability $\langle p_*(x,n|p_e)\rangle$ is exactly equal to another normatively correct inferential probability $P(x,n|p)$ when

$$((1-2d)p_e+d)^x((1-2d)(1-p_e)+d)^{n-x} = p^x(1-p)^{n-x}$$

When $d \leq p \leq 1{-}d$, this equality holds for all values of n and x when

$$(1-2d)p_e + d = p$$

or equivalently when

$$p_e = \frac{p-d}{1-2d}$$

This expression is "anti-regressive," giving values for p_e that are closer to the boundaries 0 and 1 than values of p: p_e is greater than p when $p > .5$, and less than p when $p < .5$. In other words, while noise in descriptive probability estimation (estimating the probability of A, given a sample) produces an average estimate that is regressive relative to p, noise in inferential probability estimation (estimating the probability of a sample, given $P(A) = p$) produces an inferential probability that is anti-regressive relative to p.

2.4. Conservatism in inferential judgement

Experimental studies typically investigate inferential probability estimation indirectly, using the related concept of relative probability. These studies involve describing two populations containing complementary proportions of two different types of event. Participants are told that a population has been picked at random, and they are then shown a sample of events drawn from the selected population and asked to assess the probability that the sample came from one population rather than the other. Typically these populations are "book-bags" containing poker chips, with one bag containing, for example, 70% red chips and 30% black (this is the “red bag”), and the other bag containing the complementary proportions: 30% red chips and 70% black (this is the “black bag”). Participants are told the distribution of chips in each bag. They are then shown a sequence of *n* chips and asked, after seeing each chip, to estimate the relative probability that the sample came from the red rather than the black bag (see Peterson & Beach, 1967, for examples).

Having seen a sample of *n* events containing *x* red chips, the normatively correct relative probability that the sample came from the red bag rather than the black bag is given by

$$R(x,n,p) = \frac{P(x,n|p)}{P(x,n|p) + P(x,n|1-p)} = \frac{1}{1 + \left[\frac{1-p}{p}\right]^{x}\left[\frac{p}{1-p}\right]^{n-x}}$$

(since the proportion of red chips is *p* in the red bag, and 1−*p* the black bag). As participants proceed through these tasks, they give relative probability estimates that follow the direction required by normative probability theory but are conservative: less extreme than the normatively correct values. This means that if participants see $x > n/2$ red chips in their sample, they give estimates for the probability that the sample came from the red bag that are greater than .5 but less than the normatively correct value, while if participants see $x > n/2$ black chips in their sample, they give estimates for the probability that the sample came from the black bag that are greater than .5 but less than the normatively correct value. In applying our model to this task we assume, without loss of generality, that $x > n/2$ is the number of red chips in the sample of *n* events that have been seen, and we assume $p > .5$ to be the proportion of red chips in the red bag.

The estimated relative probability, in our model, of seeing a sample of size *n* with *x* red chips coming from the red bag rather than the black bag is given by

$$R_*(x,n,p) = \frac{p_*(x,n|p)}{p_*(x,n|p) + p_*(x,n|1-p)} \tag{5}$$

Note that, since by assumption $p > .5$ and $x > n/2$, from Eq. 4 we see that $p_*(x,n|p) > p_*(x,n|1-p)$ will tend to hold (subject, of course, to random error: more specifically, the higher the values of *x* and *p* the more likely it is that this inequality will hold). This means that $R_*(x,n,p)$ will be >.5, and these noisy relative probability estimates will follow the direction required by normative probability theory, just as seen in experiments.

The value of this noisy estimate $R_*(x,n,p)$ varies randomly. By a sequence of rearrangements, we get

$$\langle R_*(x,n,p)\rangle \leq \frac{1}{1 + \left[\frac{(1-2d)(1-p)+d}{(1-2d)p+d}\right]^{x}\left[\frac{(1-2d)p+d}{(1-2d)(1-p)+d}\right]^{n-x}}$$

as an expression for the expected value of this estimate (see Appendix B for a detailed derivation of this result). Comparing our expressions for $\langle R_*(x,n)\rangle$ and *R(x,n,p),* we see that $\langle R_*(x,n)\rangle < R(x,n,p)$ when

$$\left[\frac{1+d\left(\frac{1}{p}-2\right)}{1+d\left(\frac{1}{1-p}-2\right)}\right]^{x} < \left[\frac{1+d\left(\frac{1}{p}-2\right)}{1+d\left(\frac{1}{1-p}-2\right)}\right]^{n-x} \tag{6}$$

Since by assumption we have $p > .5$ and $x > n/2$, we see that the inequality in Eq. 6 always holds, and so $.5 < \langle R_*(x,n,p)\rangle < R(x,n,p)$: Estimated relative probability follows the direction required by probability theory, but it is conservative, just as observed in people's relative probability judgments. In other words, even though the expected values for the individual inferential probability judgements $\langle p_*(x,n|p)\rangle$ and $\langle p_*(x,n|1-p)\rangle$ are each anti-regressive relative to their corresponding normative values in this model, when combined to produce an overall estimate of relative probability, this estimate is regressive and so reproduces the pattern of conservatism seen in inferential judgement.

2.5. Combined estimation and judgment tasks

We finally describe how this model applies to tasks that involve both descriptive and inferential probability estimation. We consider an iterative task that involves the repeated updating of an estimate for a hidden population probability parameter (which may itself randomly change), given a sample of events presented outcome by outcome. Such tasks were investigated in an experiment by Gallistel et al. (2014), where participants gave repeated estimates of the hidden population probability, *p*, of a stepwise nonstationary Bernoulli process that controlled the color of a circle being drawn from a concealed box. On each trial participants clicked a button to draw a new circle from the box. After being drawn, the circle evaporated, and participants could update their estimate for the hidden probability *p*. Participants were told that the box would sometimes be silently replaced by another box with a different value of *p*. Participants could update their estimates by either clicking a "The box has changed!" button (and then picking a new probability estimate), by adjusting their current probability estimate, or by making no change.

There were two main results from this experiment. First, people's probability estimates were characterised by rapid changes in the estimated value in response to changes in the underlying hidden probability, separated by periods of small adjustments in the estimate (see Fig. 1, left side). The speed of detection of a change in the underlying probability *p* depended on the degree of change: Large changes in the underlying probability were detected more rapidly than smaller changes. The median latency for detection of a change in probability estimate in response to a change in the underlying probability was around

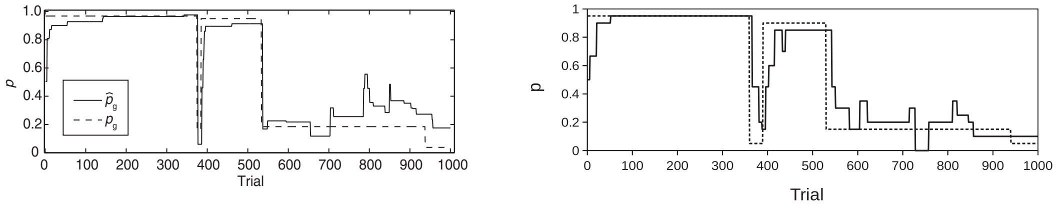

Fig. 1. (Left) Trial-by-trial true probability (dashed line) and trial-by-trial probability estimate (solid line) for Subject 4, Session 8 in Gallistel et al.'s task (from fig. 5 in Gallistel et al., 2014, page 102; p_g and $\hat{p}_g$ represent true and estimated probabilities, respectively). (Right) Trial-by-trial probability estimates produced by our model for the same set of true probabilities. These graphs illustrate the step-hold pattern seen in Gallistel et al.'s task and show that the model reproduces this pattern.

12 events. The second main result was that the relationship between the true probability p and participants' estimated probability was essentially that of identity: The median trial-by-trial probability estimates closely tracked the true hidden probability with no systematic bias.

This pattern of agreement with the true probability arises, in our model, due to the cancellation of regression in probability estimation against anti-regression in inferential judgment. Suppose we see a series of random samples drawn from a population with a parameter $p = P(A)$, and we take p_e to represent our estimate of p (which we repeatedly update as outcomes are presented in the task). This estimate p_e will be subject to random noise, and so it will have a regressive average value as in Eq. 1. Individual estimates p_e will be adjusted (in a quasi-random walk) in response to inferential probability judgment of the chance of obtaining the currently seen sequence of outcomes, given our current estimate. This inferential probability judgment will also be subject to random noise and so will be anti-regressive. This estimate p_e is least likely to be adjusted when it reaches a value maximally consistent with the average number of counted occurrences of A in the presented sample, and so it will tend to fix at that value. Due to random noise, the average number of counted occurrences of A in a sample is equal to $[(1-2d)p+d]n$, and so p_e will fix at the value for which the inferential probability $\langle p_*([(1-2d)p+d]n, n|p_e)\rangle$ is maximized. Since from Eq. 4 this inferential probability has a binomial distribution with probability $(1-2d)p_e + d$, it has its maximum value when

$$(1-2d)p_e + d = (1-2d)p + d$$

or equivalently, when $p_e = p$; when our estimate p_e for the underlying population probability equals the true value. In other words, when descriptive probability estimation and inferential probability judgment are combined, the regressive and anti-regressive effects in each should cancel out, leaving estimates that on average agree with the hidden population probability p; just as seen in mixed estimation and judgment tasks such as Gallistel et al.'s.

3. Computational simulation

We apply the model to Gallistel et al.'s task by assuming that a continuous probability estimate p_e is produced by counting the frequency of A in n just-observed events (subject to random noise). The parameter n here represents the size of short-term memory available to store just-seen events: We assume n is small, but beyond that make no assumptions about the value n (in our simulations, below, we chose n randomly for each simulated participant, uniformly in the range 5... 20).

We take x to represent the number of occurrences of A in the n most recently observed events and take x_e to represent the noisy count of that number (the count of occurrences obtained with a chance d of randomly miscounting). The expected value of x_e equals $(1-2d)x+nd$, and so the immediately observed probability of A in that sample has the expected value

$$q = (1 - 2d)\frac{x}{n} + d \tag{7}$$

On each event occurrence the model makes one of three choices, corresponding to the three choices available to participants in Gallistel et al.'s experiment. First, the model may reject the current value of p_e as inconsistent with the number of A's just observed, and update to a new estimate by setting $p_e = q$ (this choice corresponds to clicking "The box has changed!" in Gallistel et al.'s experiment). Second, the model may decide that the underlying distribution has *not* changed but that q is more consistent with the observed number of A's than p_e. In this case the model again updates to a new estimate by setting $p_e = q$: This choice corresponds to a small adjustment of the current probability estimate. Third, the model may decide not to modify p_e.

To decide whether the current estimate p_e needs to be rejected, the model considers the chance of seeing x_e occurrences of A in n samples where the probability of seeing A in those samples is actually p_e. If this chance is too low, p_e is rejected. The model assesses this chance in a simple way: by generating 100 random samples (each of size n, with A occurring randomly with probability p_e) and counting the number of A's in each sample. This counting process is subject to random error, with some probability $d < 0.5$ that A will be counted as $\neg A$, or $\neg A$ counted as A. The proportion of these samples that contain exactly x_e occurrences of A represents an estimate of the inferential probability $P_E(x_e, n|p_e)$. If this inferential probability is less than some decision criterion T_1, the model concludes that p_e should be rejected because the underlying distribution has changed. The model then changes the new estimate to q.[2]

If the current estimate is not rejected, the model next considers making an estimate adjustment. To decide whether the current estimate p_e needs to be adjusted, the model considers the inferential probability $P_E(x_e, n|q)$: the chance of seeing x_e occurrences of A in n samples drawn from a population where $P(A) = q$. As above, the model assesses this chance by generating 100 random samples (each of size n, with A occurring randomly with probability q) and counting the number of A's in each sample (subject to a rate d of random error in counting). The proportion of these samples that contain exactly x_e occurrences of A represents an estimate of the inferential probability $P_E(x_e, n|q)$. If the difference between this inferential probability and the previous inferential probability is greater than some decision criterion T_2 (that is, if $P_E(x_e, n|q) - P_E(x_e, n|p_e) > T_2$), the model decides that q is a better estimate and changes to a new estimate by setting $p_e = q$. Otherwise the current estimate p_e is left unchanged.

3.1. Results

We implemented this model and tested it by simulating Gallistel et al.'s experiment. On each run of this simulation, the model was shown a consecutive sequence of 1,000 randomly generated A or $\neg A$ events. After seeing each event, the model either rejected its current probability estimate and changed to the new estimate q; adjusted its estimate to the new estimate q; or else left its estimate unchanged. Events were generated randomly, with a hidden probability p. The value of p itself changed randomly over the sequence of 1,000 events, with the probability that p would change after a given event being set at a constant value of .005 (just as in Gallistel et al.'s experiment). The size and direction of a change in the hidden probability were determined by a random choice of the next value from a uniform distribution between 0 and 1, subject to the restriction that $p/(1-p)$, the resulting change in the odds, was no less than fourfold, just as in Gallistel et al. (2014).

To investigate the role of error in descriptive probability estimation and in inferential judgment, we designed the program so that we could set one error rate d for descriptive estimation and another rate d_s for inferential judgement. We simulated Gallistel et al.'s experiment for four different pairs of values for these parameters: Sim A ($d = 0.0, d_s = 0.0$), Sim B ($d = 0.1, d_s = 0.0$), Sim C ($d = 0.0, d_s = 0.1$), and Sim D ($d = 0.1, d_s = 0.1$). We set the criterion parameters T_1 and T_2 to 0.01 and 0.1, respectively, in all simulations, since initial tests suggested that these values produced a reasonable rate of adjustment in the model's probability estimates. Each simulation involved 500 "participants" (runs of the model), all with the same values for parameters d and d_s, and each with a value of n (the size of short-term memory) selected randomly from the range 5 ... 20. Each "participant" saw a different randomly generated sequence of 1,000 events, produced according to a different randomly generated sequence of values of p (as in Gallistel et al., 2014).

3.1.1. Rapid detection of changes

The median latency between a change in the hidden probability p and the recognition of that change by the model (via rejection of the current probability estimate) was 10 in simulations A and B, 13 in simulation C, and 12 in simulation D. These values agree with the median latency of 12 seen in Gallistel et al. (2014).

3.1.2. High hit rates and low false alarm rates

Gallistel et al. (2014) describe a method for computing hit rates and false-alarm rates in participants' responses in their experiment: They found that nine out of ten participants had hit rates in the range 0.77 ... 1 and false-alarm rates in the range 0.004 ... 0.02. We used the same method to compute hit rates and false alarm rates across all "participants" in our simulations. Average hit rates were 0.87, 0.79, 0.81, 0.76, and false-alarm rates were 0.006, 0.005, 0.005, 0.005 in simulations *A*, *B*, *C*, and *D*, respectively. These agree with the rates seen by Gallistel et al. (2014).

3.1.3. Precision

We assess the precision of the model's probability estimates by computing the root mean squared deviation (RMSD) between the model's estimate at a given event against the true probability p at that event. These RMSDs between estimated and true probabilities were 0.15, 0.17, 0.17, 0.17 for simulations A, B, C, and D, respectively. These were consistent with the corresponding RMSD's for participants in Gallistel et al.'s experiment, which ranged between 0.15 and 0.21.

These results show that, if we assume a constant rate of error $d = 0.1$ in both descriptive probability estimation and inferential probability judgment, the probability theory plus noise model produces results that agree closely with those seen in Gallistel et al. (2014). Similar agreement holds for a range of other values of d. These same values of d, however, also produce regressive effects; in our model these regressive effects produce patterns of bias such as conservatism, subadditivity, and the conjunction fallacy. In other words, this model may provide a single unified account for systematic bias away from the true probabilities (in some tasks) and for agreement with the true probabilities (in other tasks): an account that depends on a single factor—noise in reasoning.

These three aspects of the model are illustrated in the right of Fig. 1. This figure shows trial-by-trial probability estimates produced by the model for one run, with parameter values $d = 0.1, d_s = 0.1, n = 20$. Values of the true probability p were controlled to match those in Gallistel et al.'s example. Individual event occurrences in this run, however, were random and did not follow the precise sequence of event occurrences in Gallistel et al. (2014). This figure shows that the model produces the step-hold pattern seen in Gallistel et al.'s task, with large changes in the estimate when the hidden probability changes, and small adjustments, or no changes, otherwise.

3.1.4. Identity between true probability and median estimates

Recall that the noisy frequentist model predicts that noise will have different effects in different probability judgment tasks: When estimating a probability from a sample (descriptive probability estimation), noise will produce regressive effects; when estimating the likelihood of a sample given a probability (inferential probability judgment), noise will produce anti-regressive effects; and in tasks that involve both forms of estimation, these contrasting effects of noise cancel out, producing agreement with the true probability. To test these predictions, for each simulation, we calculated the median estimate produced by the model for a given hidden probability value p. The results are shown in the four graphs in Fig. 2. Graph *A* gives the results obtained when there is no noise in either descriptive estimation or inferential judgment ($d = 0.0, d_s = 0.0$); the relationship between median estimates and the true probability is one of identity here. Graph *B* gives the results with noise in descriptive estimation but not inferential judgement ($d = 0.1, d_s = 0.0$), and it

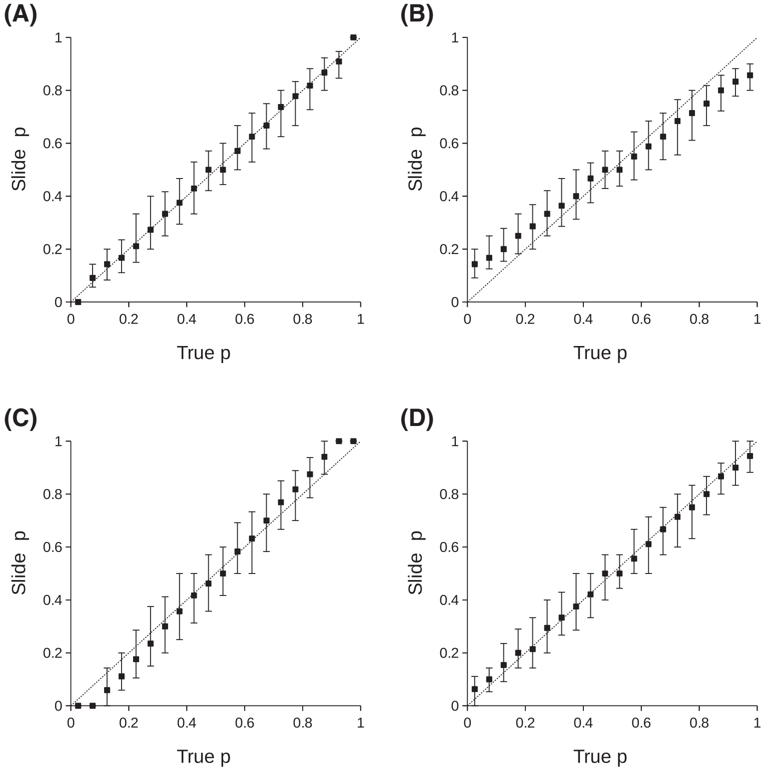

Fig. 2. Median (squares) and interquartile intervals (vertical lines) of model's probability estimates plotted against corresponding true probabilities, for different values of the noise parameters: $d = 0.0, d_s = 0.0$ (graph A), $d = 0.1, d_s = 0.0$ (graph B), $d = 0.0, d_s = 0.1$ (graph C), and $d = 0.1, d_s = 0.1$ (graph D). The dashed line represents identity. While graphs A and D (no noise, and equal noise in both descriptive and inferential probability judgment) show no significant bias away from the line of identity, graph B (noise in descriptive estimation but not inferential judgment) shows clear regression (estimated Slider p being closer to .5 than true p), while graph V (noise in inferential judgment but not descriptive estimation) shows clear anti-regression (estimated Slider p being further from .5 than true p).

shows a clear pattern of regression. Graph C gives the results with no noise in descriptive estimation but noise in inferential judgment ($d = 0.0, d_s = 0.1$), and it shows a clear anti-regressive pattern. Finally, graph D shows the results obtained when there is the same rate of noise in both components ($d = 0.1, d_s = 0.1$). The relationship between median estimates and the true probability in graph D is one of identity: The effects of noise in the two components have cancelled each other out.

4. Conclusions

Our aim in this paper is to present a general model of descriptive probability estimation, of inferential probability judgment, and of the interaction between these two processes. This model assumes that people estimate descriptive and inferential probabilities using a mechanism that follows standard frequentist probability theory, but it is subject to the biasing effects of random noise in the reasoning process. In other work, we have shown that this model accounts for patterns of bias and agreement with probability theory for various probabilistic expressions. Here we show that this model can simultaneously explain the patterns of bias seen in people's probability estimation and inferential judgment (which arise in the model due to the regressive effects of random noise) and the observed agreement with the underlying true probability in tasks such as that of Gallistel et al.'s (where the regressive effect of noise in descriptive probability estimation is counteracted by the anti-regressive effect of noise in inferential probability judgment).

This model predicts a form of "cancellation of noise" in tasks such as Gallistel et al.'s, which involve both descriptive estimation and inferential judgement. This cancellation doesn't mean that estimates in such tasks will always exactly equal the correct normative value, just that estimates will tend to settle at that value, and may move around that value (in a quasi-random walk). In other words, this model predicts that people's estimates in these tasks will tend to agree with the normatively correct value, but with random variation around that value. A similar point applies to expressions such as that in Eq. 2, which combine descriptive probability estimates in a way that cancels out the biasing effects of noise: For such expressions the model also predicts that people's estimates will on average match normatively correct value, a pattern that is strongly confirmed in experimental results (see Costello & Watts, 2014, 2016). More broadly, while our model assumes that people reason in a way that follows probability theory, it does not in general predict agreement with probability theory in people's probability estimates. Instead, it predicts that, in certain specific probability estimation tasks where the effect of noise is cancelled (Gallistel et al.'s task, and the noise-cancelling expressions mentioned above) there will be average agreement with probability theory, while in most other aspects of probability estimation there will be systematic biases away from probability theory.

Our proposal has implications for research on bias in people's decision-making in general. A common pattern in such research is to identify a systematic bias (a systematic violation of normative requirements) in people's decision-making and take that bias as evidence that people do not reason using the normatively correct procedure, but instead use some normatively incorrect heuristic (e.g., Kahneman & Tversky, 1979). Our results, however, suggest that this leap from an observed bias to an inferred heuristic (motivated by, and intended to explain, that bias) is premature. This is because random noise in reasoning can cause systematic biases in people's responses even when people are using normatively correct reasoning processes (see Budescu, Erev, & Wallsten, 1997; Erev et al., 1994, for similar arguments). To demonstrate conclusively that people are using heuristics (are using any nonnormative reasoning process), researchers must show that observed biases cannot be explained as the result of systematic effects caused by random noise.

Notes

1. For this model's account of other biases, such as subadditivity and the conjunction fallacy, see Costello and Watts (2014, 2017).
2. Note that our decision to use 100 random samples when estimating inferential probabilities here is essentially arbitrary: This number was chosen to allow us to use values for the decision criteria T_1 and T_2 that correspond to standard significance level values such as .01 and .05. Versions of the simulation that make use of much smaller numbers of samples give essentially the same results as seen here.

References

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making*, *10* (3), 157–171.

Costello, F. J., & Mathison, T. (2014). On fallacies and normative reasoning: When people's judgements follow probability theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 361–366). Quebec City, Canada: Cognitive Science Society.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480.

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133.

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304–321.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

Erlick, D. E. (1964). Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology*, *67*(5), 475–482.

Fisher, C. R., & Wolfe, C. R. (2014). Are people naïve probability theorists? A further examination of the probability theory + variation model. *Journal of Behavioral Decision Making*, *27*(5), 433–443.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, *121*(1), 96–123.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, *47*, 263–291.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578.

Peterson, C., & Beach, L. (1967). Man as an intuitive statistician. *Psychonomic Bulletin*, *68*(1), 29–46.

Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, *69*(5), 522–527.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354.

Poulton, E. (1973). Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin*, *80*(2), 113–121.

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist*, *67*(8), 603–613.

Teigen, K. H. (1973). Number and percentage estimates in sequential tasks. *Perceptual and Motor Skills*, *36* (3 suppl), 1035–1038.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

Appendix A

We take $p_*(A)$ to represent an individual estimate of the probability of A, produced by randomly sampling some set of events from memory and counting the proportion that are A (subject to random error in reading an item as A). Since $P_*(A)$ is the probability of an item being counted as A, and since these samples are drawn randomly, these estimates $p_*(A)$ will vary randomly following the binomial proportion distribution

$$\frac{Bin(N, P_*(A))}{N}$$

where N is the size of the sample drawn. A property of the binomial proportion distribution is that

$$\left\langle \frac{Bin(N, P_*(A))}{N} \right\rangle = P_*(A)$$

for any sample size N. Given this, we take $\langle p_*(A) \rangle$ to represent the expected value of estimates $p_*(A)$ independent of sample size: the value we would get if we averaged an infinite number of individual estimates $p_*(A)$, each based on a sample drawn randomly from a population with probability $P(A)$ noise rate d, and with sample size varying across samples. Let p_i represent the probability of a sample being drawn with a particular size $N = i$, and we have

$$\langle p_*(A) \rangle = \sum_{i=1}^{\infty} p_i \left\langle \frac{Bin(i, P_*(A))}{i} \right\rangle = \sum_{i=1}^{\infty} p_i P_*(A) = P_*(A) \sum_{i=1}^{\infty} p_i$$

Since the sum of probabilities p_i across all sample sizes necessarily equals 1, we thus have

$$\langle p_*(A) \rangle = P_*(A) = (1 - 2d)P(A) + d \tag{8}$$

as required.

Appendix B

We take

$$R_*(x,n,p) = \frac{p_*(x,n|p)}{p_*(x,n|p) + p_*(x,n|1-p)}$$

to represent an individual noisy estimate for the relative probability that a sample *x,n* came from a population with $P(A) = p$ rather than one with $P(A) = 1{-}p$, as in Eq. 5. The expected value of this expression is

$$\langle R_*(x,n,p)\rangle = \left\langle \frac{p_*(x,n|p)}{p_*(x,n|p) + p_*(x,n|1-p)} \right\rangle$$

For $p > .5$ this function $R_E(x,n,p)$ will be concave for all $x > n/2$ (since as x increases from $n/2$, the probability $p_*(x,n|p)$ increases while the probability $p_*(x,n|1-p)$ simultaneously falls). Since from Jensen's Inequality we have $\langle f(x)\rangle < f(\langle x\rangle)$ for concave functions (the expected value of a concave function is less than that function of the expected value of its argument), we get

$$\left\langle \frac{p_*(x,n|p)}{p_*(x,n|p)+p_*(x,n|1-p)} \right\rangle \le \frac{\langle p_*(x,n|p)\rangle}{\langle p_*(x,n|p)\rangle + \langle p_*(x,n|1-p)\rangle}$$

and so

$$\langle R_E(x,n,p)\rangle \le \frac{\langle p_*(x,n|p)\rangle}{\langle p_*(x,n|p)\rangle + \langle p_*(x,n|1-p)\rangle} = \frac{1}{1+\left[\frac{(1-2d)(1-p)+d}{(1-2d)p+d}\right]^x \left[\frac{(1-2d)p+d}{(1-2d)(1-p)+d}\right]^{n-x}}$$

as required.