Bulletin of the Psychonomic Society
1991, 29 (5), 411-414

Intuitive statistical inference: An "irrational" context effect in college students' categorization of binomial samples

B. KENT PARKER
University of West Virginia, Morgantown, West Virginia

and

CHARLES P. SHIMP
University of Utah, Salt Lake City, Utah

College students learned to earn points by categorizing binomial samples produced by two equally likely and complementary random processes. To facilitate cross-species comparisons, the procedure was made to resemble an experiment with pigeon subjects (Shimp & Hightower, 1990). Samples were 1, 2, 4, or 8 successively presented outcomes (vertical or horizontal lines) of, in effect, tossing one of two equally likely coins, one coin (A) biased in favor of vertical and the other coin (B) similarly biased in favor of horizontal. Choosing red or green stimuli presented after each sample was reinforced if coin A or B had been tossed, respectively. The statistical diagnosticity of a sample, the relative likelihood of its having been produced by a particular coin, was varied by the bias on the coins and by sample size. Subjects became more conservative in the use of statistical information in a sample as the average diagnosticity of other samples (due to the bias on the coins) decreased. Except for this "irrational" context effect, the results qualitatively resembled those in the corresponding experiment with pigeons.

The present experiment was designed to facilitate a cross-species comparison involving a simple type of fuzzy concept, a probabilistic discrimination (Estes, Burke, Atkinson, & Frankmann, 1957). We addressed the question of how subjects discriminate between two random processes after they have been given repeated experience in categorizing samples produced by those processes. Both infrahumans and humans may be assumed to confront many problems of this general type in naturalistic settings; we wished to explore whether infrahuman and human solutions to these problems are qualitatively similar. We used a laboratory setting to facilitate arranging tasks with similar sampling and stimulus characteristics.

Shimp and Hightower (1990) required pigeons to discriminate between binomial samples drawn from two complementary and equally likely Bernoulli-trial processes. On each trial, pigeons saw a sample of 1, 2, 4, or 8 observations and then made a binary categorization response that was rewarded if it corresponded to the random process that had produced the sample. The likelihood that pigeons would correctly categorize samples depended on the relative likelihood that a sample was produced by a particular process (i.e., the sample's "diagnosticity"). In particular, both the magnitude of the complementary biases in the Bernoulli-trial processes and the sample size affected performance in qualitatively appropriate ways. Although performance was, in this sense, adaptive, it was very suboptimal.

Here, we asked if college students would produce similar performances when they were given sampling problems virtually identical to those previously given pigeons. Strikingly inappropriate use of statistical evidence in humans has sometimes been attributed to presumably uniquely human deterministic mental sets (Klayman, 1984) and semantic categories (Kahneman & Tversky, 1973). Therefore, the instructions we gave the college students were designed to reduce their reliance on their verbal knowledge, in the sense that no explicit natural language interpretation, such as sampling tokens from book bags (Edwards, 1968) or diagnosing medical illnesses (Gluck & Bower, 1990), were provided. With this explicit role of language removed, we hypothesized that human performance might not be so strikingly irrational and might qualitatively resemble that of infrahumans. Our goal of a qualitative comparison did not require us to equate amount of practice: The pigeons in Shimp and Hightower's study were given thousands of trials of experience in a within-subject design, while our subjects were given several hundred in a between-subject design. These amounts of training are sufficient to achieve, if not asymptotic performance, at least a rough approximation to it in both humans and pigeons (Friedman et al., 1964; Shimp, 1973; Shimp & Hightower, 1990).

This research was supported in part by the Biomedical Sciences Support Grant Committee of the University of Utah and by NIMH Grant R01 MH 42770 to the second author. Correspondence should be addressed to Charles P. Shimp, Department of Psychology, University of Utah, Salt Lake City, UT 84112.

METHOD

Subjects

The subjects were students enrolled in an introductory psychology course at the University of Utah. They participated in the study for a small increment in credit toward their course grades. This credit was not contingent upon their performance. The subjects were not known to the experimenters and were arbitrarily assigned to experimental groups before their arrival. The data from 40 subjects are presented here. Data from another subject are not presented, because the subject left the experiment due to a scheduling conflict before the 1-h session was over.

Apparatus

The subjects sat facing three Industrial Electronics Engineers inline readouts identical to those used in standard Lehigh Valley Electronics pigeon chambers. Each subject sat about 4 ft away from the display, which was about 1½ ft above eye level for most subjects. The subject pressed three buttons on a small panel placed on the arm of the chair. Each button was 28 × 20 mm and required a minimum force of about .83 N for operation. The three buttons were mounted in a horizontal row across the bottom of the panel. Above the center button was a standard electromechanical cumulative counter. This incremented by one every time a subject made a correct response, as defined below. Environmental events were programmed and responses were recorded by a DEC PDP-8/E computer.

Procedure

The subjects were seated in front of the stimulus display panel, and instructions (see below) were read to them. Each subject was given a 3-min practice session and served in a single 60-min experimental session, after which the subject was debriefed.

Instructions. The instructions read to the subjects were as follows:

> Your goal in this experiment is to obtain as many points as you can. The points you earn are shown on this counter (experimenter indicates the counter). To get points, you have to correctly decide which buttons to press on this panel (experimenter indicates left and right buttons) after you watch a series of lights here (experimenter indicates). On each trial, you will see a series of stimuli appearing in this center position. Each stimulus will be either blue with a vertical line or dark with a horizontal line. Each time a line appears, you have to press the middle button until the line goes out. After some number of lines, there will be no more and the side lights will come on (experimenter indicates). Then you have to press either the left or the right button.
>
> If you are correct, you will immediately earn a point on the counter; if you are incorrect, you must go through the same series of lines again and then push the other button to collect the point for that trial. It is impossible for you to be correct 100% of the time.
>
> When you press a button, just tap it once and release it, like this (experimenter indicates). Do not hold it down. Just to learn what it feels like to press the button, please press it once now.
>
> Are there any questions? (If so, experimenter reads relevant part of the instructions again.)
>
> We will now go through a 3-minute practice period. At the end of 3 minutes, you will have another opportunity to ask questions about the procedure. (Experimenter starts program and halts it after 3 min.)
>
> Do you have any questions? (If so, experimenter reads the relevant part of the instructions.) In 30 minutes, the printer will print for a minute or so and I will return to remove the printout. While it is printing, you will have a rest period. Then the experiment will run for another 30 minutes, after which you will be done, except for a short questionnaire. After you fill out the questionnaire, I'll briefly review for you what the experiment is about. Are you ready? Alright, now I'll start the experiment.

Coin-tossing interpretation. The experimental procedure is conveniently viewed in terms of a coin-tossing interpretation of Bernoulli-trial processes. On each of a series of trials, a subject in effect saw a sample of heads and tails obtained by 1, 2, 4, or 8 tosses of one of two coins. The coin tossed for any given sample was selected with a probability of 0.5. Coin A had a probability of tails P_A; coin B had the complementary probability of tails $P_B = 1 - P_A$. The value of P_A was always greater than 0.5, so that coin A was always biased in favor of tails and coin B was always biased in favor of heads. Coins A and B were equally likely to be tossed 1, 2, 4, or 8 times. After observing the outcomes of the 1, 2, 4, or 8 tosses on a trial, a subject was in effect asked which coin was tossed, in the sense that if coin A were tossed, one response was correct—a response to whichever side button on the response panel corresponded to the location of the red stimulus on the left and right inline readouts. If coin B were tossed, another response was correct—a response to whichever side button on the response panel corresponded to the location of the green stimulus on the left and right readouts. It will be convenient to speak of vertical-biased coin A, where a choice of red was correct, or horizontal-biased coin B, where a choice of green was correct. Thus, it will also be convenient to speak of the *vertical-biased red choice* and the *horizontal-biased green choice*.

A session, which was a series of discrete trials, lasted 1 h. Each trial consisted of a study component, a test component, and, if necessary, a correction component. Each of these components is described below in detail.

Study component. The study component involved successive presentations of the stimuli (heads or tails) on a center location. The two stimuli representing the outcomes of tossing coin A or B were a white vertical line on a blue background (a "tail") and a white horizontal line on a dark background (a "head"), respectively. A trial began with the presentation on the center location of the stimulus representing the outcome of the first toss of coin A or B, whichever had been randomly selected for that trial. The stimulus remained on until the subject responded to the center button after a minimum observation time of 0.5 sec (i.e., there was a fixed-interval 0.5-sec schedule arranged for terminating an observation). A response to the center button after 0.5 sec either terminated the study component, if the sample size for the trial was one, or started an interobservation interval of 0.25 sec, if the sample size was greater than one. During this interval, all lights were off. If the sample size was two or greater, the stimulus representing the outcome of the second toss of the same coin was then presented on the center location. Note that the probability of a head was constant throughout a trial, because the same coin was tossed for every observation in a sample. The number of observations in any sample was 1, 2, 4, or 8; as noted above, all four possible sample sizes were equally likely.

A trial began with the presentation of the first heads or tails on the center stimulus location and ended with the delivery of a point by incrementing the cumulative point count. There was then a 5-sec intertrial interval before the beginning of the next trial.

Test component. The termination of the last observation in a sample initiated a 0.1-sec interval, after which two side lights appeared red and green. The side on which a particular color appeared varied randomly. A subject's task was to respond to the color corresponding to the random process that had produced the preceding sample: the task may be interpreted to have been to respond to the color that represented the coin tossed on that trial, red for tails-biased (vertical-biased) coin A and green for heads-biased (horizontal-biased) coin B. If a subject responded to the color corresponding to the coin tossed, the lights went out and a point was delivered. An intertrial interval then elapsed before the beginning of the next trial.

Correction component. A response to the incorrect side button began a 1-sec correction interval, during which none of the three lights was on and after which the same sequence of observations was repeated again. The same sample was repeated after any subsequent error until the subject made the correct response (the position of which stayed the same) and a point was delivered.

Definition of groups. Four groups of 10 subjects each varied only in terms of P_A. Groups 1, 2, 3, and 4 had P_A values of .90, .60, .95, and .75, respectively, and were run in this sequential order. The subjects were assigned to groups in their order of appearance. The mean number of trials per group was 261, 251, 296, and 259, for Groups 1, 2, 3, and 4, respectively.

RESULTS

The relative likelihood of categorizing a sample as diagnosing the vertical-biased red process A was estimated by the relative frequency of choosing the red button. A relative frequency for each sample type was calculated

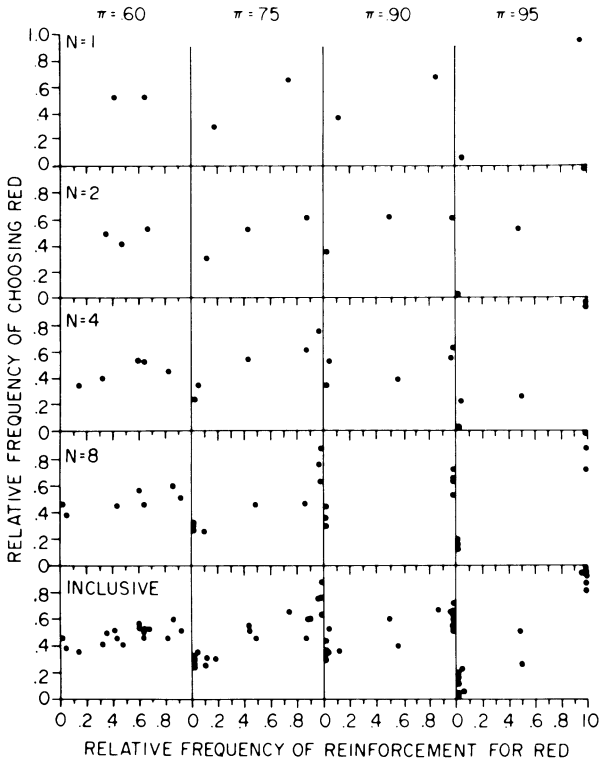

Figure 1. The relative frequency of choosing the button correlated with red as a function of the obtained relative frequency of reinforcement for red, for each experimental condition (columns) and each sample size (rows). A choice of red may be interpreted as categorizing the previous sample of heads and tails as having been produced by tossing the coin biased in favor of tails. Each point represents the average of responses for all 10 students. The bottom row is inclusive; it includes all of the points for each sample size.

for each group by summing the number of red choices across all 10 subjects and then dividing by the total number of trials on which the subjects in that group saw that sample type. Responses counted in the calculations included only the first choice on each trial; they did not include any responses during correction components.

Figure 1 shows the relative frequency of inferring that a sample was produced by vertical-biased red process A as a function of the obtained relative frequency with which that sample actually was produced by process A (the number of reinforced responses to red divided by the total number of reinforced responses to red and green). Each point in Figure 1 represents a different type of sample with a specific number of observations and a specific number of heads. For example, for a sample size of two, there were three possible types of samples: no heads and two tails, one head and one tail, or two heads and no tails. The relative frequencies of choosing red after these three types of samples are represented by the three points in the next-to-the-top row in Figure 1.

To reduce effects due to averaging over changing choice probabilities, Figure 1 (and also Figure 2 below) show performances in the second half of each session. Also, Figures 1 and 2 display results for trial types that occurred at least five times over the second half of a group's performance. (One property of the binomial sampling distribution is that different sample types occur with different frequencies. Thus, with the present arrangement, a few trial types occurred very infrequently.) A total of seven points were deleted due to this criterion.

Figure 1 shows that, on the whole, categorization of a sample as belonging to the vertical-biased red category A generally increased as a function of the degree to which it diagnosed that category. However, while categorization clearly depended on statistical diagnosticity, it did so to a degree that was far below optimal: categorization was in this sense "conservative" (Edwards, 1968) relative to the optimal strategy that was always to categorize a sample in terms of the relatively more likely category; in Figure 1, such a strategy would produce a step function rising from zero to one at a diagnosticity of 0.5

An inspection of the bottom row of Figure 1, each panel of which includes all the points of the four panels above it, reveals that the function relating categorization to diagnosticity became steeper as the bias on the coins became more extreme. This means that the use of a *given* diagnosticity became less efficient (more conservative) as the overall diagnostic context became less clear (as the bias of the random processes approached 0.5). If a sample with a given diagnosticity was surrounded on other trials by samples having relatively less diagnosticity, less use was made of the diagnosticity of that sample. Categori-

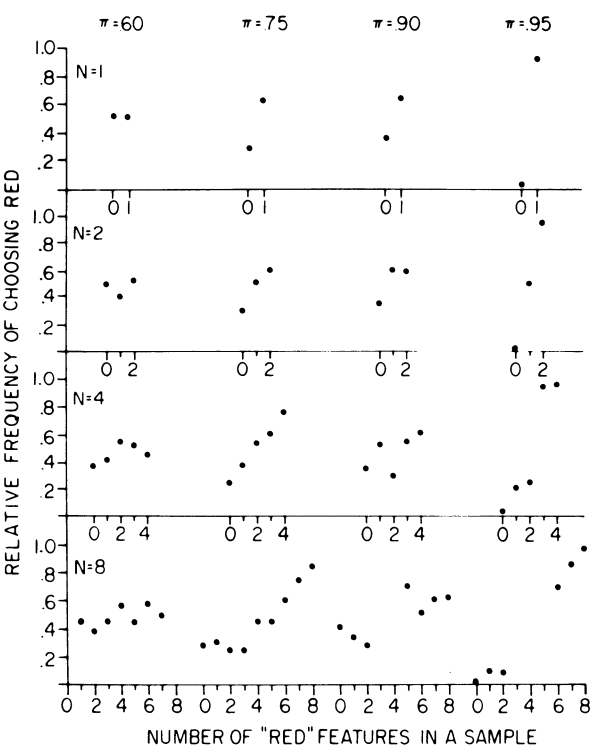

Figure 2. The relative frequency of choosing red as a function of the number of tails in a sample, for sample sizes 1, 2, 4, and 8, in rows 1, 2, 3, or 4, respectively.

zation became more conservative as overall diagnosticity was reduced.

Figure 2 reveals a similar picture. It shows that the relative frequency of the tails-biased red choice generally increased as a function of the number of tails in a sample, but in a distinctly suboptimal manner relative to the step functions that would be produced by optimal inference. Also, inference was most conservative when the bias was .60, intermediate when it was .70 or .90, and least, or most nearly optimal, when it was .95.

DISCUSSION

The task that the subjects faced here was one of "intuitive" statistical inference, since they had to categorize statistical evidence without recourse to explicit computational methods. Intuitive statistical inference in humans often dramatizes how statistical evidence can be misused (Gluck & Bower, 1990; Kahneman & Tversky, 1973). Linguistic knowledge and unequal base rates are often held responsible for this misuse. In the present case, this misuse took the form of conservatism in the use of statistical information and may not have been attributable largely to linguistic knowledge, because the instructions the subjects received were minimal: They involved no verbal interpretations inviting the use of common heuristics or other rules or strategies. Neither was the misuse attributable to base rates, since the two random processes were equiprobable (i.e., the two coins were equally likely to be tossed). Perhaps misuse of intuitive statistical information in humans is due to underlying elementary, nonverbal processes. If this were so, it might help to explain why our subjects' misuse of statistical data resembles that obtained by Shimp and Hightower (1990) with pigeon subjects, in the sense that the human and pigeon subjects both showed conservatism in the use of the same statistical information contained in the same visual displays. Both the animals and humans were sensitive to the variables upon which optimal statistical inference depends, but to a very less-than-optimal degree.

While the pigeon and human performances were alike in the sense that both functions relating categorization to statistical diagnosticity showed conservatism, this conservatism was not entirely alike in each case. In pigeons, conservatism did not appear to depend on the coins' bias, whereas in the present human data, it did. This conservatism is "irrational" in the sense that it reduces the probability of a correct categorization; therefore, it can be said that, with the present binomial sampling arrangement, the humans displayed a form of irrationality that the pigeons did not.

REFERENCES

EDWARDS, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: Wiley.

ESTES, W. K., BURKE, C. J., ATKINSON, R. C., & FRANKMANN, J. P. (1957). Probabilistic discrimination learning. *Journal of Experimental Psychology*, **54**, 233-239.

FRIEDMAN, M. P., BURKE, C. J., COLE, M., KELLER, L., MILLWARD, R. B., & ESTES, W. K. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 250-316). Stanford, CA: Stanford University Press.

GLUCK, M. A., & BOWER, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General*, **119**, 105-109.

KAHNEMAN, D., & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237-251.

KLAYMAN, J. (1984). Learning from feedback in probabilistic environments. In K. Borcherding, B. Brehmer, C. Vlek, & W. A. Wagnenaar (Eds.), *Research perspectives on decision making under uncertainty* (pp. 81-92). Amsterdam: North-Holland.

SHIMP, C. P. (1973). Probabilistic discrimination learning in the pigeon. *Journal of Experimental Psychology*, **87**, 292-304.

SHIMP, C. P., & HIGHTOWER, F. A. (1990). Intuitive statistical inference: How pigeons categorize binomial samples. *Animal Learning & Behavior*, **18**, 401-409.

(Manuscript received March 22, 1991.)