Intuitive statistical inference: Categorization of binomial samples depends on sampling context

CHARLES P. SHIMP, KARREN A. LONG, and THANE FREMOUW
University of Utah, Salt Lake City, Utah

Pigeons categorized binomial samples. One of two "coins" was tossed on each trial, and birds learned to infer from observing the outcomes which of the two equally likely coins had been tossed. Outcomes ("heads" or "tails") appeared as successively presented red or green center keys. Coin R was biased in favor of red, and coin G was similarly biased in favor of green. A categorization consisted of a choice of a left or right side key and was reinforced with food if it was to the key (left for coin R and right for coin G) corresponding to the coin that produced that trial's sample. Coin bias and minimum sample size required for reinforcement were experimentally manipulated. When sample size was greatest ($n = 8$), categorizing a sample as having been produced by coin R tended to undermatch the probability that the sample was produced by coin R. When sample size was smallest ($n = 1$), categorizing a sample overmatched, provided that the context did not include other trials with large samples. This context effect reconciles an otherwise inconsistent literature on intuitive statistical inference in pigeons but suggests a new and difficult goal for research—the general clarification of the effects of sampling context on inference.

Many real-world discriminations may be probabilistic in nature (Brunswick, 1939; Gigerenzer & Murray, 1987; Staddon, 1988; Stephens & Krebs, 1986). For instance, foraging animals in the wild presumably face many ambiguous environments where decisions about whether or not to remain in one patch depend on incomplete evidence. Naturalistic foraging situations, however, may involve formidably complex statistical problems, including patch assessments, changing environments, and distributions of travel times. While it is important to study these complex discriminations in naturalistic settings, we feel that it is also useful to study simplified versions in the laboratory. Here, we specified simple random processes, let these processes generate samples, and then studied how well subjects could learn to discriminate between these processes on the basis of the different likelihoods with which they produced these samples. Reward for a categorization of a sample in such a probabilistic discrimination (Estes, Burke, Atkinson, & Frankmann, 1957) depends on whether the categorization corresponds to the process that produced it.

The probabilistic discriminations we will describe may be interpreted as experiments in intuitive statistical inference because a subject, without benefit of explicit computational methods, had to categorize samples in a

manner logically equivalent to inferring the populations from which they were drawn.

If we are to study simplified probabilistic discriminations, or intuitive inference, what problems have high priority? If real-world discriminations are statistical in nature, then it becomes important to know how discriminations depend on the statistical diagnosticity of samples, or, more broadly, how organisms categorize ambiguous evidence. Thus, perhaps the most basic goal in the study of intuitive statistical inference is to characterize how the accuracy of intuitive statistical inference depends on the statistical diagnosticity of a sample, where diagnosticity is defined as the relative likelihood that the sample is generated by a particular random process. We find it quite surprising that relatively little is known about this function relating correct inference to diagnosticity even when samples are produced by the simplest random processes. The existing literature on probability learning and probabilistic discrimination in nonhuman animals does not give a clear picture of this dependency between discrimination and diagnosticity for even the simplest random processes. Consider this literature as it pertains, for example, to pigeon subjects.

Pigeons can infer rather accurately from which of two populations random samples are drawn both in a simple probability-learning task, if only a single observation is presented in accordance with a binomial distribution (Shimp, 1966), and in a probabilistic-discrimination task, if up to four observations in a sample are presented simultaneously and in accordance with complementary multinomial distributions (Shimp, 1973). (The multinomial is a generalization of the binomial distribution. The binomial involves two possible outcomes, the probabilities of which are complementary and constant over inde-

This research was supported in part by NIMH Grant R01 MH42770. The authors would like to thank Ann Rudge for help in conducting the experiment, Andrew Shimp and Laurie Ingebritsen for help in data analysis, and William Harris for help with computer programming. Correspondence should be addressed to C. P. Shimp, Psychology Department, University of Utah, Salt Lake City, UT 84112.

—Accepted by previous editor, Vincent M. LoLordo

pendent trials. The multinomial is similar except that there are several possible outcomes on each trial.) The relative frequency with which pigeons categorized samples, either binomial with one observation or multinomial with up to four simultaneously presented observations, as belonging to the more likely category exceeded the probability-matching level and approximated maximizing.

However, in a more recent experiment (Shimp & Hightower, 1990) with sequentially presented observations, pigeons categorized binomial samples much less accurately. On different occasions within the same task, pigeons observed and categorized samples with randomly selected numbers of sequentially presented observations (either 1, 2, 4, or 8), in a highly nonoptimal manner, with the likelihoods of appropriate categorizations falling below the probability-matching levels in the direction of chance performance (see Shimp & Hightower's, 1990, Figure 1). Furthermore, the function relating categorization to diagnosticity was independent of sample size. This latter result was quite surprising given the pervasive roles of sample size in mathematical statistics and human inference. One would have expected that the function relating categorization to statistical diagnosticity would have depended on sample size. One might also have expected categorization accuracy to decline as sample size increased because, in a task with sequentially presented observations, animals would be expected to forget more of larger samples than of smaller samples.

Thus, the available literature defines two chief questions. First, why was categorization of ambiguous evidence so much more consistent with optimality theory in the experiments by Shimp (1966, 1973) than in the experiment by Shimp and Hightower (1990)? Second, why was the function relating categorization to diagnosticity independent of sample size in the experiment by Shimp and Hightower (1990)? Our tactic was to try to answer these two questions about how nonhuman animals make inferences in the simplest of sampling tasks as a stepping stone to understanding how they deal with the more complex sampling environments of the natural world.

GENERAL METHOD

Apparatus

Five standard three-key Lehigh Valley Electronics pigeon chambers were interfaced with a Digital Equipment Corporation PDP-11/73 computer, which arranged all experimental events and recorded the data. The pigeon chambers had ventilator fans, and white noise helped to mask extraneous sounds.

Procedure

The experimental procedure is conveniently described in terms of a coin-tossing interpretation of a Bernoulli-trials process (Feller, 1950). On each of a series of trials, a subject in effect saw a sample of "heads" (a red center key) and "tails" (a green center key) obtained by n tosses of one of two coins. Coin R had probability of red, $P_R > .50$, so that it was biased in favor of red, and coin G had the complementary probability of red, $P_G < .50$, so that it was biased in favor of green. The probability of red on a trial, either P_R

for coin R or P_G for coin G, was constant over the several observations in a sample for a given trial since the same coin was tossed for every observation in a given sample. The number, n , of observations within any given trial varied over trials in the way described below.

A subject sequentially observed the outcomes of these coin tosses and, in effect, was asked to report on each trial which coin, R or G, had been tossed on that trial. That is, a subject's task was to respond to the side key corresponding to the random process that had produced the observed sample—to respond to the key that represented the coin tossed on that trial. Pecks on side keys in this sense categorized the sample in terms of which coin was tossed: a response to the left key was correct if red-biased coin R were tossed, and a response to the right key was correct if green-biased coin G were tossed.

Sessions lasted 45 min and consisted of a series of discrete trials. The sequence of events within trials is described below.

Sequential organization of a trial. Each trial began with all three keys lit white. A peck to the center key meant, in plain English, "I want to see more data." A peck to either side key meant, "I want to categorize the sample I have just observed." The consequences of each type of response are described below.

Consequences of a center-key peck. A peck to the center white key, which meant "I want to see more data," turned off all three white lights and presented the next outcome of tossing whichever coin was being tossed on that trial. This observation, either a red or a green light, appeared on the center key. After a minimum observation duration elapsed, the first peck to the center key initiated an interobservation interval during which all lights were off and after which all three lights again appeared white.

Consequences of a side-key peck. A peck to a side key had different consequences depending on whether a minimum sample size had been reached. This number was the minimum number of observations that had to be presented on the center key before a side-key response would produce a reinforcer and terminate a trial. A bird was free to categorize a coin before having seen this minimum number of observations; however, if it did so, that categorization response could not be reinforced. We will separately describe the side-key consequences for the two cases corresponding to whether or not this number had been reached.

Sufficient data presented before categorization. If the minimum sample size had been obtained, a side-key peck, which meant "I want to categorize the sample I have just observed," delivered a reinforcer (2.0 sec of access to mixed grain), if the response was to the side key corresponding to the coin tossed on that trial. In this case, one could say that the bird correctly guessed the coin that had been tossed on that trial. Otherwise, if the bird guessed incorrectly and the side-key peck was not to the side key corresponding to the coin tossed on that trial, the peck started a correction interval during which the houselight flashed on and off every 0.1 sec (for details, see "Correction procedure" section below).

Sufficient data not presented before categorization. If the subject pecked a lit side key before having produced the minimum required number of observations, the consequences depended on whether the response was the first side-key response after the last center-key peck. That is, the consequences depended on whether the subject had last requested to see more data or had last tried, without enough data, to categorize the sample. If the last response had been to the center key, then a side-key peck turned off all three keys and the houselight for a time equal to the interstimulus interval, after which the house light came back on and all three keys again appeared white. However, if a response to a lit side key followed a response to a lit side key, whether to the same or different key, it had no programmed consequences. The procedure in this sense demanded another request for more data before it permitted a subject to make another categorization of the sample after already having categorized it. This procedure applied when a subject

had already responded to a side key before having seen the minimum number of observations.

These complex arrangements ensured that no informative feedback about the correct response on a trial was provided before the required minimum sample had been delivered. They also ensured that the same sample could not be categorized more than once. (A reader will note, however, that a subject could make more than one categorization response within the same trial—for example, by categorizing a sample before having produced the minimum sample size required for reinforcement, then asking for more data, and finally categorizing the larger sample. Experiment 3 evaluated the possibility that having made a categorization early in a sample affected a later categorization of a larger sample in the same trial.)

Correction procedure. A correction interval began if the subject had most recently asked to see more data and then guessed incorrectly which coin had been tossed. More precisely, this interval began if the bird had most recently pecked the center key and then responded to the incorrect side key after having seen the minimum number of observations. During the 2.5-sec correction interval, the key lights were off and the houselight blinked on and off every 0.1 sec. The two side keys reappeared white after the interval ended. A response to the correct side key produced reinforcement and ended the trial. A response to the incorrect side key started the correction interval over again. Subsequent errors repeated this cycle until the subject made the correct response (the position of which stayed the same since it corresponded to the coin being tossed on that trial) and a reinforcer was delivered.

Pretraining. Total pretraining lasted approximately 7 months while the experimental procedure was developed and subjects learned to respond differentially to red and green. The subjects were first given color-discrimination training, with a randomly selected red or green color presented according to a fixed-interval 5-sec schedule: The first center-key response after 5 sec elapsed turned off the center key and turned on white side keys. A peck to the left (right) key was reinforced after a red (green) center-key color. The average relative frequency of a correct response over the last 5 days of this pretraining was .98 after red and .96 after green. The subjects subsequently performed for several months on a variety of preliminary versions of the procedure. The main procedural developments during this time were the implementation of various contingencies for side-key responses occurring before the minimum sample size was obtained and for incorrect categorizations.

EXPERIMENT 1

Our first purpose was to see if it was possible to replicate the close approximation to maximizing that was obtained by Shimp (1966, 1973) but with a probabilistic-discrimination task with successive observations. We began with a task involving very small samples. Experiment 1 constrained sample size by placing a lower bound on the number of observations in samples preceding reinforced categorizations. Several different biases on the coins were used to determine the generality of the resulting categorization functions.

Method

Subjects. Five pigeons were maintained at 80% of their free-feeding weights, plus or minus 10 g. Grit and water were always available in the home cages, where there was a 14:10-h light:dark cycle. Three birds were experimentally naive, and 2 had served in previous experiments involving discriminations among alphanumeric characters.

Apparatus. The apparatus was as described above.

Procedure. The procedure was as described above, with the values of P_R equal to .95, .90, .80, .70, and .90 in Conditions 1, 2, 3, 4, and 5, respectively. Conditions lasted 15, 15, 14, 15, and 19 days, in Conditions 1, 2, 3, 4, and 5, respectively. The intertrial interval was 20 sec, the correction interval was 2.5 sec, the interobservation interval was 0.1 sec, and the minimum observation time was 0.75 sec. The minimum required number of observations before a categorization response could be reinforced was fixed at two.

Results

We define any response to the left or right key in a trial as one that categorized the observed sample on that trial, provided that it was the first side-key response after a given number of observations of stimuli on the center key. For instance, if a bird categorized a sample with a single observation, no subsequent categorization response could occur until the bird produced a sample with at least two observations. Defined this way, a response that categorized a sample was one that preceded any informative feedback about the correct response on that trial. A categorization before the required minimum number of observations simply produced three white keys. All such categorization responses were included in the analysis since they all preceded any informative feedback. However, there could be only one categorization response in a trial after the required minimum number of center-key pecks since it produced informative feedback in the form of either a reinforcer or the correction-procedure interval. All responses during the correction procedure were ignored in the data analysis.

The relative frequency of categorizing a particular sample as one produced by coin R (i.e., of categorizing it by a response to the "red-biased" left key) equaled the number of times that subjects chose left after that sample divided by the total number of side-key pecks after that sample. The obtained statistical diagnosticity of a sample, the likelihood the sample was produced by a particular coin, equaled the number of trials on which that sample was produced by that coin divided by the total number of trials on which that sample was categorized.

Figure 1 shows the relative frequency of choosing the left key after a sample, or the relative frequency of categorizing the sample as one produced by coin R, as a function of that sample's statistical diagnosticity. The results are averaged over the 5 birds and, in order to delete performances that might depend heavily on the previous condition, the last 5 days of a condition. Each point corresponds to a different sample type. That is, each point represents a different type of sample with a specific number of observations and a specific number of "heads." For example, for sample size 2, there were three possible types of samples: no red and two greens, one red and one green, or two reds and no greens. The relative frequency of choosing left after these three types of samples is represented by the three points in the third column in Figure 1. Rows correspond to values of P_R , and columns correspond to sample size, or the number of center-key observations

Figure 1. The relative frequency of choosing the left key, interpreted as categorizing a sample as one produced by tossing coin R (a Bernoulli-trials process with probability of "heads" P_R) as a function of the obtained relative frequency with which that sample was actually produced by coin R instead of by complementary coin G. Coins R and G were biased in favor of red and green, respectively. Only categorizations of samples with at least two observations could be reinforced. Different rows (corresponding to different conditions) show the results for different values of the probability of heads, P_R . Different columns show the results for samples with different numbers of observations. A number of observations equal to 0 means that a categorization occurred before any observation was presented. Panels in the bottom row are inclusive and combine results in the corresponding columns. Diagonal lines represent the theoretical matching function.

of red and green before a side-key categorization response. Recall that throughout these conditions, the required minimum number of observations was two and categorization of samples smaller than two were not followed by informative feedback. The bottom row is inclusive: each panel shows all the points from the panels above it. It should be noted that all the results presented below should be interpreted in terms of group averages. No claims are made here about individual performances, not because individual performances do or do not conform to group averages but because complexities of the im-

plications of binomial sampling for the frequencies of occurrence of particular samples would make any such claims in the present case too delicate to be worthwhile.

The bottom, inclusive panels clearly show that the function relating categorization to diagnosticity depended on sample size or amount of data. When a sample was categorized before the subject saw an observation (left-most panel), categorization was, of course, at chance. When categorization was based on one observation, the function shows overmatching; that is, performance deviated from probability matching in the direc-

tion of maximizing—in fact, every point deviated from matching toward maximizing. (Probability matching in discrete-trials, probability-learning, and probabilistic-discrimination experiments refers, of course, to an approximate equality between the relative frequency with which an animal chooses an alternative in the presence of a stimulus and the relative frequency with which that alternative is reinforced in the presence of that same stimulus.) When categorization was based on two observations, the function approximated matching. When categorization was based on three (right-most panel), there is perhaps a suggestion of undermatching, with nearly all the points deviating from matching toward indifference.

It is of some interest to notice how the required sample size controlled average sample size. The average sample sizes preceding categorizations were 2.51, 2.32, 2.21, 2.10, and 2.27 in Conditions 1, 2, 3, 4, and 5, respectively. The mean just slightly exceeded the value of two that was the minimum required for reinforcement and, in this sense, performance was highly adaptive. The subjects were able to determine what sample size was required for reinforcement in the context of producing and categorizing binomial samples.

Discussion

Experiment 1 created a sampling context with only small samples and produced two noteworthy results. First, categorization of the smallest samples (those with only one observation) was very accurate and exceeded the probability-matching level in the direction of maximizing. This result is consistent with the earlier overmatching results of probability-learning (Shimp, 1966) and probabilistic-discrimination (Shimp, 1973) experiments. Thus, Experiment 1 replicated earlier results that approximated maximizing.

Second, even within the very narrow range of small sample sizes produced in Experiment 1, categorization accuracy depended on sample size. Categorization accuracy shifted from overmatching with one observation to a less accurate level with merely two or three observations. In the case of samples with two observations, categorization very closely approximated matching, but if viewed as a whole—that is, if we look at categorization of samples with one, two, or three observations—the function relating categorization to diagnosticity suggests matching may be only an arbitrary point on a function ranging from indifference to maximizing. This dependency of categorization on sample size conflicts with Shimp and Hightower (1990; see their Figure 1), where no such dependency was found.

EXPERIMENT 2

The results of Experiment 1 suggest a straightforward role for sample size in intuitive statistical inference when observations are presented sequentially. If we interpret sequential observations as stimuli that a subject has to remember, then it follows naturally that larger samples will

be more difficult to remember. Such an interpretation would lead us to expect that categorization of larger samples would be poorer than that of smaller ones, given equal diagnosticities, which of course is the result obtained in Experiment 1. One would expect that if sample size were made still larger than those in Experiment 1, categorization would become even poorer. In Experiment 2, we therefore examined how accuracy of inference depends on sample size across a broader range of sample sizes.

In Experiment 2, we also examined the possibility that sample-size effects could explain the difference between the accurate inference based on samples of one observation in Shimp (1966) and in the present Experiment 1, on the one hand, and the much less accurate inference in Shimp and Hightower (1990) even in those cases in which samples were very small, on the other hand. We attempted to reconcile these results by examining the role of sampling context. We evaluated the idea that birds might learn the sizes of samples they repeatedly experience and therefore learn to expect additional observations after the first ones in cases where most samples have several observations. A bird might not attend as carefully as it otherwise would to the earliest observations in a sample if it expected subsequent observational opportunities. Such a process could resolve the problem that the results of the present Experiment 1, of Shimp (1966), and of Shimp (1973) conflict with the data in Shimp and Hightower's (1990) Figure 1, because in the former three experiments, samples either had only one or two observations presented sequentially or had several observations presented simultaneously. Only in Shimp and Hightower (1990) were observations presented sequentially in a manner that would strongly encourage a bird to attend less carefully to early observations in the expectation that there would be several subsequent observational opportunities. In that experiment, samples with 1, 2, 4, and 8 observations were equally likely. Therefore, three fourths of the time, after one observation, a subject could expect at least one more observational opportunity. Thus, in Experiment 2, we attempted to show that inference based on one observation becomes less accurate if the context includes other, larger samples.

Method

Subjects. A different group of 4 subjects served in Experiment 2. Of the 4, 3 were experimentally naive and 1 had previously served in an experiment involving discriminations among alphanumeric characters.

Apparatus. The apparatus was that described in the General Method section.

Procedure. The pretraining and procedure were as described in the General Method section. At the end of pretraining, the value of P_R was .95. Then, throughout Experiment 2, it was fixed at .75, a value halfway between the nonpredictive value of .50 and the perfectly predictive value of 1.0. The minimum number of observations required before a categorization response could be reinforced was varied over conditions. It was 2, 3, 4, 6, and 8, in Conditions 1, 2, 3, 4, and 5, respectively. Conditions lasted 15, 14, 15, 19, and 15 days, respectively. All other procedural probabilities and intervals were as in Experiment 1.

Figure 2. The relative frequency of choosing the left key, interpreted as categorizing a sample as one produced by coin R, as a function of the obtained relative frequency with which the sample was produced by coin R, the probability of which, P_R , was .75. Different rows correspond to different conditions. Only categorizations following a minimum number of observations, varied over conditions, were reinforced. These minimum numbers were 2, 3, 4, 6, and 8 for rows 1, 2, 3, 4, and 5, respectively. Different columns correspond to samples with different numbers of observations.

Results

Figure 2 shows the relative frequency of choosing the left key, or categorizing a sample as one produced by coin R, as a function of the sample's diagnosticity. The results are averaged over the 4 birds and the last 5 days of a condition. Columns correspond to different numbers of observations preceding a categorization response, and rows correspond to different values of the variable experimentally varied over conditions, the minimum number of observations required in a sample before a categorization could be reinforced.

Condition 1, represented in the top row, shows that, as in Experiment 1, maximizing was obtained with samples of one observation when the required sample size was two. Also as in Experiment 1, accuracy was markedly reduced even for samples with as few as two or three observations: the panel for samples of three observations shows functional relations consistent with a matching function. In short, the top row of Figure 2 shows that Experiment 2 replicated the essential results of Experiment 1. Subsequent rows of Figure 2 show effects of varying the minimum required sample. The second through fifth rows in Figure 2 show that when the required sample size was 3, 4, 6, or 8, maximizing was not obtained with even a single observation. In general, as sample size increased beyond one, the categorization function evolved into a very orderly undermatching curve, perhaps best illustrated in

the right-most panel, for sample size 8, of the bottom row where eight observations were required for reinforcement.

The minimum sample size that was required for reinforcement was varied in Experiment 2, so we may compare the minimum required with the average sample sizes that the subjects produced. When the minimum requirement was 2, 3, 4, 6, and 8, the corresponding average sample sizes that the birds produced were 2.18, 2.86, 3.51, 4.70, and 6.08, respectively. The required minimum was slightly exceeded at the smallest values and progressively undershot at the larger ones.

Discussion

Categorization was very accurate on trials when samples were never required to have more than two observations and subjects categorized samples after only a single observation. Categorization on those trials exceeded the probability-matching level in the direction of maximizing and so agreed with results in the present Experiment 1, Shimp (1966), and Shimp (1973), where samples either had a single observation or had all observations presented simultaneously.

Two phenomena appeared when sample size increased. One involved categorization of larger samples themselves, and the other involved categorization of smaller samples in a context including other, larger samples. We will consider each in turn.

First, when sample size increased, accuracy decreased, as in Experiment 1. In Experiment 2, moreover, we can see that the trend that began across samples of one to three observations in Experiment 1 continued with still larger samples. In particular, categorization accuracy rapidly declined toward that reported in Shimp and Hightower's (1990) Figure 1. Categorization accuracy resembling theirs was obtained here when categorization followed samples with six, seven, or eight observations when eight were required for reinforcement. Results similar to theirs also began to emerge even for samples with four observations when four were required. Thus, Experiment 2 replicated the results of Shimp and Hightower (1990) for samples of four and eight observations.

Second, categorization accuracy was worse after even small samples with only one or two observations if larger samples appeared on other trials. This sampling context effect reveals an effect of the presence of larger sequential samples on categorization of even very small ones. This context effect reconciles the nearly optimal performance obtained when all samples are very small, as in the present Experiment 1 and in Shimp (1966), and when all observations are presented at once (Shimp, 1973), with the very nonoptimal performance on small samples shown in the top two rows of Shimp and Hightower's (1990) Figure 1. In short, we can now see that when samples with just one or two sequentially presented observations appear in a context that includes larger samples, as in Shimp and Hightower (1990), accuracy of categorization of even those smaller samples suffers.

This same context effect appears to explain why sample size played no apparent role in Shimp and Hightower (1990, Figure 1). We can now see that this context effect probably operated there to reduce categorization accuracy for small samples down to about that of larger samples. It therefore falsely appeared that sample size does not, in general, affect categorization. We can now see, that is, that the task in Shimp and Hightower (1990) was one in which the context effect revealed here would obscure an effect of sample size.

It will be noted that all of these results can be interpreted in terms of the possibility that a bird attends less closely, or with a lower probability, to an observation when the task teaches the subject that further observational opportunities are likely.

In summary, Experiments 1 and 2 provide a coherent description of available data on birds' accuracy of categorizing binomial and multinomial samples with either successive or simultaneous presentations of observations.

EXPERIMENT 3

Recall that in Experiments 1 and 2 a bird could produce and categorize more than one sample within a single trial. It is possible that this aspect of Experiments 1 and 2 could affect the results. Experiment 3 was therefore conducted to determine if an earlier categorization of a smaller sample within a trial could somehow interfere with or otherwise affect a subsequent categorization

of a larger sample within the same trial. Such interference could produce, for instance, the shift observed in Experiments 1 and 2 from better to poorer inference as sample size increased. Any such possible effects were removed in Experiment 3 by the simple method of including only the first categorization in each trial in the calculation of the relative frequencies of categorizations.

In addition, Experiment 3 roughly doubled the amount of training given subjects in order to evaluate the possibility that performance was relatively poor on samples with several observations in Experiment 2 merely because of inadequate training.

Method

Subjects and Apparatus. The subjects and apparatus were as in Experiment 2.

Procedure. Between Experiments 2 and 3, the effects of several temporal variables, including observation time and delay between sample exposure and categorization, were investigated. These manipulations reduced overall accuracy of categorization. Therefore, before initiating Experiment 3, we conducted a simple red-green color discrimination until correct performance exceeded 95% for all 4 birds.

The basic experimental procedure of Experiment 2 was then used for 29 days (an error prevented the inclusion of a planned 30th day), with a minimum number of observations required equal to eight and a value of P_R equal to .75. All other parameters were as in Experiment 2. Only the first categorization response in each trial was recorded.

Results

Figure 3 shows the average relative frequency of categorizing a sample as one produced by coin R as a function of the sample's diagnosticity. Each point is an average over the last 15 days of the 29 days of training. (This average over the last 15 of 29 days instead of over the last 5 of 15 days improves reliability and, at the same time, permits us to check that effects described in Experiments 1 and 2 are preserved after more extended training.) In the calculation of the relative frequency of categorization, only the first categorization in a trial was counted, so that categorization could not in part reflect interference by earlier categorizations. Figure 3 shows the same two phenomena as those found in Experiments 1 and 2. First, categorizations became increasingly nonoptimal as sample size increased. The right-most panel in Figure 3 closely parallels the corresponding panel in Figure 2. Second, Figure 3 shows essentially the same contextual phenomenon as the corresponding part of Figure 2—namely, categorizations were clearly nonoptimal, even with samples with one observation, since larger samples appeared on other trials. As in Experiment 2, matching could be produced by intermediate sample sizes (see, for example, the panel for a sample size of four in Figure 3).

Discussion

The decrease in accuracy of categorization in Experiments 1 and 2 as samples became larger was not an artifact of counting categorizations of larger samples after those of smaller ones within the same trial. The same de-

Figure 3. The relative frequency of choosing the left key after a sample as a function of the obtained relative frequency with which that sample was produced by coin R. Only categorizations of samples with at least eight observations could be reinforced. Different columns correspond to samples with different numbers of observations, ranging from zero on the left to eight on the right.

crease in accuracy of categorizing larger samples was obtained in Experiment 3 even though only the first categorization in each trial was counted. Categorization of larger binomial samples with successive observations is indeed less accurate than that of smaller ones. Similarly, the effect of sample size on accuracy of categorization obtained in Experiments 1 and 2 is unlikely to have been a result of inadequate training, since Experiment 3 doubled the amount of training.

GENERAL DISCUSSION

The long-standing speculation that many real-world discriminations are statistical in nature motivates finding out how well discrimination depends on the statistical diagnosticity of samples or, more broadly, on how organisms can learn to categorize ambiguous evidence. Indeed, the emphasis, in the past few decades, on fuzzy, ill-defined, or naturalistic concepts (Estes, 1994; Medin, 1975; Wittgenstein, 1953), rather than on well-defined, binary concepts of the type previously popular (Bourne & Restle, 1959; Whitehead & Russell, 1910), no doubt reflects the same growing interest in how organisms categorize ambiguous evidence. This interest has produced a quite wonderful theoretical literature on the categorization performance of humans (Estes, 1994; Gluck & Bower, 1988; Medin & Edelson, 1988) that has yet to be duplicated with nonhuman animals.

The existing literature on the performance of nonhuman animals in probability-learning and probabilistic-discrimination experiments does not even give a clear picture of the dependency between discrimination and diagnosticity. The existing literature, however, is oriented around the dichotomous question of whether animals match or maximize (Graf, Bullock, & Bitterman, 1964; Shimp, 1966, 1973; Williams, 1988). The present results question the traditional orienting perspective that there is only a binary range of outcomes. Rachlin, Green, and Tormey (1988) reached a conclusion similar to this; they suggested that neither matching nor maximizing is exclusively correct. We go one step further: They feel matching is a useful tool, whereas we interpret the pres-

ent data to mean that, at least in the present task, matching is a wholly arbitrary special case. The present results show that there is a dependency of choice behavior on sample size and on sampling context, so that this dichotomy is simplistic. A binomial probabilistic task can certainly be arranged to produce a good approximation to either matching, if training is not extended too long (Graf et al., 1964), or maximizing (Shimp, 1966, the present Experiment 1, and Experiment 2 with one observation in a sample and a sampling context with no large samples). However, from the broader perspective of the present Figures 1, 2, and 3, these two outcomes—matching and maximizing—seem wholly arbitrary, with the more general result being a family of functions ranging from optimizing to indifference, depending on specific sampling characteristics of the task.

Thus, the traditional empirical question of whether certain nonhuman animals match or maximize, along with the various theoretical positions that motivate this question, needs revision. There is a need for more flexible and more context-sensitive theories of acquired adaptive performances (for related arguments, see Gallistel, 1990; Shimp, 1989, 1992; and Staddon & Bueno, 1991) because no theory currently available for nonhuman performances can handle the existing variety of functional relations between discrimination accuracy and diagnosticity of ambiguous evidence. Several otherwise useful accounts of various aspects of probabilistically reinforced choice behavior, including matching, melioration, momentary maximizing, and molar maximizing (Williams, 1988), fail to handle the way categorization of binomial samples depended here on sample size and on sampling context. While any or all of these might prove capable of impressive future generalization, at the present time, none seems to handle the present data. Matching (Herrnstein, 1970; Williams, 1988) does not seem to explain the effects of sample size or sampling context and seems to us to overly emphasize the conceptual status of matching, which seemed here to be merely an arbitrary special case. Melioration (Herrnstein, 1982) has the same problems. Similarly, momentary maximizing (Hinson & Staddon, 1983; Shimp,

1966) can handle neither the context effect nor the systematic sampling effects. It is not obvious how either momentary maximizing or molar maximizing (Staddon & Motheral, 1978) can handle any of the present results, except perhaps the few special cases of optimal performances with small sample sizes and appropriate sampling context. We believe that the failure of these theories for one of the simplest of all possible probabilistic discriminations defines a situation where conceptual poverty may be hindering progress in understanding how animals categorize ambiguous evidence. Even the otherwise powerful theories of human performances (Estes, 1994; Gluck & Bower, 1988; Medin & Edelson, 1988) do not seem to deal with the temporal features of a task, which we can see from the present data appear to be central to a general understanding of intuitive inference. It would seem that a richer, more flexible, and more powerful theory than any currently in existence is needed to guide future empirical research in this area. In any case, no matter what an adequate theory for these phenomena will look like, one inescapable precondition for its development is a consistent database. Now that the present results show how the major inconsistency in the existing literature can be resolved, it may prove more constructive to examine the possible underlying mechanisms for discrete-trials, probabilistically reinforced choice behavior.

These mechanisms may include both attention and memory. Consider first how the effect of sample size may depend on memory. As the number of observations in a sample increases, an animal may find it increasingly difficult to remember the number of "heads" and "tails" and accordingly may misjudge the sample's diagnosticity. In this case, even if the subject knew the diagnosticity of the observed sample, it might categorize the sample incorrectly because it would use the remembered, not the observed, sample's diagnosticity. Or, the subject might not be able to learn the diagnosticity of a larger sample as easily as that of a smaller sample. Finally, contexts in which samples of a particular size appear relatively infrequently might disadvantage learning about them simply by virtue of the relatively few opportunities to experience them. This possibility could explain the poor performance on small samples in several of the contexts in the present experiments where small samples were relatively infrequent.

In any case, it is clear that one avenue along which a theory for intuitive statistical inference might develop would involve at least two separable components: a mechanism for discriminating the numbers of different types of component elements in a stimulus compound, and a mechanism for learning and discriminating reward likelihoods associated with those numbers. A numerical competence of the former type has been demonstrated by Honig and Stewart (1993), who showed that pigeons are able to discriminate between two categories of stimuli that differ only in terms of the relative frequencies of two component stimuli. Postdiscrimination generaliza-

tion gradients suggested that pigeons can discriminate "relative numerosity," although how to characterize the mechanism underlying such a competence is not yet clear. A numerical competence of the latter type presumably underlies performance of pigeons in probabilistic discrimination experiments, where pigeons discriminate likelihoods with which the relative frequencies of "heads" and "tails" in binomial samples diagnose different categories. The literature on probability learning historically was viewed as focused on this competence. That literature still provides several different potential mechanisms to describe the learning of likelihoods associated with different relative frequencies of component stimuli (Myers, 1976; Neimark & Estes, 1967). It is not at all clear, however, how such mechanisms could handle the sampling context effects identified in the present experiments.

Recall that this effect of sampling context may be due to an attentional mechanism. That is, the effect can be interpreted in terms of a tendency for the subject to attend less carefully to an initial observation when it expects subsequent observational opportunities. Whether a theory ultimately explains the context effect in terms of attention, as we have chosen to do here, or some other process, it is clear that the theory will have to be able to handle context effects in a way current accounts do not.

While the detailed articulation of an adequate theory for intuitive statistical inference in nonhuman animals is not yet a reality, available data do encourage the metatheoretical assumption that much everyday behavior will be susceptible to interpretation in terms of such a theory. The orderly functional relations obtained here are those one might expect if organisms were generally adapted to deal with subtle but orderly statistical relations in the natural world. Now that it is clear that at least some nonhuman animals can handle the constant-probability situation underlying the binomial and multinomial distributions, it becomes a matter for future research to examine the case, presumably more common in the natural world, where animals have to categorize samples deriving from sampling probabilities that vary in much more complex ways. For instance, naturalistic sampling probabilities presumably often vary from moment to moment so that sequential features of a sample could be highly diagnostic. In any case, it is to be hoped that laboratory research on the categorization of samples in ever more complex sampling situations eventually will be able to inform and support research on naturalistic foraging, where it is widely assumed that animals can estimate the profitability of a patch (Devenport & Devenport, 1993; Stephens & Krebs, 1986).

REFERENCES

- BOURNE, L. E., & RESTLE, F. (1959). A mathematical theory of concept identification. *Psychological Review*, *66*, 278-296.
- BRUNSWICK, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology*, *25*, 175-197.
- DEVENPORT, J. A., & DEVENPORT, L. D. (1993). Time-dependent deci-

- sions in dogs (*Canis familiaris*). *Journal of Comparative Psychology*, **107**, 169-173.
- ESTES, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- ESTES, W. K., BURKE, C. J., ATKINSON, R. C., & FRANKMANN, J. P. (1957). Probabilistic discrimination learning. *Journal of Experimental Psychology*, **54**, 233-239.
- FELLER, W. (1950). *An introduction to probability theory and its applications* (Vol. 1). New York: Wiley.
- GALLISTEL, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.
- GIGERENZER, G., & MURRAY, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- GLUCK, M. A., & BOWER, G. H. (1988). Evaluating an adaptive network model for human learning. *Journal of Memory & Language*, **27**, 166-195.
- GRAF, V., BULLOCK, D. H., & BITTERMAN, M. E. (1964). Further experiments on probability-matching in the pigeon. *Journal of the Experimental Analysis of Behavior*, **7**, 151-157.
- HERRNSTEIN, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, **13**, 243-266.
- HERRNSTEIN, R. J. (1982). Melioration as behavioral dynamism. In M. L. Commons, R. J. Herrnstein, & H. Rachlin (Eds.), *Quantitative analyses of behavior: Vol. 2. Matching and maximizing accounts* (pp. 433-458). Cambridge, MA: Ballinger.
- HINSON, J. M., & STADDON, J. E. R. (1983). Hill-climbing by pigeons. *Journal of the Experimental Analysis of Behavior*, **39**, 25-47.
- HONIG, W. K., & STEWART, K. E. (1993). Relative numerosity as a dimension of stimulus control: The peak shift. *Animal Learning & Behavior*, **21**, 346-354.
- MEDIN, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 131-169). New York: Academic Press.
- MEDIN, D. L., & EDELSON, S. M. (1988). Problem structure and the use of base rate information from experience. *Journal of Experimental Psychology: General*, **117**, 68-85.
- MYERS, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 3, pp. 131-205). Hillsdale, NJ: Erlbaum.
- NEIMARK, E. D., & ESTES, W. K. (Eds.) (1967). *Stimulus sampling theory*. San Francisco: Holden-Day.
- RACHLIN, H., GREEN, L., & TORMEY, B. (1988). Is there a decisive test between matching and maximizing? *Journal of the Experimental Analysis of Behavior*, **50**, 113-123.
- SHIMP, C. P. (1966). Probabilistically reinforced choice behavior in pigeons. *Journal of the Experimental Analysis of Behavior*, **9**, 443-455.
- SHIMP, C. P. (1973). Probabilistic discrimination learning in the pigeon. *Journal of Experimental Psychology*, **87**, 292-304.
- SHIMP, C. P. (1989). Contemporary behaviorism versus the old behavioral straw man in Gardner's *The Mind's New Science: A History of the Cognitive Revolution*. *Journal of the Experimental Analysis of Behavior*, **51**, 163-171.
- SHIMP, C. P. (1992). Computational behavior dynamics: An interpretation of Nevin (1969). *Journal of the Experimental Analysis of Behavior*, **57**, 289-299.
- SHIMP, C. P., & HIGHTOWER, F. A. (1990). Intuitive statistical inference: How pigeons categorize binomial samples. *Animal Learning & Behavior*, **18**, 401-409.
- STADDON, J. E. R. (1988). Learning as inference. In R. C. Bolles & M. D. Beecher (Eds.), *Evolution and learning* (pp. 59-77). Hillsdale, NJ: Erlbaum.
- STADDON, J. E. R., & BUENO, J. L. O. (1991). On models, behaviorism and the neural basis of learning. *Psychological Science*, **2**, 3-11.
- STADDON, J. E. R., & MOTHERAL, S. (1978). On matching and maximizing in operant choice experiments. *Psychological Review*, **85**, 436-444.
- STEPHENS, D. W., & KREBS, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- WHITEHEAD, A. N., & RUSSELL, B. (1910). *Principia mathematica* (Vol. 1). Cambridge: Cambridge University Press.
- WILLIAMS, B. A. (1988). Reinforcement, choice, and response strength. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' Handbook of experimental psychology: Vol. 2. Learning and cognition* (pp. 167-244). New York: Wiley.
- WITTGENSTEIN, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). New York: Macmillan.

(Manuscript received April 13, 1994;
revision accepted for publication November 11, 1994.)