ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 3, 1–11 (1968)

Intuitive Statistical Inferences about Diffuse Hypotheses[1]

CAMERON R. PETERSON AND RICHARD G. SWENSSON
University of Michigan

When data are sampled from a population and subjects revise probability estimates about which population is being sampled, their revisions are less than the optimal amount calculated by using Bayes's theorem; they are conservative. The experiments reported here used binomial populations with proportions that were either defined precisely by a display or defined diffusely by a sample of data. The experimenter randomly selected one of two populations and then sampled data from the selected population. The subjects made very nearly Bayesian revisions on the basis of the first datum sampled, but became markedly conservative when the task required aggregating evidence across a sequence of data. This result was independent of whether population proportions were defined precisely or diffusely.

Man comes to know his world largely by inference. He observes only a portion of his environment, and draws general conclusions on the basis of such observations. This generalization from limited observations is analogous to statistical inference, the process by which a statistician uses samples of data as a basis for making inferences about parent populations. In order to investigate the process by which man makes inferences from samples to populations, several experiments have used the following paradigm: A set of alternative hypotheses specify different populations and the subject is shown data sampled from one of them. Upon observing each datum, the subject becomes more or less sure of which hypothesis is correct and reflects this change of opinion by revising probabilities he has assigned to the hypotheses. Such a revision of probabilities is interpreted as intuitive statistical inference. The experiments have usually shown that the intuitive inferences are conservative; subjects revise the probability estimates less than the optimal amount as prescribed by formal Bayesian models of statistical inference (e.g., Edwards, Lindman, and Phillips, 1965).

[1]The research reported here was conducted in the Engineering Psychology Laboratory, Institute of Science and Technology, The University of Michigan. The research was supported by USPHS Fellowship MF 12.012-02, and by the Air Force Office of Scientific Research under Contract AF 49(638)-1731.

These previous experiments investigated probability revision when the hypotheses were specific, i.e., each hypothesis specified a particular value for the population parameter. In nonlaboratory situations, however, hypotheses are formed not only by explicit definition, but often by the less exact knowledge provided by data previously sampled from the hypothesized populations. A hypothesis about a population known only indirectly through observing a sample of data is not specific; it is diffuse because a sample of data provides a somewhat vague estimate of the population parameter.

The experiments reported below investigated intuitive inferences made about diffuse as well as about specific hypotheses.

EXPERIMENT I

In order to contrast the effects of specific and diffuse hypotheses upon intuitive inferences, a single datum was sampled from one of two possible binomial populations; subjects used that datum as a basis for making an inference about the population from which the datum was sampled. The binomial probability of the datum was P for one population and $1—P$ for the other. The value of P was defined explicitly for specific hypotheses but only vaguely, by a sample of data from the population, for diffuse hypotheses.

The quality of performance on the inference task was measured by comparing each subject's inference with the corresponding optimal inference. Bayes's theorem specifies the optimal inference as follows:

$$\frac{P(H_a|D)}{P(H_b|D)} = \frac{P(D|H_a)}{P(D|H_b)} \frac{P(H_a)}{P(H_b)} \tag{1}$$

where $P(H_i)$ is the probability of hypothesis H_i prior to the observation of a datum, $P(D|H_i)$ is the probability of sampling datum D if H_i is true, and $P(H_i|D)$ is the revised probability of H_i based upon the information provided by D. Equation 1 specifies that the odds of H_a to H_b should be revised from the right-hand ratio, the prior odds, to the left-hand ratio, the posterior odds, as a result of sampling datum D. The optimal amount of revision depends upon the middle ratio, the likelihood ratio. Since prior odds and estimated posterior odds are obtained from each subject for each inference, the amount of his probability revision is calculated by inferring what likelihood ratio would have led to that revision from prior to posterior odds. The logarithm of the likelihood ratio (LLR) is the appropriate measure of revision (see Peterson, Schneider, and Miller, 1965). A comparison of the theoretical LLR with the inferred LLR of subjects provides a measure of performance on the inference task.

Method

Experimental Design

For specific hypotheses, the experimenter directly displayed the compositions of each binomial population; he displayed the value of each $P(D|H)$. The process of establishing $P(D|H)$ for each diffuse hypothesis was somewhat more complex. Instructions first specified that a population proportion had been selected by random procedure prior to the experiment such that all proportions between 0 and 1 were equally likely. A random sample, hereafter called the defining sample, of binary events from the selected population further defined the diffuse hypothesis. Under these conditions, if a defining sample of size n contains r examples of datum D, the probability that the next datum drawn will also be D is equal to $(r+1)/(n+2)$, which is the expected value of the population proportion (see the appendix for a derivation of this conclusion). This expected value serves as $P(D|H)$ for each diffuse hypothesis.

The contrast in the sources of $P(D|H_i)$ for the specific and diffuse hypotheses illustrates an important characteristic of diffuseness. For specific hypotheses $P(D|H_i)$ is defined explicitly. The population proportion is known precisely. In the diffuse case, however, $P(D|H_i)$ is an expected value. It is the expected value of the probability distribution over all possible population proportions; the true parameter cannot be known precisely because the knowledge derives from sampling only a portion of the entire population.

When a sample of data diffusely defines the population parameter the definition can be based on various sized samples. Therefore, the number of data used in defining diffuse hypotheses was varied from 1 to 19 as an independent variable. Furthermore, previous research has shown that the quality of intuitive inference varies with diagnostic value of data (Peterson and Miller, 1965). Accordingly, for both specific and diffuse hypotheses, the experiment used four levels of diagnosticity equivalent to the symmetrical binomial proportions of .60–.40, .67–.33, .75–.25, and .90–.10.

Subjects. Fifteen paid University of Michigan men students served as subjects in groups of either 3, 4, or 5 each.

Apparatus and Procedure

As the physical data generating model associated with each hypothesis, subjects were instructed to imagine sampling at random with replacement from a large urn filled with red and blue poker chips. The procedure contained three stages; the first required inferences about specific hy-

potheses, the second about diffuse hypotheses, and the third about specific hypotheses again. For each inference task the experimenter selected one of two imaginary urns, either the predominantly red urn or the predominantly blue urn, by the toss of a fair coin. One chip was then sampled at random from the selected urn (the experimenter explained that all sampling had been done prior to the experiment and simply read the sample from a list). On the basis of the sample, the experimenter instructed the subjects to estimate from which urn it was more likely that the chip had been sampled and how many times more likely. Thus, the estimates were in the form of posterior odds. The subjects were to base inferences on their knowledge about the situation: the relative composition of the two urns, the fact that one of the two was chosen by a random procedure, and the information provided by the datum sampled from the chosen urn. Each subject then recorded his estimated odds on an answer sheet and the experimenter progressed to the next inference task.

For specific hypotheses, a disk display represented the composition of each urn, i.e., the relative proportions of red and blue chips. In the 60–40 case, for example, the display indicated 60% red chips in one urn and 40% red chips in the other. For the diffuse hypotheses, the experimenter instructed the subjects to imagine a very large urn containing thousands of red and blue chips. The proportion of red chips in the urn was selected by a random procedure prior to the experiment, such that all percentages were equally likely. Then the experimenter provided the subjects with information about the population proportion by reading off and displaying an imaginary sequence of chips drawn from the urn.

The subjects needed only to observe the display in order to understand the instructions about specific hypotheses. Their comprehension of the instructions about diffuse hypotheses was evaluated by requiring each subject to make a point estimate of the value of the population proportion after observing each chip in the defining sample. Each subject made his point estimate by sliding a marker, along a scale indexed from 0% red to 100% red, to the point that divided the possible percentages of red into two intervals. He was to select the point that made it equally likely that the true percentage of red was contained in either interval.[2] The subject began by setting his marker at 50%. Upon observing each

[2] The point that partitions a probability distribution into two equal areas is the median of that distribution. An earlier experiment (Peterson and Miller, 1964) and a pilot test of the present procedure both indicated that it is more appropriate for subjects to estimate medians than means. Medians are close to the means used in this experiment and so were used for the point estimates for reasons of experimental convenience.

sampled chip, he repartitioned the scale into two equally likely intervals based upon the cumulative impact of the data in the defining sample.

The crucial inference task with diffuse hypotheses came at the end of each defining sample. The experimenter instructed the subjects that there was a second urn which contained the reverse composition of the first urn (from which the defining sample was drawn). The experimenter selected either the first or second urn by the toss of a coin and then drew a single chip out of the selected urn. On the basis of the chip drawn, the subject indicated whether it was more likely that the chip was sampled from the first or the second urn and how many times more likely. This procedure continued through all of the diffuse hypothesis conditions.

The third stage of the experimental procedure then replicated the first stage, again using specific hypotheses, in an effort to increase the stability of the measures.

Results

All inferences were evaluated by comparing absolute values of inferred log likelihood ratios with corresponding theoretical values. The log likelihood ratio is in the form of log $P/(1—P)$. In an effort to achieve comparability of results this transformation was used for all data analyses throughout the experiment.

Comprehension of Diffuse Hypotheses

Recall that inferences with diffuse hypotheses required the subjects to use two classes of samples. First, populations were defined diffusely by means of what we have called *defining* samples, and then an *informational* sample of a single datum provided information about which population had been selected. The left graph in Fig. 1 refers to inferences based on the first kind of sample. It indicates how accurately the subjects comprehended data that defined, diffusely, each population. All data points were taken from the final estimate of each defining sample. The abscissa indicates the n of the defining sample and the ordinate refers to the absolute value of the log $P/(1—P)$ transformation of the point estimate of the proportion of red chips in the urn. This transformation increases from zero as the median of the probability distribution over P departs from 0.5. Solid lines connect data points and the nearly horizontal dashed lines show corresponding theoretical values.

The subjects generally comprehended the implications of defining samples very well. There is a close parallel between theoretical and data values. The major exception to this principle resides in the two smallest defining samples, with n's of 1 and 2, respectively. Here the response

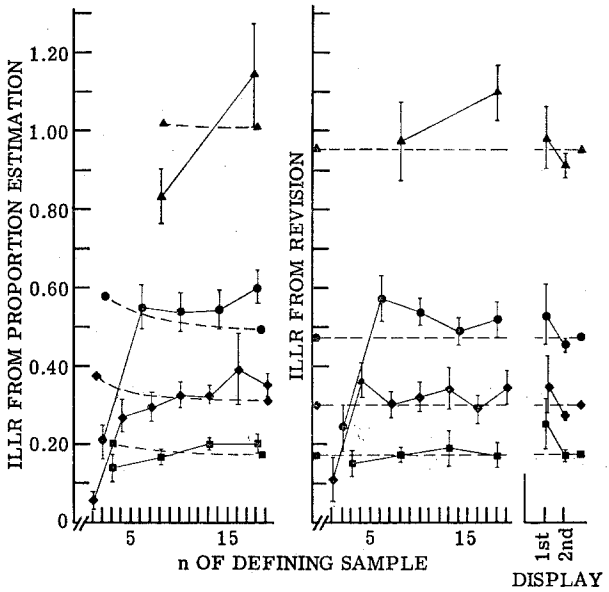

FIG. 1. Inferred log likelihood ratio as a function of the n of the defining sample for diffuse hypotheses. For specific (displayed) hypotheses, probability revisions were made either before (1st) or after (2nd) the diffuse hypothesis task. The four different symbols refer to the four different levels of diagnostic value. Dashed lines refer to theoretical values. Vertical lines show plus and minus one standard error of each mean.

functions fall sharply below corresponding theoretical values. The subjects responded as if a defining sample of only 1 chip or 2 chips contained very little information about the population proportion.

INFERENCES WITH DIFFUSE AND SPECIFIC HYPOTHESES

The middle panel in Fig. 1 refers to inferences between two diffuse hypotheses, inferences based on informational samples. Most were nearly optimal. Inferences were substantially conservative in only two instances, again with defining samples of 1 and 2. Comparison of both graphs suggests that the conservatism in these two instances can be accounted for by a suboptimal comprehension of the diffuse hypotheses, i.e., by earlier conservatism in interpreting the implications of the defining sample (this correspondence is analogous to the consistency in the revision of probability estimates reported by Beach, 1966; by Peterson, Ulehla, Miller, Bourne, and Stilson, 1965; and by Peterson, DuCharme, and Edwards, in press).

The near optimality in making inferences about most diffuse hypotheses was also evident when subjects made inferences about specific hypotheses, hypotheses for which the experimenter displayed the precise

composition of the population. The results of inferences about these specific hypotheses are displayed on the right hand side of the right graph in Fig. 1. The inferences labeled 1st were those made prior to the diffuse hypothesis portion of the experiment and those labeled 2nd were made after the diffuse hypothesis portion. The second set of inferences appears slightly more conservative than the first, but the overriding result is that inferences with specific hypotheses were nearly optimal, just as were inferences about diffuse hypotheses.

Discussion

The subjects evidenced a high level of performance on inference tasks throughout the experiment. They made nearly optimal inferences about specific hypotheses; they correctly comprehended the defining samples for most diffuse hypotheses; and they made nearly optimal inferences about those diffuse hypotheses. The only exception to this high level of performance was in the case in which the defining samples contained only one or two data. In those cases, conservative inferences can be explained by corresponding conservative interpretations of the defining samples. Why those interpretations were conservative remains unexplained; perhaps people are unwilling to believe that only one or two data provide much information for the definition of populations.

Why did this experiment fail to yield the conservatism found in the previous experiments on intuitive inference? Experiment II was conducted for the purpose of answering that question.

EXPERIMENT II

The results of several experiments indicate that part of conservative human inference is due to a difficulty in the aggregation of evidence across a sequence of data (e.g., Edwards, 1966; Phillips, 1966; Schum, 1966). This may explain why Exp. I found so little of the conservatism prevalent in previous experiments. Each of the informational samples in Exp. I contained only a single datum whereas informational samples in many earlier experiments contained sequences of up to 48 data (Edwards *et al.*, 1965; Peterson *et al.*, 1965; Phillips and Edwards, 1966). Accordingly, Exp. II replicated most of Exp. I except that all informational samples were extended to contain sequences of data and subjects revised their estimates about the identity of the selected population after each datum in the sequence.

Method

The apparatus and the procedure of Exp. II were essentially the same as those of Exp. I except for the deletion of several defining samples of intermediate size. The essential change in Exp. II was that each

informational sample now contained a sequence of five data; subjects revised their odds estimates about the population from which the data were being sampled after the presentation of each datum in the sequence.

Subjects. Eighteen men students of the University of Michigan served as paid subjects in groups of 3, 4, or 5 each.

Results

The left graph in Fig. 2 displays log likelihood ratios from inferences based on the first datum in each sequence. Inferences about specific hypotheses are slightly more conservative than those in Exp. I; all data points fall below the theoretical values. The amount of conservatism,

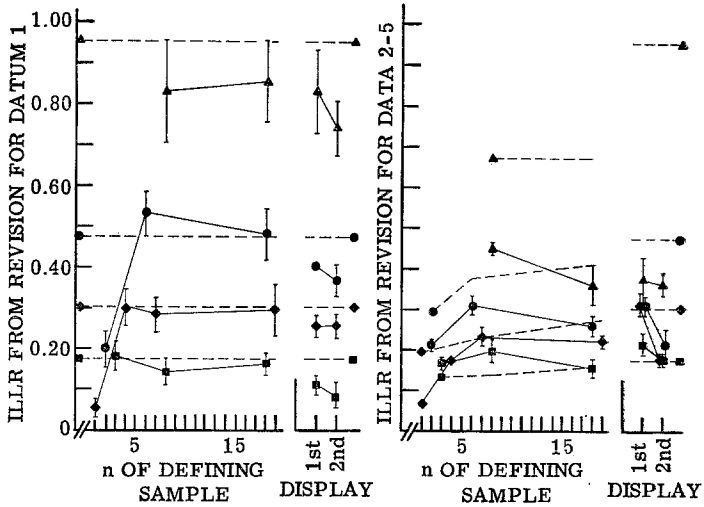

FIG. 2. Inferred log likelihood ratio as a function of n of the defining sample for diffuse hypotheses. For specific (displayed) hypotheses, probability revisions were made either before (1st) or after (2nd) the diffuse hypothesis task. The four different symbols refer to the four different levels of diagnostic value. Dashed lines refer to theoretical values. Vertical lines show plus and minus one standard error of each mean.

however, is overshadowed by the relatively small distance separating data points from corresponding theoretical lines. Inferences about diffuse hypotheses based on the first datum replicated the results of Exp. I; inferences were nearly optimal except with defining samples of only 1 or 2 data.

The right hand graph of Fig. 2 is based upon the mean across Trials 2–5 of absolute inferred log likelihood ratio for each trial. Here is the kind of conservatism that has pervaded previous experiments. In the case of specific hypotheses, the average amount of inference was generally

less than that called for theoretically, and the degree of conservatism was much greater for high theoretical log likelihood ratios. As in previous experiments, intuitive inferences failed to discriminate adequately among the experimental conditions; the data points frequently overlap each other.

Results are a little less clear in the case of diffuse hypotheses. It is the nature of these diffuse hypotheses that the theoretical values of log likelihood ratios do not remain constant across trials of an informational sample (calculations of these theoretical values are based on equations presented in the appendix). The mean absolute value of the theoretical ratios across Trials 2–5 are indicated by points connected by dashed lines. These values are, for the most part, less than the corresponding theoretical values for specific hypotheses. Results show that the average amount of inference for diffuse hypotheses was more conservative for Trials 2–5 than for Trial 1. However, subjects seemed to discriminate among the experimental conditions better with diffuse hypotheses than with specific hypotheses. Even though theoretical values became less separated during Trials 2–5, the relative amounts of revision remained separated.

Discussion

Experiments I and II yield three clear conclusions. First, subjects make excellent inferences from data when their task is to learn about the composition of populations by observing sampled data. They reflect nearly optimal understanding of defining samples except when the sample size is very small. The reason for this exception remains unexplained.

Second, intuitive inferences about diffuse hypotheses are just as nearly optimal, and sometimes more so, than intuitive inferences about specific hypotheses. The fact that the former are considerably more difficult to deal with conceptually failed to deteriorate the performance of subjects in these experiments. Perhaps diffuse hypotheses are more compatible with human inference because they are more representative of nonlaboratory inference tasks.

Third, these experiments support the conclusion of some earlier experiments—that at least one locus of conservative inference resides in the task of aggregating evidence across sequential samples. Intuitive inferences were nearly optimal when based upon only a single datum or upon the first datum in a sequence. Subsequent data in sequences, however, elicited suboptimal, conservative inferences. The mounting evidence that the task of aggregating evidence leads to conservative inferences calls for more research aimed at an explanation of the phenomenon.

APPENDIX

A diffuse hypothesis in the present context means that the value of the binomial population proportion, p, is not known precisely. Instead, the probability that each of the possible values of p, from 0 to 1.0, is the true population proportion is given by a beta density function. The beta distribution is the natural conjugate to the binomial distribution, which means that if one's opinions about p are described by a beta distribution and a random sample is taken from the binomial population then the Bayesian posterior opinion about p will be described by a beta distribution with new parameters.

In this experiment subjects were told initially that the value of p was selected randomly such that all values from 0 to 1.0 were equally likely. This procedure defines a rectangular probability distribution over p which can be described by a beta density function (Raiffa and Schlaifer, 1961, p. 263, Eq. 9-9) having parameters: $r' = 1$ and $n' = 2$. Applying the results of Raiffa and Schlaifer (p. 263, Eq. 9-10), if a defining sample in the present experiment contained r red chips and n-r blue chips the posterior beta distribution over p would be

$$P(p|r,n;r' = 1, n' = 2) = \frac{r!(n-r)!}{n!} p^r(1-p)^{n-r} \tag{1}$$

The probability of drawing r_s red chips in n_s draws from a diffuse hypothesis defined by Eq. 1 is shown by Raiffa and Schlaifer (P. 265, Eq. 9-18 and P. 237, Eq. 7-76) to be

$$P(r_s|r,n;r' = 1,n' = 2;n_s) = \frac{(r_s + r)!(n_s + n - r_s - r)!n_s!(n+1)!}{r_s!r!(n_s - r_s)!(n-r)!(n_s + n + 1)!} \tag{2}$$

The probability of drawing a single red chip is found by setting $r_s = n_s = 1$ and is equal to $(r+1)/(n+2)$.

In the odds revision portion of the experiment, Urn 2 was known to have the reverse composition of Urn 1. This means that the beta parameters for the diffuse hypothesis over p for Urn 2 are the reverse of those which described Urn 1 (i.e., $n_2 = n_1$, $r_2 = n_1$—r_1). This likelihood ratio in favor of Urn 1 over Urn 2, based upon a sample of r_s red chips in n_s draws, is therefore;

$$\begin{aligned} L(r_s,n_s|H_1:H_2) &= \frac{P(r_s|r,n;r' = 1,n' = 2;n_s)}{P(r_s|n-r,n;r' = 1,n' = 2;n_s)} \\ &= \frac{(r_s + r)!(n_s + n - r_s - r)!}{(r_s + n - r)!(n_s - r_s + r)!} \end{aligned} \tag{3}$$

For a revision sample consisting of a single chip, $n_s = 1$ and Eq. 3 simplifies to $(r + 1)/(n—r + 1)$ or $(n—r + 1)/(r + 1)$, depending upon whether a red ($r_s = 1$) or blue ($r_s = 0$) chip is sampled.

REFERENCES

BEACH, L. R. Accuracy and consistency in the revision of subject probabilities. *IEEE Transactions on Human Factors in Electronics,* 1966, **7**, 29–37.

EDWARDS, W. Non-conservative probabilistic information processing systems. ESD-TDR-66-404, University of Michigan, Institute of Science and Technology Report 05893-22-F, August 1966.

EDWARDS, W., LINDMAN, H., AND PHILLIPS, L. D. Emerging technologies for making decisions. *In* T. M. Newcomb (Ed). *New directions in psychology II.* New York: Holt, Rinehart and Winston, 1965, 261–325.

PETERSON, C. R., DUCHARME, W., AND EDWARDS, W. Sampling distributions and probability revisions. *Journal of Experimental Psychology* (in press).

PETERSON, C. R., AND MILLER, A. Mode, median and mean as optimal strategies. *Journal of Experimental Psychology,* 1964, **68**, 363–367.

PETERSON, C. R., AND MILLER, A. J. Sensitivity of subjective probability revision. *Journal of Experimental Psychology,* 1965, **70**, 117–121.

PETERSON, C. R., SCHNEIDER, R. J., AND MILLER, A. J. Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology,* 1965, **69**, 522–527.

PETERSON, C. R., ULEHLA, Z. J., MILLER, A. J., BOURNE, L. K., JR., AND STILSON, D. W. Internal consistency of subjective probabilities. *Journal of Experimental Psychology,* 1965, **70**, 526–533.

PHILLIPS, L. D. Some components of probabilistic inference. Technical Report No. 1. Human Performance Center, University of Michigan. January 1966.

PHILLIPS, L. D., AND EDWARDS, W. Conservatism in a simple probability inference task. *Journal of Experimental Psychology,* 1966, **72**, 346–354.

RAIFFA, H., AND SCHLAIFER, R. *Applied statistical decision theory.* Boston: Graduate School of Business Administration, Harvard University, 1961, Pp. 237, 265.

SCHUM, D. A. Prior uncertainty and amount of diagnostic evidence as variables in a probabilistic inference task. *Organizational Behavior and Human Performance,* 1966, **1**, 31–54.

RECEIVED: May 22, 1967