Bulletin of the Psychonomic Society
1985, 23 (6), 509-512

Averaging rules and adjustment processes in Bayesian inference

LOLA L. LOPES
University of Wisconsin, Madison, Wisconsin

Two empirically well-supported research findings in the judgment literature are (1) that human judgments often appear to follow an averaging rule, and (2) that judgments in Bayesian inference tasks are usually conservative relative to optimal judgments. This paper argues that both averaging and conservatism in the Bayesian task occur because subjects produce their judgments by using an adjustment strategy that is qualitatively equivalent to averaging. Two experiments are presented that show qualitative errors in the direction of revisions in the Bayesian task that are well accounted for by the simple adjustment strategy. Also noted is the tendency for subjects in one experiment to evaluate sample evidence according to representativeness rather than according to relative likelihood. The final discussion describes task variables that predispose subjects toward averaging processes.

The Bayesian inference task is usually instantiated in terms of the "bookbag and poker chips paradigm" in which there are two competing hypotheses involving populations of binary events. For example, there may be two bookbags, one containing 70 red poker chips and 30 blue poker chips (the "red bag") and another containing 30 red chips and 70 blue chips (the "blue bag"). In most experiments, the experimenter ostensibly selects a bag at random and then draws samples from it one or more times. These samples are shown to the subject, usually sequentially, and the subject rates the strength of his or her belief about which bookbag was sampled.

According to Bayes's theorem, the probability that the bookbag sampled is red, say, given sample D1, is given by:

$$p(HR|D1) = \frac{p(D1|HR)p(HR)}{p(D1|HR)p(HR) + p(D1|HB)p(HB)} \quad (1)$$

where p(HR) and p(HB) are the prior probabilities of the red and blue bookbags, respectively, and p(HR|D1) is the posterior probability of HR. If a second sample is introduced, the process reiterates with the values for p(HR|D1) and p(HB|D1) replacing the old prior probabilities. That is,

$$p(HR|D1 \ \& \ D2)$$

$$\frac{p(D2|HR)p(HR|D1)}{p(D2|HR)p(HR|D1) + p(D2|HB)p(HB|D1)} \quad (2)$$

The writing of this paper and the research reported were supported by the Office of Naval Research (Contract N00014-81-K-0069). I am grateful to Gregg Oden and Martin Tolcott for helpful comments on an earlier draft.

The author's mailing address is: Department of Psychology, University of Wisconsin, Madison, WI 53706.

In general, human responses to this task are conservative relative to Bayesian responses; that is, human responses fall nearer neutral than Bayesian responses. Three conceptually distinct explanations have been given for conservatism: (1)*Misperception* which locates the error in the process of estimating the diagnostic impact of sample data; (2) *misaggregation* which locates the error in the process of integrating the information from multiple samples into composite responses; and (3) *response bias* which treats conservatism as an artifact of the response scale reflecting subjects' tendency to avoid extreme responses.

Although all three sources of error are likely to occur in Bayesian tasks, misaggregation is particularly important because it appears to figure most prominently in producing conservatism (Edwards, 1968). It probably also is easier to teach people improved methods for aggregation than it is to eliminate the other errors (Eils, Seaver, & Edwards, 1977; Lopes, 1982).

Averaging and Conservatism

Early evidence linking averaging and conservatism came from experiments showing that subjects' ratings in Bayesian tasks were often more like estimates of population proportion than like inferences from Bayes's rule (Beach, Wise, & Barclay, 1970; Shanteau, 1970, 1972). Shanteau hypothesized that the data could be modeled by a rule in which the response at each serial position is given by a weighted average of the scale values of the previous and current sample events. Such a rule would be conservative relative to Bayes's rule since averages always lie within the range of stimulus values, whereas Bayesian inferences are often more extreme.

Although the averaging rule fits inference judgments quantitatively, even better evidence for averaging came from qualitative experiments (Shanteau, 1975; Troutman & Shanteau, 1977), which showed that presentation of neutral or nondiagnostic information following previously presented diagnostic information causes subjects to revise

their judgments toward neutral. This result is expected under averaging but not under Bayes's theorem, which specifies that neutral information ought to have no impact on prior judgments.

Further evidence linking averaging and conservatism came from experiments by Eils et al. (1977), who hypothesized that, although untutored subjects appear to average, they might be better at judging the mean log likelihood ratio for a set of samples than they are at judging the cumulative log likelihood ratio, which is the more typical Bayesian judgment. The data supported the hypothesis: Log odds inferred from mean certainty judgments were closer to veridical than odds inferred from cumulative certainty judgments.

Averaging and Serial Adjustment Processes

Although averaging rules have been successful in accounting for judgment data, little attention has been directed at finding out why averaging occurs. The present research hypothesizes that averaging occurs because subjects adopt an adjustment strategy in which they integrate new information into "old" composite judgments by adjusting the old value as necessary to make the new value lie somewhere between the old value and the value of the new information. Such a process would be qualitatively equivalent to averaging, even though subjects would not average in any arithmetical sense of that term.

The hypothesized adjustment process differs both quantitatively and qualitatively from Bayes's theorem. This can be seen by rewriting Equation 2 as follows:

$$p(HR \mid D1 \,\&\, D2) = \frac{1}{1 + \dfrac{p(D1 \mid HB)p(HB)}{p(D1 \mid HR)p(HR)} \dfrac{p(D2 \mid HB)}{p(D2 \mid HR)}} \tag{3}$$

Focusing for illustration on HR, notice that if sample D2 favors HR [i.e., $p(D2 \mid HB)/p(D2 \mid HR) < 1$], then the value of $p(HR \mid D1 \,\&\, D2)$ will exceed the value of $p(HR \mid D1)$, because the denominator term is made smaller. Thus, if a new sample favors HR, then adjustments must be toward *increased* support for HR. Likewise, if a new sample favors HB, then adjustments must be toward *decreased* support for HR.

Under averaging, however, adjustments are not always made in the direction of increased support for the hypothesis favored by new evidence. To illustrate, consider a subject who is judging between HR and HB using a rating scale on which increased confidence in HR is associated with larger numbers. If the first sample favors HR moderately strongly (i.e., 5 red and 3 blue), then the subject should make an initial judgment at some value favoring HR. If the next sample also favors HR, but more strongly (e.g., 7 red and 1 blue), the subject should respond by adjusting upward toward, but not past, the value of the second sample. Such an adjustment would be directionally in accord with both the averaging rule and the Bayesian rule because $p(D2 \mid HB)/p(D2 \mid HR) < 1$, but it would be quantitatively conservative because the Bayesian response would necessarily be more extreme than the response to either sample taken alone.

If the samples were reversed, however, so that the weaker follows the stronger, the two rules make different predictions about the direction of the adjustment. Under the Bayesian rule, the adjustment would still be upward: Although the new sample is less favorable to HR than the old, the ratio $p(D2 \mid HB)/p(D2 \mid HR)$ is still less than one, so that $p(HR \mid D1 \,\&\, D2)$ should increase. Under averaging, however, the adjustment would be downward (i.e., toward neutral), because the value of the new sample is less favorable to HR than the value of the judgment based on the first sample. Thus, errors should be evident in the direction of subjects' adjustments when weaker sample evidence favoring a particular hypothesis follows stronger evidence favoring the same hypothesis.

EXPERIMENTAL TESTS OF THE DIRECTIONAL HYPOTHESIS

Two experiments were run to test the directional hypothesis previously described. Because the experiments were essentially identical except for the stimulus designs, they are discussed together.

Experimental Tasks

Subjects made imaginary judgments concerning the maintenance of milling machines based on samples of parts. Subjects in Experiment 1 judged whether machines needed maintenance or not. They were instructed that machines which are working properly are about as likely to produce parts that are a little too large as a little too small ($H_{50/50}$). Broken machines, on the other hand, were described as tending to produce parts of which about 75% are a little too large and 25% are a little too small ($H_{75/25}$).

Subjects in Experiment 2 judged which of two maintenance procedures a machine needs. They were instructed that machines needing one procedure tend to produce parts of which about 75% are a little too small and 25% are a little too large ($H_{25/75}$), whereas machines needing the other procedure tend to produce parts of which about 75% are a little too large and 25% are a little too small ($H_{75/25}$).

Stimulus Designs

The stimulus design for Experiment 1 was a 7 × 7 (first-sample × second-sample) factorial design in which the levels of both factors comprised the same seven sample distributions. The distributions were, for large and small parts, respectively: 3/7, 4/6, 5/5, 6/4, 7/3, 8/2, and 9/1. The design for Experiment 2 was the same but with nine levels on each factor: 1/9, 2/8, 3/7, 4/6, 5/5, 6/4, 7/3, 8/2, and 9/1.

Procedure

Subjects were run individually, using a computer controlled video terminal. In Experiment 1, each stimulus display comprised a box showing a sample of parts with a notation indicating whether this was the first or second sample. At the bottom of the display was a rating scale anchored at the left by "MILLING NORMALLY (50%)" and at the right by "MILLING TOO LARGE (75%)." The display for Experiment 2 was identical except that the rating scale was anchored with "MILLING TOO SMALL (75% SMALL)" at the left and "MILLING TOO LARGE (75% LARGE)" at the right.

The procedure for each trial was identical. Subjects read the information for the first sample and then rated the probability of the two hypotheses by moving a rating arrow along the response scale. After subjects signaled their response, the first sample was replaced by the second sample, and subjects revised their rating. Subjects then signaled their final response and initiated the next trial by returning the response arrow to the middle of the scale. There were 15 trials for practice followed by 2 replications of the stimulus design.

Subjects

The subjects for the two experiments were 41 and 39 student volunteers, respectively. In Experiment 1 they were all males; in Experiment 2 they were evenly divided on sex.

Results and Discussion

Ratings of single samples. To test the adjustment hypothesis, it is necessary to determine what values the subjects attached to the various sample types. This can be done by looking at responses to first samples. Averaged data are given in Table 1. Ratings have been scaled to run between 0 and 1.

It is best to begin with the results for Experiment 2. Ratings of the likelihood of $H_{75/25}$ increased essentially linearly from 1/9 samples to 9/1 samples, with 5/5 samples being rated as neutral. The close correspondence between ratings and proportion of large parts suggests that subjects produced initial ratings by using the sample proportion as a judgmental "anchor" (Tversky & Kahneman, 1974). The results of Experiment 1 are more difficult to understand. Nominally, the data are reasonable: Samples of 3/7 through 6/4 were rated as supporting $H_{50/50}$, and samples of 7/3 through 9/1 were rated as supporting $H_{75/25}$. But the data deviated from the norm ordinally in that 5/5 samples were judged to be more supportive of $H_{50/50}$ than either 3/7 or 4/6 samples.

Inspection of single subject data revealed that 38 out of 41 subjects could be assigned to one of three groups. The first group (15 subjects) ordered the samples according to likelihood ratio, with 3/7 samples being taken as stronger evidence for $H_{50/50}$ than 4/6 samples, and these in turn as stronger evidence than 5/5 samples. The second group (12 subjects) ordered the samples in inverse order to the norm: 5/5 samples were taken to be the strongest evidence of $H_{50/50}$, followed by 4/6 and 3/7 samples. These subjects appeared to judge the samples according to how representative they are of a 50/50 generating process (Kahneman & Tversky, 1972). The third group (11 subjects) ordered the samples 5/5, 3/7, 4/6. This mixed group appeared to be influenced by representativeness only if samples were "perfectly" representative.

Table 1
Mean Ratings of First Samples

Experiment 1		Experiment 2	
Sample L/S	Rating	Sample L/S	Rating
3/7	.20	1/9	.08
4/6	.23	2/8	.16
5/5	.16	3/7	.24
6/4	.45	4/6	.31
7/3	.73	5/5	.49
8/2	.85	6/4	.67
9/1	.94	7/3	.74
		8/2	.82
		9/1	.91

Note—A rating of 0 indicates complete confidence in $H_{50/50}$ for Experiment 1 and $H_{25/75}$ for Experiment 2. A rating of 1 indicates complete confidence in $H_{75/25}$. L/S = large/small.

The fact that many subjects in Experiment 1 appeared to use representativeness rather than relative likelihood in evaluating samples causes no problems for testing the adjustment hypothesis other than making it necessary to perform separate tests for the various groups. But the result is puzzling. Although Wallsten and Barton (1982) reported a similar result, the present effect did not occur for Experiment 2 or for the samples supporting $H_{75/25}$ in Experiment 1. One possibility is that subjects may have believed mistakenly that normal machines always produce 50/50 samples. This is unlikely, however, since subjects were explicitly instructed otherwise. A better possibility seems to be that the task violated the conventional semantics of what it means for a machine to be working normally. Certainly it is odd linguistically to say that the best evidence for the normal functioning of a machine is that it produces a sample with an abnormally large number of small parts. Thus, subjects may have slipped into using a hybrid strategy in which samples favoring $H_{75/25}$ were evaluated with respect to both hypotheses, whereas samples favoring $H_{50/50}$ were evaluated only with respect to $H_{50/50}$.

Adjustments for second samples. Table 2 gives the proportions of directionally correct adjustments (relative to Bayesian norms) for strong-weak sample orders and weak-strong sample orders. Seven subjects have been eliminated from the analysis of Experiment 1 (6 from the likelihood ratio group and 1 from the mixed group), and 15 subjects have been eliminated from the analysis of Experiment 2. Although these subjects' data would strongly support the adjustment hypothesis, it seemed best to exclude them because individual analyses revealed that their final ratings significantly reflected only the value of the second sample.

The results are in Table 2. In both experiments, adjustments were almost always made in the correct direction for weak-strong sample pairs and much less often for strong-weak pairs, $p < .05$ for likelihood ratio subjects, and $p < .01$ for all other groups. Note that the mean proportion of directionally correct adjustments is somewhat greater for the likelihood ratio group of Experiment 1 and for Experiment 2, generally, than for the representativeness group and the mixed group of Experiment 1. This is because these groups include subjects who were qualitatively Bayesian (i.e., subjects who demonstrated both ordinally proper evaluation of the relative likelihoods of the various samples and directionally proper adjustment following the second sample). There were five such subjects in each experiment.

The prevalence of subjects who based their final judgment only on the second sample was unexpected. Although, such subjects can be interpreted as obeying an

Table 2
Proportion of Directionally Correct Adjustments for Homogeneous Sample Pairs

	Strong-Weak	Weak-Strong	F_{diff}	(df)
	Experiment 1			
Likelihood ratio group	.56	.86	5.65*	(1,8)
Representativeness group	.30	.86	74.14†	(1,11)
Mixed group	.31	.94	94.02†	(1,9)
	Experiment 2			
	.40	.92	80.93†	(1,23)

Note—In Experiment 1 there were 18 strong-weak and 18 weak-strong comparisons per subject, and in Experiment 2 there were 24 strong-weak and 24 weak-strong comparisons per subject.
*$p < .05$. †$p < .01$.

averaging rule with extremely strong recency, it is more likely that they differed qualitatively from subjects who used both samples. The most obvious possibility, of course, is that these subjects mistakenly understood each sample to constitute a new trial. Although this possibility cannot be ruled out, several factors argue against it, including that (a) the instructions stressed that both samples of each pair came from the same machine, (b) the stimulus display reminded subjects that samples were either "first" or "second" samples, and (c) the trial structure was clearly divided into a "within machine" portion and a "between machine" portion.

A second possibility is that these subjects thought that the second sample from a machine ought not to be integrated with the first, but rather ought to be substituted. This might appear reasonable if samples were often disparate enough to suggest that they could not both be correct. Although there is no hard evidence to support this possibility, the fact that there were more such subjects in Experiment 2 than in Experiment 1 lends some credence to the view, because in Experiment 2 there were both more sample pairs in which the samples supported opposing hypotheses as well as a greater potential range in the magnitude of difference between samples.

DISCUSSION

The data demonstrate that the adjustments subjects make in the Bayesian task are consistent with averaging and inconsistent with Bayes's theorem. The view taken here is that such errors reflect subjects' use of a serial adjustment strategy in which new information is integrated with old information by adjusting the old judgment value toward the value of the new information. A basic corollary of this view is that subjects do not "choose" judgment rules in the sense that they "decide" how information ought to be combined. Rather, they choose adjustment processes that seem to "fit" the task, both in terms of the ease with which the process can be executed mentally and in terms of the degree to which the process generates plausible judgments.

In the case of Bayesian inference, a basic question concerns why subjects who are presented with a sample that they would independently judge to support a particular hypothesis, sometimes reduce their confidence in that hypothesis. This question is especially intriguing because there are some tasks in which subjects seem to follow exactly the sort of ratio rule that is specified by Bayes's theorem.

One potentially important stimulus variable may be whether the stimuli are explicitly "marked" for one hypothesis or another. For example, in an experiment by Oden (1979), subjects produced ratio data when judging whether a presented character was a T or an F. The featural dimensions that were varied were (1) the length of a horizontal bar at the upper left of the character whose presence is characteristic of T and not of F, and (2) the length of a horizontal bar at the middle right whose presence is characteristic of F and not of T. With stimuli such as these it is unlikely that a subject would ever adjust incorrectly because the evidence itself would absolutely mark the appropriate hypothesis.

In the Bayesian task, however, the stimuli are less explicitly marked. For example, although a sample of six small and four large parts in Experiment 2 would tend to suggest that the machine is "milling too small," there is nothing to rule out that the machine is "milling too large." Thus, there is nothing to prevent subjects from responding primarily to diagnostic strength rather than diagnostic sign and adjusting toward neutral (i.e., away from the hypothesis that is, in fact, better supported by the data).

Although the present experiments were not designed to examine weighting strategies, it is worth noting that another difference between simple averaging tasks and tasks in which ratio data are produced appears to be whether the weight of a stimulus element is directly related to its diagnostic strength. This potential task difference is supported by experiments reported elsewhere (Lopes, 1982), which show that if the relative diagnostic strength of samples is made salient to subjects by instructing them to process the more diagnostic sample first, the judgments show the kind of differential weighting characteristic of Bayesian judgments.

REFERENCES

BEACH, L. R., WISE, J. A., & BARCLAY, S. (1970). Sample proportions and subjective probability revisions. *Organizational Behavior & Human Performance*, **5**, 183-190.

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

EILS, L. C., SEAVER, D. A., & EDWARDS, W. (1977). *Developing the technology of probabilistic inference: Aggregating by averaging reduces conservatism* (Research Report 77-3). Los Angeles: University of Southern California, Social Science Research Institute.

KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-454.

LOPES, L. L. (1982). *Procedural debiasing* (Technical Report WHIPP 15). Madison, WI: Wisconsin Human Information Processing Program.

ODEN, G. C. (1979). A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 336-352.

SHANTEAU, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology*, **85**, 181-191.

SHANTEAU, J. C. (1972). Descriptive versus normative models of sequential inference judgment. *Journal of Experimental Psychology*, **93**, 63-68.

SHANTEAU, J. C. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, **39**, 83-89.

TROUTMAN, C. M., & SHANTEAU, J. C. (1977). Inferences based on non-diagnostic information. *Organizational Behavior & Human Performance*, **19**, 43-55.

TVERSKY, A., & KAHNEMAN, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

WALLSTEN, T. S., & BARTON, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 361-384.

(Manuscript received for publication June 10, 1985.)