2 Base rates in Bayesian inference

Michael H. Birnbaum

What is the probability that a randomly drawn card from a well-shuffled standard deck would be a heart? What is the probability that the German soccer team will win the next world championships?

These two questions are quite different. In the first, we can develop a mathematical theory from the assumption that each card is equally likely. If there are 13 cards each of hearts, diamonds, spades, and clubs, we calculate that the probability of drawing a heart is 13/52, or 1/4. We test this theory by repeating the experiment again and again. After a great deal of evidence (that 25% of the draws are hearts), we have confidence in using this model of past data to predict the future.

The second case (soccer) refers to a unique event that either will or will not occur, and there is no way to calculate a proportion from the past that is clearly relevant. One might examine records of the German team and those of rivals, and ask if the Germans seem healthy – nevertheless players change, conditions change, and it is never really the same experiment. This situation is sometimes referred to as one of *uncertainty*, and the term *subjective probability* is used to refer to psychological strengths of belief.

However, people are willing to use the same term, probability, to express both types of ideas. People gamble on both types of predictions – on repeatable, mechanical games of chance (like dice, cards, and roulette) with known risks, and on unique and uncertain events (like sports, races, and stock markets). In fact, people even use the term "probability" *after* something has happened (a murder, for example), to describe belief that an event occurred (e.g., that this defendant committed the crime). To some philosophers, such usage seemed meaningless. Nevertheless, Reverend Thomas Bayes (1702–1761) derived a theorem for inference from the mathematics of probability. Some philosophers conceeded that this theorem could be interpreted as a calculus for rational formation and revision of beliefs in such cases (see also Chapter 3 in this volume).

BAYES' THEOREM

The following example illustrates Bayes' theorem. Suppose there is a disease that infects one person in 1000, completely at random. Suppose there is a blood test for this disease that yields a "positive" test result in 99.5% of cases of the disease and gives a false "positive" in only 0.5% of those without the disease. If a person tests "positive", what is the probability that he or she has the disease? The solution, according to Bayes' theorem, may seem surprising.

Consider two hypotheses, H and not-H (denoted H'). In this example, they are the hypothesis that the person is sick with the disease *(H)* and the complementary hypothesis *(H')* that the person does not have the disease. Let D refer to the datum that is relevant to the hypotheses. In this example, D is a "positive" result and D' is a "negative" result from the blood test.

The problem stated that 1 in 1000 have the disease, so $P(H) = 0.001$; that is, the prior probability (before we test the blood) that a person has the disease is 0.001, so $P(H') = 1 - P(H) = 0.999$.

The conditional probability that a person will test "positive" given that the person has the disease is written as $P(\text{"positive"} \mid H) = 0.995$, and the conditional probability that a person will test "positive" given he or she is not sick is $P(\text{"positive"} \mid H') = 0.005$. These conditional probabilities are called the *hit rate* and the *false alarm rate* in signal detection, also known as *power* and *significance* (α). We need to calculate $P(H \mid D)$, the probability that a person is sick, given the test was "positive". This calculation is known as an *inference.*

The situation in the disease example above is as follows: we know $P(H)$, $P(D \mid H)$ *and* $P(D \mid H')$, and we want to calculate $P(H \mid D)$. The definition of conditional probability:

$$P(H \mid D) = \frac{P(H \cap D)}{P(D)} \tag{1}$$

we can also write, $P(H \cap D) = P(D \mid H)\, P(H)$. In addition, D can happen in two mutually exclusive ways, either with H or without it, so $P(D) = P(D \cap H) + P(D \cap H')$. Each of these conjunctions can be written in terms of conditionals, therefore:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D \mid H)\, P(H) + P(D \mid H')\, P(H')} \tag{2}$$

Equation 2 is Bayes' theorem. Substituting the values for the blood test problem yields the following result:

$$P(\textit{sick} \mid \text{"}\textit{positive}\text{"}) = \frac{(0.995)(0.001)}{(0.995)(0.001) + (0.005)(0.999)} = 0.166.$$

Does this result seem surprising? Think of it this way: Among 1000 people, only 1 is sick. If all 1000 were tested, the test will likely give a "positive" test to the sick person, but it would also give a "positive" to about 5 others (0.5% of 999 healthy people, about 5, should test positive). Thus, of the six who test "positive," only one is actually sick, so the probability of being sick, given a "positive" test, is only about one in six. Another way to look at the answer is that it is 166 times greater than the probability of being sick given no information (0.001), so there has indeed been considerable revision of opinion given the positive test.

An on-line calculator is available at the following URL:

http://psych.fullerton.edu/mbirnbaum/bayes/bayescalc.htm

The calculator allows one to calculate Bayesian inference in either probability or *odds*, which are a transformation of probability, $\Omega = p/(1 - p)$. For example, if probability = 1/4 (drawing a heart from a deck of cards), then the odds are 1/3 of drawing a heart. Expressed another way, the odds are 3 to 1 against drawing a heart.

In odds form, Bayes' theorem can be written:

$$\Omega_1 = \Omega_0 \left(\frac{P(D|H)}{P(D|H')} \right) \tag{3}$$

where Ω_1 and Ω_0 are the revised and prior odds, and the ratio of hit rate to false alarm rate, $\frac{P(D|H)}{P(D|H')}$ is also known as the likelihood ratio of the evidence. For example, in the disease problem, the odds of being sick are 999:1 against, or approximately 0.001. The ratio of hit rate to false alarm rate is 0.995/.005 = 199. Multiplying prior odds by this ratio gives revised odds of 0.199, about 5 to 1 against. Converting odds back to probability, $p = \Omega/(1 + \Omega) = 0.166$.

With a logarithmic transformation, Equation 3 becomes additive – prior probabilities and evidence should combine independently; that is, the effect of prior probabilities and evidence should contribute in the same way, at any level of the other factor.

Are humans Bayesian?

Psychologists have wondered if Bayes' theorem describes how people revise their beliefs (Birnbaum, 1983; Birnbaum & Mellers, 1983; Edwards, 1968; Fischhoff, Slovic, & Lichtenstein, 1979; Kahneman & Tversky, 1973; Koehler, 1996; Lyon & Slovic, 1976; Pitz, 1975; Shanteau, 1975; Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1982; Wallsten, 1972). The psychological literature can be divided into three periods. Early work supported Bayes' theorem as a rough descriptive model of how humans combine and update evidence, with the exception that people were described

as *conservative*, or less influenced by either base rate or evidence than Bayesian analysis of the objective evidence would warrant (Edwards, 1968; Wallsten, 1972).

The second period was dominated by Kahneman and Tversky's (1973) assertions that people do not use base rates or respond to differences in validity of sources of evidence. It emerged that their conclusions were viable only with certain types of experiments (e.g., Hammerton, 1973), but those experiments were easy to do, so many were done. Perhaps because Kahneman and Tversky (1973) did not cite the body of previous work that contradicted their conclusions, it took some time for those who followed in their footsteps to become aware of the contrary evidence and to rediscover how to replicate it (Novemsky & Kronzon, 1999).

More recent literature supports the early research showing that people do indeed utilize base rates and source credibility (Birnbaum, 2001; Birnbaum & Mellers, 1983; Novemsky & Kronzon, 1999). However, people appear to combine this information by an averaging model (Birnbaum, 1976, 2001; Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976; Troutman & Shanteau, 1977). The Scale-Adjustment Averaging Model of source credibility (Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979), is not consistent with Bayes' theorem and it also explains "conservatism".

Averaging model of source credibility

The averaging model of source credibility can be written as follows:

$$R = \frac{\sum_{i=0}^{n} w_i s_i}{\sum_{i=0}^{n} w_i} \tag{4}$$

where R is the predicted response, w_i the weights of the sources (which depend on the source's perceived credibility), and s_i is the scale value of the source's testimony (which depends on what the source testified). The initial impression reflects prior opinion (w_0 and s_0). For more on averaging models see Anderson (1981).

In problems such as the disease problem quoted earlier, there are three or more sources of information; first, there is the prior belief, represented by s_0; second, base rate is a source of information; third, the test result is another source of information. For example, suppose that weights of the initial impression and of the base rate are both 1, and the weight of the diagnostic test is 2. Suppose the prior belief is 0.50 (no opinion), scale value of the base rate is 0.001, and the scale value of the "positive" test is 1. This model predicts that the response in the disease problem is as follows:

$$\frac{1 \times 0.5 + 1 \times 0.001 + 2 \times 1}{1 + 1 + 2} = 0.63$$

Thus, this model can predict neglect of the base rate, if people put more weight on witnesses than on base rates.

Birnbaum and Stegner (1979) extended this model to describe how people combine information from sources varying in both validity and bias. Their model also involves configural weighting, in which the weight of a piece of information depends on its relation to other information. For example, when the judge is asked to identify with the buyer of a car, the judge appears to place more weight on lower estimates of the value of a car, whereas people asked to identify with the seller put more weight on higher estimates.

The most important distinction between Bayesian and averaging models is that in the Bayesian model, each piece of independent information has the same effect no matter what the current state of evidence. In the averaging models, however, the effect of any piece of information is inversely related to the number and total weight of other sources of information. In the averaging model, unlike the Bayesian model, the directional impact of information depends on the relation between the new evidence and the current opinion.

Although the full story is beyond the scope of this chapter, three aspects of the literature can be illustrated by data from a single experiment, which can be done two ways – as a within-subjects or a between-subjects study. The next section describes a between-subjects experiment, like the one in Kahneman and Tversky (1973); the section following it will describe how to conduct and analyze a within-subjects design, like that of Birnbaum and Mellers (1983).

EXPERIMENTS: THE CAB PROBLEM

Consider the following question, known as the *cab problem* (Tversky & Kahneman, 1982, pp. 156–157):

> A cab was involved in a hit and run accident at night. There are two cab companies in the city, with 85% of cabs being Green and the other 15% Blue cabs. A witness testified that the cab in the accident was "Blue." The witness was tested for ability to discriminate Green from Blue cabs and was found to be correct 80% of the time. What is the probability that the cab in the accident was Blue as the witness testified?

Between-subjects vs within-subjects designs

If we present a single problem like this to a group of students, the results show a strange distribution of responses. The majority of students (about

three out of five) say that the answer is "80%", apparently because the witness was correct 80% of the time. However, there are two other modes: about one in five responds "15%", the base rate; a small group of students give the answer of 12%, apparently the result of multiplying the base rate by the witness's accuracy, and a few people give a scattering of other answers. Supposedly, the "right" answer is 41%, and few people give this answer.

Kahneman and Tversky (1973) argued that people ignore base rate, based on finding that the effect of base rate in such inference problems was not significant. They asked participants to infer whether a person was a lawyer or engineer, based on a description of personality given by a witness. The supposed neglect of base rate found in this *lawyer–engineer* problem and others came to be called the "base-rate fallacy" (see also Hammerton, 1973). However, evidence of a fallacy evaporates when one does the experiment in a slightly different way using a within-subjects design, as we see below (Birnbaum, 2001; Birnbaum & Mellers, 1983; Novemsky & Kronzon, 1999).

There is also another issue with the cab problem and the lawyer–engineer problem as they were formulated. Those problems were not stated clearly enough that one can apply Bayes' theorem without making extra assumptions (Birnbaum, 1983; Schum, 1981). One has to make arbitrary, unrealistic assumptions in order to calculate the supposedly "correct" solution.

Tversky and Kahneman (1982) gave the "correct" answer to this cab problem as 41% and argued that participants who responded "80%" were mistaken. They assumed that the percentage correct of a witness divided by percentage wrong equals the ratio of the hit rate to the false alarm rate. They then took the percentage of cabs in the city as the prior probability for cabs of each colour being in cab accidents at night. It is not clear, however, that both cab companies even operate at night, so it is not clear that percentage of cabs in a city is really an appropriate prior for being in an accident.

Furthermore, we know from signal-detection theory that the percentage correct is not usually equal to hit rate, nor is the ratio of hit rate to false alarm rate for human witnesses invariant when base rate varies. Birnbaum (1983) showed that if one makes reasonable assumptions about the witness in these problems, then the supposedly "wrong" answer of 80% is actually a better solution than the one called "correct" by Tversky and Kahneman.

The problem is to infer how the ratio of hit rate to false alarm rate (in Eq. 3) from the values given for the witness is affected by the base rate. Tversky and Kahneman (1982) implicitly assumed that this ratio is unaffected by base rate. However, experiments in signal detection show that this ratio changes in response to changing base rates. Therefore this complication must be taken into account when computing the solution (Birnbaum, 1983).

Birnbaum's (1983) solution treats the process of signal detection with reference to normal distributions on a subjective continuum, one for the signal and another for the noise. If the observer changes his or her "Green/

Blue" response criterion to maximize percent correct, then the solution of 0.80 is not far from what one would expect if the witness was an ideal observer (for details, see Birnbaum, 1983).

Fragile results in between-subjects research

But perhaps even more troubling to behavioural scientists was the fact that the null results deemed evidence of a "base-rate fallacy" proved very fragile to replication with different procedures (see Gigerenzer & Hoffrage, 1995, and Chapter 3). In a within-subjects design, it is easy to show that people attend to both base rates and source credibility.

Birnbaum and Mellers (1983) reported that within-subjects and between-subjects studies give very different results (see also Fischhoff et al., 1979). Whereas the observed effect of base rate may not be significant in a between-subjects design, the effect is substantial in a within-subjects design. Whereas the distribution of responses in the between-subjects design has three modes (e.g., 80%, 15%, and 12% in the above cab problem), the distribution of responses in within-subjects designs is closer to a bell shape. When the same problem is embedded among others with varied base rates and witness characteristics, Birnbaum and Mellers (1983, Fig. 2) found few responses at the former peaks; the distributions instead appeared bell-shaped.

Birnbaum (1999a) showed that in a between-subjects design, the number 9 is judged to be significantly "bigger" than the number 221. Should we infer from this that there is a "cognitive illusion" a "number fallacy", a "number heuristic", or a "number bias" that makes 9 seem bigger than 221?

Birnbaum (1982, 1999a) argued that many confusing results will be obtained by scientists who try to compare judgements between groups who experience different contexts. When they are asked to judge both numbers, people say 221 is greater than 9. It is only in the between-subjects study that significant and opposite results are obtained. One should not compare judgements between groups without taking the context into account (Birnbaum, 1982).

In the complete between-subjects design, context is completely confounded with the stimulus. Presumably, people asked to judge (only) the number 9 think of a context of small numbers, among which 9 seems "medium", and people judging (only) the number 221 think of a context of larger numbers, among which 221 seems "small".

DEMONSTRATION EXPERIMENT

To illustrate findings within-subjects, a factorial experiment on the cab problem will be presented. This study is similar to one by Birnbaum (2001). It varies the base rate of accidents in which Blue cabs were involved (15%, 30%, 70%, or 85%) and the credibility of a witness (medium or high). The

participants' task is to estimate the probability that the car in the accident was a Blue cab. All methodological details are given in Text box 2.1.

Text box 2.1 Method of demonstration experiment

Instructions make base rate relevant and give more precise information on the witnesses. Instructions for this version are as follows:

> A cab was involved in a hit-and-run accident at night. There are two cab companies in the city, the Blue and Green. Your task is to judge (or estimate) the probability that the cab in the accident was a Blue cab.
>
> You will be given information about the percentage of accidents at night that were caused by Blue cabs, and the testimony of a witness who saw the accident. The percentage of night-time cab accidents involving Blue cabs is based on the previous 2 years in the city. In different cities, this percentage was either 15%, 30%, 70%, or 85%. The rest of night-time accidents involved Green cabs. Witnesses were tested for their ability to identify colours at night. They were tested in each city at night, with different numbers of colours matching their proportions in the cities.
>
> The MEDIUM witness correctly identified 60% of the cabs of each colour, calling Green cabs "Blue" 40% of the time and calling Blue cabs "Green" 40% of the time.
>
> The HIGH witness correctly identified 80% of each colour, calling Blue cabs "Green" or Green cabs "Blue" on 20% of the tests.
>
> Both witnesses were found to give the same ratio of correct to false identifications on each colour when tested in each of the cities.

Each participant received 20 situations, in random order, after a warmup of 7 trials. Each situation was composed of a base rate, plus testimony of a high-credibility witness who said the cab was either "Blue" or "Green", testimony of a medium-credibility witness (either "Blue" or "Green"), or there was no witness. A typical trial appeared as follows:

> 85% of accidents are Blue cabs & medium witness says "Green".

The dependent variable was the judged probability that the cab in the accident was Blue, expressed as a percentage. The 20 experimental trials were composed of the union of a 2 × 2 × 4, Source Credibility (Medium, High) by Source Message ("Green", "Blue") by Base Rate (15%, 30%, 70%, 85%) design, plus a one-way design with four levels of Base Rate and no witness.

Complete materials can be viewed at the following URL:

> http://psych.fullerton.edu/mbirnbaum/bayes/CabProblem.htm

The following results are based on data from 103 undergraduates who were recruited from the university "subject pool" and who participated via the worldwide web.

Results and discussion

Mean judgements of probability that the cab in the accident was Blue are presented in Table 2.1. Rows show effects of Base Rate, and columns show combinations of witnesses and their testimony. The first column shows that if Blue cabs were involved in only 15% of cab accidents at night and the high-credibility witness said the cab was "Green", the average response was only 29.1%. When Blue cabs were involved in 85% of accidents, however, the mean judgement was 49.9%. The last column of Table 2.1 shows that when the high-credibility witness said that the cab was "Blue", mean judgements were 55.3% and 80.2% when base rates were 15% and 85%, respectively.

Analysis of variance tests the null hypotheses that people ignored base rate or witness credibility. The ANOVA showed that the main effect of Base Rate was significant, $F(3, 306) = 106.2$, as was Testimony, $F(1, 102) = 158.9$. Credibility of the witness has both significant main effects and interactions with Testimony, $F(1, 102) = 25.5$, and $F(1, 102) = 58.6$, respectively. As shown in Table 2.1, the more diagnostic the witness, the greater the effect of that witness's testimony. These results show that we can reject the hypotheses that people ignored base rates and validity of evidence.

The critical value of $F(1, 60)$ is 4.0, with $\alpha = 0.05$, and the critical value of $F(1, 14)$ is 4.6. Therefore, the observed F-values are more than 10 times their critical values. Because F values are approximately proportional to n for true effects, one should be able to reject the null hypotheses of Kahneman and Tversky (1973) with only 15 participants. However, the purpose of this research is to evaluate models of how people combine evidence, which requires larger samples in order to provide clean results. Experiments conducted via the worldwide web allow one to test large numbers of participants quickly at relatively low cost in time and effort (see Birnbaum, 2001). Therefore, it is best to collect more data than are necessary just to show statistical significance.

Table 2.2 shows Bayesian calculations, simply using Bayes' theorem to calculate with the numbers given. (Probabilities are converted to

Table 2.1 Mean judgements of probability that the cab was Blue (%)

	Witness credibility and witness testimony				
Base rate	*High credibility "Green"*	*Medium credibility "Green"*	*No witness*	*Medium credibility "Blue"*	*High credibility "Blue"*
15	29.1	31.3	25.1	41.1	55.3
30	34.1	37.1	36.3	47.4	56.3
70	46.0	50.3	58.5	60.9	73.2
85	49.9	53.8	67.0	71.0	80.2

Each entry is the mean inference judgement, expressed as a percentage.

Table 2.2 Bayesian predictions (converted to percentages)

	Witness credibility and witness testimony				
Base rate	*High credibility "Green"*	*Medium credibility "Green"*	*No witness*	*Medium credibility "Blue"*	*High credibility "Blue"*
15	4.2	10.5	15.0	20.9	41.4
30	9.7	22.2	30.0	39.1	63.2
70	36.8	60.9	70.0	77.8	90.3
85	58.6	79.1	85.0	89.5	95.8

percentages.) Figure 2.1 shows a scatterplot of mean judgements against Bayesian calculations. The correlation between Bayes' theorem and the data is 0.948, which might seem "high". It is this way of graphing the data that led to the conclusion of "conservatism", as described in Edwards' (1968) review.

Conservatism described the fact that human judgements are less extreme than Bayes' theorem dictates. For example, when 85% of accidents at night involved Blue cabs and the high-credibility witness said the cab was "Blue", Bayes' theorem gives a probability of 95.8% that the cab was Blue; in contrast, the mean judgement was only 80.2%. Similarly, when base rate was

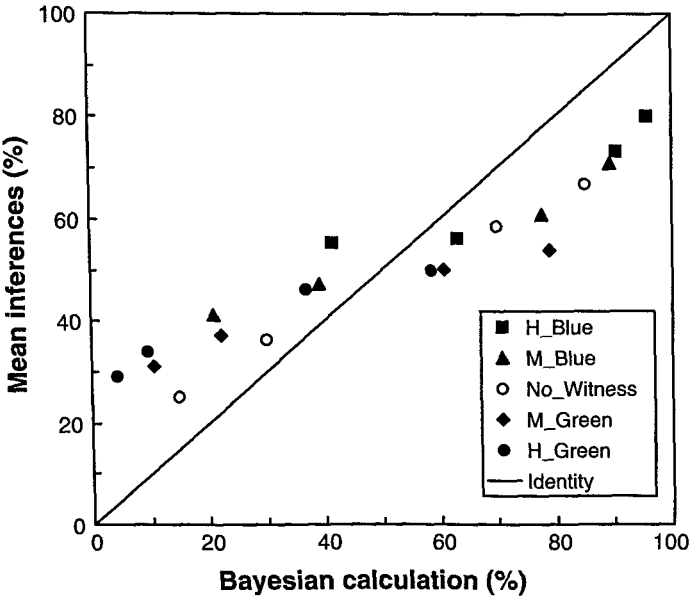

Figure 2.1 Mean inference that the cab was Blue, expressed as a percentage, plotted against the Bayesian solutions, also expressed as percentages (H = high-, M = medium-credibility witness).

15% and the high-credibility witness said the cab was "Green", Bayes' theorem calculates 4.2% and the mean judgement was 29.1%.

A problem with this way of graphing the data is that it does not reveal patterns of systematic deviation, apart from regression. People looking at such scatterplots are often impressed by "high" correlations. Such correlations of fit with such graphs easily lead researchers to wrong conclusions (Birnbaum, 1973). The problem is that "high" correlations can coexist with systematic violations of a theory. Correlations can even be higher for worse models! See Birnbaum (1973) for examples showing how misleading correlations of fit can be.

In order to see the data better, they should graphed as in Figure 2.2, where they are drawn as a function of base rate, with a separate curve for each type of witness and testimony. Notice the unfilled circles, which show judgements for cases with no witness. The cross-over between this curve and others contradicts the additive model, including Wallsten's (1972) subjective Bayesian (additive) model and the additive model rediscovered by Novemsky and Kronzon (1999). The subjective Bayesian model utilizes Bayesian formulas but allows the subjective values of probabilities to differ

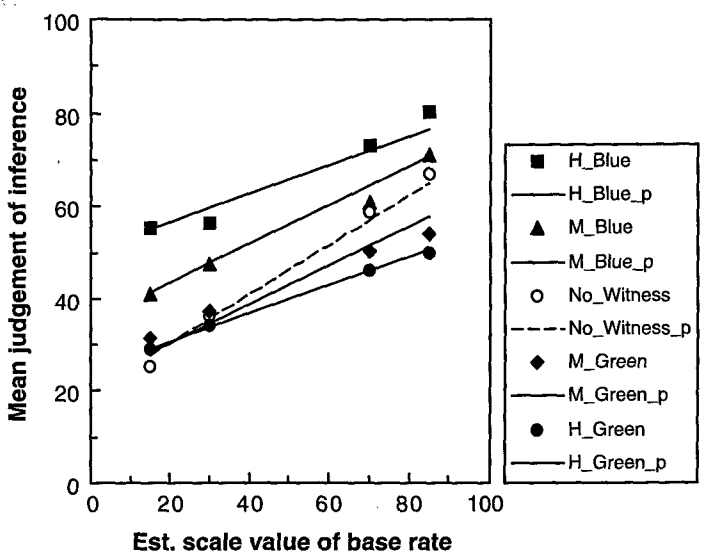

Figure 2.2 Fit of averaging model: Mean judgements of probability that the cab was Blue, plotted as a function of the estimated scale value of the base rate. Filled squares, triangles, diamonds, and circles show results when a high-credibility witness said the cab was "Green", a medium-credibility witness said "Green", a medium-credibility witness said "Blue", or a high-credibility witness said "Blue", respectively. Solid lines show corresponding predictions of the averaging model. Open circles show mean judgements when there was no witness, and the dashed line shows corresponding predictions (H = high-, M = Medium-credibility witness, p = predicted).

from objective values stated in the problem. Instead, the crossover interaction indicates that people are averaging information from base rate with the witness's testimony. When subjects judge the probability that the car was Blue given only a base rate of 15%, the mean judgement was 25.2%. However, when a medium-credibility witness also said that the cab was "Green", which should exonerate the Blue cab and thus *lower* the inference that the cab was Blue, the mean judgement actually *increased* from 25.1% to 31.3%.

Troutman and Shanteau (1977) reported analogous results. They presented non-diagnostic evidence (which should have no effect) that caused people to become less certain. Birnbaum and Mellers (1983) showed that when people have a high opinion of a car, and a low credibility source says the car is "good", it actually makes people think the car is worse. Birnbaum and Mellers (1983) also reported that the effect of base rate is reduced when the source is higher in credibility. These findings are consistent with averaging rather than additive models.

Model fitting

In the old days, one wrote special computer programs to fit models to data (Birnbaum, 1976; Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979). However, spreadsheet programs such as *Excel* can now be used to fit such models without requiring programming. Methods for fitting models via the Solver in *Excel* are described in detail for this type of study in Birnbaum (2001, Ch. 19).

Each model has been fitted to the data in Table 2.1, by minimizing the sum of squared deviations. Lines in Figure 2.2 show predictions of the averaging model. Estimated parameters are as follows: weight of the initial impression, w_0, was fixed to 1; estimated weights of the base rate, medium-credibility witness, and high-credibility witness were 1.11, 0.58, and 1.56 respectively. The weight of base rate was intermediate between the two witnesses, although it "should" have exceeded the high-credibility witness.

Estimated scale values of base rates of 15%, 30%, 70%, and 85% were 12.1, 28.0, 67.3, and 83.9 respectively, close to the objective values. Estimated scale values for testimony ("Green" or "Blue") were 31.1 and 92.1 respectively. The estimated scale value of the initial impression was 44.5. This 10-parameter model correlated 0.99 with mean judgements. When the scale values of base rate were fixed to their objective values (reducing the model to only six free parameters), the correlation was still 0.99.

The sum of squared deviations (SSD) provides a more useful index of fit in this case. For the null model, which assumes no effect of base rate or source validity, SSD = 3027, which fits better than objective Bayes' theorem (plugging in the given values), with SSD = 5259. However, for the subjective Bayesian (additive) model, SSD = 188, and for the averaging model, SSD = 84. For the simpler averaging model (with subjective base rates set to their

objective values), SSD = 85. In summary, the assumption that people attend only to the witness's testimony does fit better than the objective version of Bayes' theorem; however, its fit is much worse than the subjective (additive) version of Bayes' theorem. The averaging model, however, provides the best fit, even when simplified by the assumption that people take the base-rate information at face (objective) value.

OVERVIEW AND CONCLUSIONS

The case of the "base-rate fallacy" illustrates a type of cognitive illusion to which scientists are susceptible when they find non-significant results. The temptation is to say that because I have found no significant effects (of different base rates or source credibilities), there are therefore no effects. However, when results fail to disprove the null hypothesis, they do not prove the null hypothesis. This problem is particularly serious in between-subjects research, where it is easy to get non-significant results, or significant but silly results such as "9 seems bigger than 221".

The conclusions by Kahneman and Tversky (1973) that people neglect base rate and credibility of evidence are quite fragile. One must use a between-subjects design and use only certain wordings. Because I can show that the number 9 seems "bigger" than 221 with this type of design, I put little weight on such fragile between-subjects findings. In within-subjects designs, even the lawyer–engineer task shows effects of base rate (Novemsky & Kronzon, 1999). Although Novemsky and Kronzon argued for an additive model, they did not include the comparisons needed to test the additive model against the averaging model of Birnbaum and Mellers (1983). I believe that had these authors included appropriate designs, they would have been able to reject the additive model. They could have presented additional cases in which there were witness descriptions but no base-rate information, base-rate information but no witnesses (as in the dashed curve of Figure 2.2), different numbers of witnesses, or witnesses with varying amounts of information or different levels of expertise in describing people. Any of these manipulations would have provided of tests between the additive and averaging models.

In any of these manipulations, the implication of the averaging model is that the effect of any source (e.g., the base rate) would be inversely related to the total weight of other sources of information. This type of analysis has consistently favoured averaging over additive models in source credibility studies (e.g., Birnbaum, 1976, Fig. 3; Birnbaum & Mellers, 1983, Fig. 4C; Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976, Figs. 2B & 3).

Edwards (1968) noted that human inferences might differ from Bayesian inferences for any of three basic reasons – misperception, misaggregation, or response distortion. People might not absorb or utilize all of the evidence, people might combine the evidence inappropriately, or they might express

their subjective probabilities using a response scale that needs transformation. Wallsten's (1972) model was an additive model that allowed misperception and response distortion, but which retained the additive Bayesian aggregation rule (recall that the Bayesian model is additive under monotonic transformation). This additive model is the subjective Bayesian model that appears to give a fairly good fit in Figure 2.1.

When proper analyses are conducted, however, it appears that the aggregation rule violates the additive structure of Bayes' theorem. Instead, the effect of a piece of evidence is not independent of other information available, but instead is diminished by total weight of other information. This is illustrated by the dashed curve in Figure 2.2, which crosses the other curves.

Birnbaum and Stegner (1979) decomposed source credibility into two components, expertise and bias, and distinguished these from the judge's bias, or point of view. Expertise of a source of evidence affects its weight, and is affected by the source's ability to know the truth, reliability of the source, cue correlation, or the source's signal-detection d'. In the case of gambles, weight of a branch is affected by the probability of a consequence. In the experiment described here, witnesses differed in their abilities to distinguish Green from Blue cabs.

In the averaging model, scale values are determined by what the witness says. If the witness said it was a "Green" cab, it tends to exonerate the Blue cab driver, whereas if the witness said the cab was "Blue", it tends to implicate the Blue cab driver. Scale values of base rates were nearly equal to their objective values. In judgements of the value of cars, scale values are determined by estimates provided by sources who drove the car and by the "blue book" values. (The blue book lists the average sale price of a car of a given make, model, and mileage, so it is like a base rate and does not reflect any expert examination or test drive of an individual vehicle.)

Bias reflects a source's tendency to over- as opposed to under-estimate judged value, presumably because sources are differentially rewarded or punished for giving values that are too high or too low. In a court trial, bias would be affected by affiliation with defence or prosecution. In an economic transaction, bias would be affected by association with buyer or seller. Birnbaum and Stegner (1979) showed that source's bias affected the scale value of that source's testimony.

In Birnbaum and Meller's (1983) study, bias was manipulated by changing the probability that the source would call a car "good" or "bad" independent of the source's diagnostic ability. Whereas expertise was manipulated by varying the difference between hit rate and false alarm rate, bias was manipulated by varying the sum of hit rate plus false alarm rate. Their data were also consistent with the scale-adjustment model that bias affects scale value.

The judge, who combines information, may also have a type of bias, known as the judge's *point of view*. The judge might be combining information to determine buying price, selling price, or "fair price". An example of a

"fair" price is when one person damages another's property and a judge is asked to give a judgement of the value of damages so that her judgement is equally fair to both people. Birnbaum and Stegner (1979) showed that the source's viewpoint affects the configural weight of higher- or lower-valued branches. Buyers put more weight on the lower estimates of value and sellers place higher weight on the higher-valued estimates. This model has also proved quite successful in predicting judgements and choices between gambles (Birnbaum, 1999b).

Birnbaum and Mellers (1983, Table 2) drew a table of analogies that can be expanded to show that the same model appears to apply not only to Bayesian inference, but also to numerical prediction, contingent valuation, and a variety of other tasks. To expand the table to include judgements of the values of gambles and decisions between them, let viewpoint depend on the task to judge buying price, selling price, "fair" price, or to choose between gambles. Each discrete probability (event) consequence branch has a weight that depends on probability (or event). The scale value depends on the consequence. Configural weighting of higher- or lower-valued branches depends on identification with the buyer, seller, independent, or decider.

Much research has been developing a catalogue of cognitive illusions, each to be explained by a "heuristic" or "bias" of human thinking. Each time a "bias" is named, one has the cognitive illusion that it has been explained. The notion of a "bias" suggests that if the bias could be avoided, people would suffer no illusions. A better approach to the study of cognitive illusions would be one more directly analogous to the study of visual illusions in perception. Visual illusions can be seen as consequences of a mechanism that allows people to judge actual sizes of objects with different retinal sizes at different distances. A robot that judged size by retinal size only would not be susceptible to the Mueller-Lyer illusion. However, it would also not satisfy size constancy. As an object moved away, it would seem to shrink. So, rather than blame a "bias" of human reasoning, we should seek the algebraic models of judgement that allow one to explain both illusion and constancy with the same model.

SUMMARY

- Early research that compared intuitive judgements of probability and Bayesian calculations concluded that people were "conservative", in that their judgements were closer to uncertainty than dictated by the formula.
- Based on poor studies, it was later argued that people neglect or do not attend to base rates or source validity when making Bayesian inferences.
- Evidence for the so-called "base-rate fallacy" and source neglect is very fragile and does not replicate except in very restricted conditions. When base rates, source, expertise, and testimony are manipulated

within-subjects, judges do utilize the base rates and attend to source expertise.

- The subjective Bayesian model provides a better fit than the objective model, because it can account for "conservatism" and the nearly additive relationship between base rate and source's opinion.
- However, the data show two phenomena that rule out additive or subjective Bayesian formulations: The effect of the base rate is inversely related to the number and credibility of other sources.
- The data are better described by Birnbaum and Stegner's (1979) scale-adjustment averaging model than by the other models.

FURTHER READING

Reviews of this literature from different viewpoints are presented by Edwards (1968), Tversky and Kahneman (1982), Koehler (1996), and in Chapter 3 of this volume. Birnbaum (1983) showed that the so-called "normative" Bayesian analysis presented by Tversky and Kahneman (1982) made an implausible assumption that made their conclusions unwarranted. Birnbaum and Mellers (1983) showed how to apply the model of Birnbaum and Stegner (1979) to the Bayesian inference task. The model fit here is a special case of that model, which also describes effects of the bias of sources and the viewpoint of the judge.

ACKNOWLEDGEMENT

Support was received from National Science Foundation Grants, SES 99-86436, and BCS-0129453.

REFERENCES

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Birnbaum, M. H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 79, 239–242.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology*, 89, 417–429.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, 96, 85–94.

Birnbaum, M. H. (1999a). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, 4, 243–249.

Birnbaum, M. H. (1999b). Testing critical properties of decision making on the Internet. *Psychological Science, 10*, 399–407.

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet.* Upper Saddle River, NJ: Prentice Hall.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792–804.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37*, 48–74.

Birnbaum, M. H., Wong, R., & Wong, L. (1976). Combining information from sources that vary in credibility. *Memory & Cognition, 4*, 330–336.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23*, 339–359.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency format. *Psychological Review, 102*, 684–704.

Hammerton, M. A. (1973). A case of radical probability estimation. *Journal of Experimental Psychology, 101*, 252–254.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Koehler, J. J. (1996). The base-rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53.

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40*, 287–298.

Novemsky, N., & Kronzon, S. (1999). How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. *Journal of Behavioral Decision Making, 12*, 55–69.

Pitz, G. (1975). Bayes' theorem: Can a theory of judgment and inference do without it? In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1, pp. 131–148). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance, 27*, 153–196.

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

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744.

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance, 19*, 43–55.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). New York: Cambridge University Press.

Wallsten, T. (1972). Conjoint-measurement framework for the study of probabilistic information processing. *Psychological Review, 79*, 245–260.

APPENDIX

The complete materials for this experiment, including HTML that collects the data are available via the WWW from the following URL:

http://psych.fullerton.edu/mbirnbaum/bayes/resources.htm

A sample listing of the trials, including warmup, is given below.

Warmup trials: Judge the probability that the cab was Blue.

Express your probability judgement as a percentage and type a number from 0 to 100.

W1 15% of accidents are Blue Cabs & high witness says "Green".

(There were six additional "warmup" trials that were representative of the experimental trials.)

Please re-read the instructions, check your warmups, and then proceed to the trials below.

Test trials: What is the probability that the cab was Blue?

Express your probability judgement as a percentage and type a number from 0 to 100.

1 85% of accidents are Blue Cabs & medium witness says "Green".
2 15% of accidents are Blue Cabs & medium witness says "Blue".
3 15% of accidents are Blue Cabs & medium witness says "Green".
4 15% of accidents are Blue Cabs & there was no witness.
5 30% of accidents are Blue Cabs & high witness says "Blue".
6 15% of accidents are Blue Cabs & high witness says "Green".
7 70% of accidents are Blue Cabs & there was no witness.
8 15% of accidents are Blue Cabs & high witness says "Blue".
9 70% of accidents are Blue Cabs & high witness says "Blue".
10 85% of accidents are Blue Cabs & high witness says "Green".
11 70% of accidents are Blue Cabs & high witness says "Green".
12 85% of accidents are Blue Cabs & medium witness says "Blue".
13 30% of accidents are Blue Cabs & medium witness says "Blue".
14 30% of accidents are Blue Cabs & high witness says "Green".
15 70% of accidents are Blue Cabs & medium witness says "Blue".
16 30% of accidents are Blue Cabs & there was no witness.
17 30% of accidents are Blue Cabs & medium witness says "Green".
18 70% of accidents are Blue Cabs & medium witness says "Green".
19 85% of accidents are Blue Cabs & high witness says "Blue".
20 85% of accidents are Blue Cabs & there was no witness.