The Use of Base Rates in Bayesian Inference Prediction

Brown J lee
Washington State University, Fullerton and Decision Research Center
Email: mbirnbaum@fullerton.edu

Introduction

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

These two questions are quite different. In the first case, we can develop a mathematical theory based on the assumption that each card is equally likely to be drawn. 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 can 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 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 happen, and there is no way to calculate a proportion from the past that is clearly relevant. One might look at the records of the German team and those of rivals, and we can ask if the German team seems healthy and ready to play, but players change, the 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 sometimes used to refer to the psychological strength of belief that the event will happen.

Despite these distinctions, people are willing to use the same term, probability, to express both types of ideas. People are willing to gamble on both types of predictions—on repeatable, mechanical games of chance (like dice, cards, and roulette) and on unique events (like horse races and other sporting contests). In fact, people are even willing to use the same language *after* something has happened (a murder, for example), to discuss the "probability" that a particular event occurred (e.g., this defendant committed the crime).

The Rev. Thomas Bayes (1702-1761) derived a theorem for inference from the mathematics of probability. Philosophers recognized that this theorem could be interpreted as a calculus for rational thought.

Bayes' Theorem

To illustrate Bayes' theorem, it helps to work through an example. 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 people who do not have the disease. If a person tested "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 $P(H) = .001$; i.e., the prior probability (before we test the blood) that a person has the disease is .001, so $P(H') = 1 - P(H) = 0.999$.

The conditional probability that a person will test "positive" given that person has the disease is written as P("positive"| H) = .995, and the conditional probability that a person will test "positive" given he or she is not sick is P("positive"| H') = .005. These probabilities are often called *the HIT RATE* and *FALSE ALARM RATE* in signal detection, and they are also known as *power* and *significance* (α) in statistics. We

 ISSN 2047-3745

need to calculate *P(H| D)*, the probability that a person is sick, given the test was "positive." This calculation is known as an *inference*.

The conditional probability of *A* given *B* is not the same as the conditional probability of *B* given *A*. For example, the probability that someone is male given he or she is a member of the U.S. senate is quite high because there are few women in the senate. However, the probability that a person is a member of the U.S. senate given that person is male is quite low, since there are so few senators and so many males. Conditional probability is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad (2)$$

So, if *A* is the set of U.S. Senators (of which there are 100), and *B* is the set of males (of which there are billions), we see that the probability of *A* given *B* will be quite small, but the probability of *B* given *A* can be quite high.

The situation is as follows: we know *P(H), P(D|H) and P(D|H')*, and we want to calculate P(H|D). From the definition of conditional probability:

$$P(H \mid D) = \frac{P(H \cap D)}{P(D)} \quad (3)$$

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, so by substitution the formula is as follows:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D \mid H)P(H) + P(D \mid H')P(H')} \quad (4)$$

Equation 4 is known as Bayes' theorem. Substituting the values for the blood test problem yields the following result:

$$P(sick \mid \text{" } positive\text{"}) = \frac{(.995)(.001)}{(.995)(.001) + (.005)(.999)} = .166.$$

Does this result seem surprising? Think of it this way: Among 1000 people, there is only one sick person. If all 1000 were given the test, the test would probably give a "positive" test result to that one person, and it would also give a "positive" test result to about five others (of the 999 healthy people 0.5% should test positive). Thus, of the six people who test "positive," only one is really sick, so the probability of being sick, given a "positive" test result, is only about one in six. Another way to look at the .166 is that it is 166 times bigger than the probability of being sick given no information about the person (.001), so there has indeed been considerable revision of opinion given the positive test.

To facilitate working with Bayes' theorem, I have written an on-line calculator, which is available at the following URL:

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

The calculator allows one to work with Bayes' theorem in either probability or *odds* form. One can express probabilities in the form of odds, $\Omega = p/(1 - p)$. For example, if the probability of drawing a Heart from a deck of cards is 1/4, then the odds are 1/3 of drawing a Heart; i.e., we expect one Heart for every three non-Hearts. Expressed another way, the odds are 3 to 1 against drawing a Heart. In odds form, Bayes' theorem can be written:

$$\Omega_1 = \Omega_0 \left(\frac{P(D \mid H)}{P(D \mid H')}\right) \quad (5)$$

where Ω_1 and Ω_0 are the revised and prior odds, and the ratio of hit rate to false alarm rate is also known as the likelihood ratio of the evidence. In this form, the revision of opinion based on new information is

 ISSN 2047-3745

multiplicative. With a logarithmic transformation, Equation 5 becomes additive—the effect of prior probabilities and evidence should combine by a process that satisfies independence; the effect of prior probabilities and evidence should contribute in the same way, independent of the level of the other factor.

Are Humans Bayesian?

Psychologists have wondered if Bayes' theorem describes how people form and revise their beliefs (Birnbaum, 1983; Birnbaum & Mellers, 1983; Edwards, 1968; Fischhoff, Slovic, & Lichtenstein, 1979; Gigerenzer & Hoffrage, 1995; Hammerton, 1973; Kahneman & Tversky, 1973; Koehler, 1996; Lyon & Slovic, 1976; Novemsky & Kronzon, 1999; Pitz, 1975; Shanteau, 1975; Slovic & Lichtenstein, 1971; Troutman & Shanteau, 1977; 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 base rate and evidence than Bayesian analysis of the objective evidence would warrant (Edwards, 1968; Wallsten, 1972).

The early research was followed by a period 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 turns out 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 rediscover how to replicate it (Novemsky & Kronzon, 1999).

More recent literature confirms the earliest results and shows 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 a model that is not additive (Birnbaum, 1976; 2001; Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976; Troutman & Shanteau, 1977). This model is not consistent with Bayes theorem and it also explains behavior that has been described as conservative.

Averaging Model of Source Credibility

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

$$R = \frac{\sum_{i=0}^{n} w_i s_i}{\sum_{i=0}^{n} w_i} \qquad (6)$$

where R is the predicted response, w_i the weights of the sources (which depend on the sources' 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 the person's prior opinion and is represented by w_0 and s_0. Birnbaum and Stegner (1979) presented an extension of this averaging model to describe how people combine information from sources that vary in both perceived validity and bias. Their model also involves configural weighting that depends on the point of view of the judge.

The most important distinction between Bayesian and the family of 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, 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, the directional effect of information depends on the relation between the new evidence and the current opinion.

Although the full scope of these models 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-

 ISSN 2047-3745

subjects experiment in which each participant judges combinations of evidence and base rates, or as a between-subjects experiment in which the participant only receives one condition.

Experiments

Consider the following question, known as the *Cab Problem* (Tversky & Kahneman, 1982):

> 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 Results: Not Significant

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 60%) say that the answer is 80%, apparently because the witness was correct 80% of the time. However, there are two other modes: about twenty percent respond 15%, the base rate; and another small group of students (about 5%) give the answer of 12%, apparently the result of multiplying the base rate by the witness's accuracy; a few people give a variety of other answers.

Kahneman and Tversky (1973) argued that people do not attend to base rates at all, based on finding that the effect of base rate in similar 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 similar problems 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 (Birnbaum, 2001; Birnbaum & Mellers, 1983; Novemsky & Kronzon, 1999).

There is also another issue with the cab problem and lawyer-engineer problem in their early forms. Unfortunately, those problems were not defined clearly enough that one can apply Bayes' theorem (Birnbaum, 1983; Schum, 1981). One has to make arbitrary assumptions that are not realistic in order to plug numbers into the equation.

Tversky and Kahneman (1982) argued that the correct answer to the above cab problem is 0.41, and they maintained that participants who responded .80 were in error. However, Birnbaum (1983) showed that if one makes reasonable assumptions about the behavior of the witness in the cab problem or in the lawyer-engineer problem, then the supposedly "wrong" answer of .8 is actually a better computation than the solutions given as "correct" by Kahneman and Tversky (1973; Tversky & Kahneman, 1982).

The problem is to infer how the ratio of hit rate to false alarm rate (Equation 5) for the witness is affected by the base rate. Tversky and Kahneman (1982) implicitly assumed that this ratio is unaffected by base rate when they computed their solutions. However, evidence from experiments in signal detection shows that witnesses change their ratios in response to changing base rates. Therefore the use of human witnesses in the lawyer-engineer and cab problems introduces a complication that must be taken into account when computing the solution from the theorem. This complication applies even when the witness is a machine (e.g., a light bulb tester), if humans are allowed to adjust the machine for optimal performance (Birnbaum, 1983).

But perhaps even more troubling to behavioral scientists was the fact that the results deemed evidence of a "base rate fallacy" proved very fragile to replication with different procedures. 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. Whereas the 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-

 ISSN 2047-3745

subjects design has three modes (at the witness hit rate, base rate, and product of these), the distribution of responses to the same problem embedded in a within-subjects design gives a more bell-shaped distribution. When the same case is embedded in a within-Ss design, Birnbaum and Mellers (1983, Figure 2) found few responses at the peaks found in the between-Ss version.

Indeed, 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 judgments between groups who experience different contexts. When people are asked to judge both numbers, no one says 9 is greater than 221. It is only in the between-subjects study that such results are obtained. The problem is that one cannot compare judgments between groups without taking the context into account (Birnbaum, 1982).

It is easy to get non-significant results or odd results between groups who experience different contexts. In the complete between-Ss design, the context is completely confounded with the stimulus. Presumably, people asked to judge the number 9 think of a context of small numbers among which 9 seems "medium" and people who judge only the number 221 think of a context of larger numbers, among which 221 seems "small."

Within-Subjects Method

To illustrate what is found within-subjects, a factorial experiment on the Cab problem will be presented. The instructions here make the base rate relevant and give more precise information on characteristics of the witnesses. This study is similar to one reported by Birnbaum (2001). 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 colors at night. They were tested in each city at night, with different numbers of colors
> matching their proportions in the cities.
> "The MEDIUM witness correctly identified 60% of the cabs of each color, 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 color, 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 when tested in each of the cities."

Each participant received 20 situations, in random order, after a warmup of 7 trials. Each situation is composed of a base rate with testimony of a high credibility witness who said the cab was either "Blue" or "Green", or testimony of a medium credibility witness who said it was 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."

 ISSN 2047-3745

The dependent variable was always to judge the probability that the cab in the accident was Blue. The 20 experimental trials were composed of the union of a 2 by 2 by 4, Source Credibility by Source Message by Base Rate design combined with 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
Data for this chapter come from 103 undergraduates who were recruited from the university "subject pool" who participated via the WWW.

Results

Mean judgments of probability that the cab in the accident was Blue are presented in Table 1. Rows show the effect of Base Rate, and the columns show combinations of witnesses and their testimony. The first column shows that if Blue cabs are involved in only 15% of accidents at night and the high credibility witness says the cab was "Green", the average response is only 29.1%. When Blue cabs were involved in 85% of accidents, however, the mean judgment was 49.9%. The last column of Table 1 shows that when the high credibility witness said that the cab was "Blue," mean judgments are 55.3% and 80.2% when base rates were 15% and 85%, respectively.

Table 1. Mean Judgments of Probability that the Cab was Blue (%).

	Witness Credibility and Witness Testimony				
Base Rate	H_Green	M_Green	No_Witness	M_Blue	H_Blue
15	29.13	31.26	25.11	41.09	55.31
30	34.12	37.13	36.31	47.37	56.31
70	45.97	50.25	58.54	60.89	73.20
85	49.91	53.76	66.96	70.98	80.19

Note: Each entry is the mean inference judgment, expressed as a percentage.

Analysis of variance tests the null hypotheses that people ignored base rate or witness credibility. The ANOVA shows that the main effect of Base Rate is significant ($F(3,306) = 106.2$), as is 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 1, the more diagnostic is the witness, the greater the effect of that witness's testimony. These results show that we can reject the assertions that people ignore utilize base rates and validity of evidence.

Table 2 shows the Bayesian calculations for the same conditions, simply using Bayes' theorem to calculate with the numbers given. (The probabilities have been converted to percentages.) Figure 1 shows a graph of the mean judgments as a function of the 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 of this literature.

Table 2. Bayesian Predictions (converted to percentages)

	Witness Credibility and Witness Testimony				
Base Rate	H_Green	M_Green	No_Witness	M_Blue	H_Blue
15	4.23	10.53	15.00	20.93	41.38
30	9.68	22.22	30.00	39.13	63.16
70	36.84	60.87	70.00	77.78	90.32
85	58.62	79.07	85.00	89.47	95.77

 ISSN 2047-3745

Conservatism described the fact that human judgments are less extreme than Bayes' theorem. For example, when 85% of accidents at night involve Blue cabs and the high credibility witness says the cab was "Blue", Bayes' theorem gives a probability of 95.8% that the cab was Blue; in contrast, the mean judgment is only 80.2%. Similarly, when base rate is 15% and the high credibility witness says the cab was "Green", Bayes' theorem calculates 4% and the mean judgment is 29%.

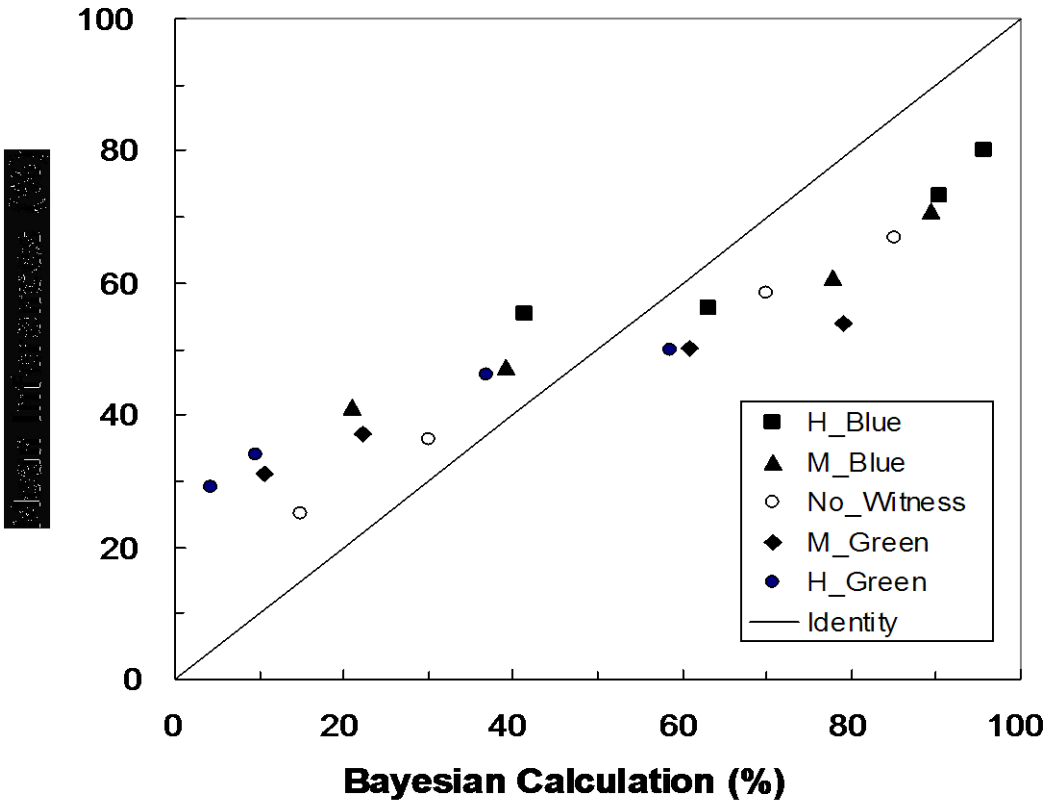

Figure 1. Mean inference that the cab was Blue, expressed as a percentage, plotted as a function of the Bayesian solutions, also expressed as percentages.

A problem with this way of graphing the data is that it does not reveal patterns of systematic deviation. People looking at such scatterplots are often impressed by "high" correlations. In fact, such correlations of fit, along with such graphs may lead researchers to systematically wrong conclusions (Birnbaum, 1974). The problem is that "high" correlations can coexist with systematic violations of a theory and correlations can even be higher for worse models, when prior measures are used.

In order to see the data better, they should graphed as in Figure 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 judgments in cases with no witness. The cross-over between this curve and others contradicts the additive model, including Wallsten's (1972) subjective Bayesian model and the additive model rediscovered by Novemsky and Kronzon (1999).

 ISSN 2047-3745

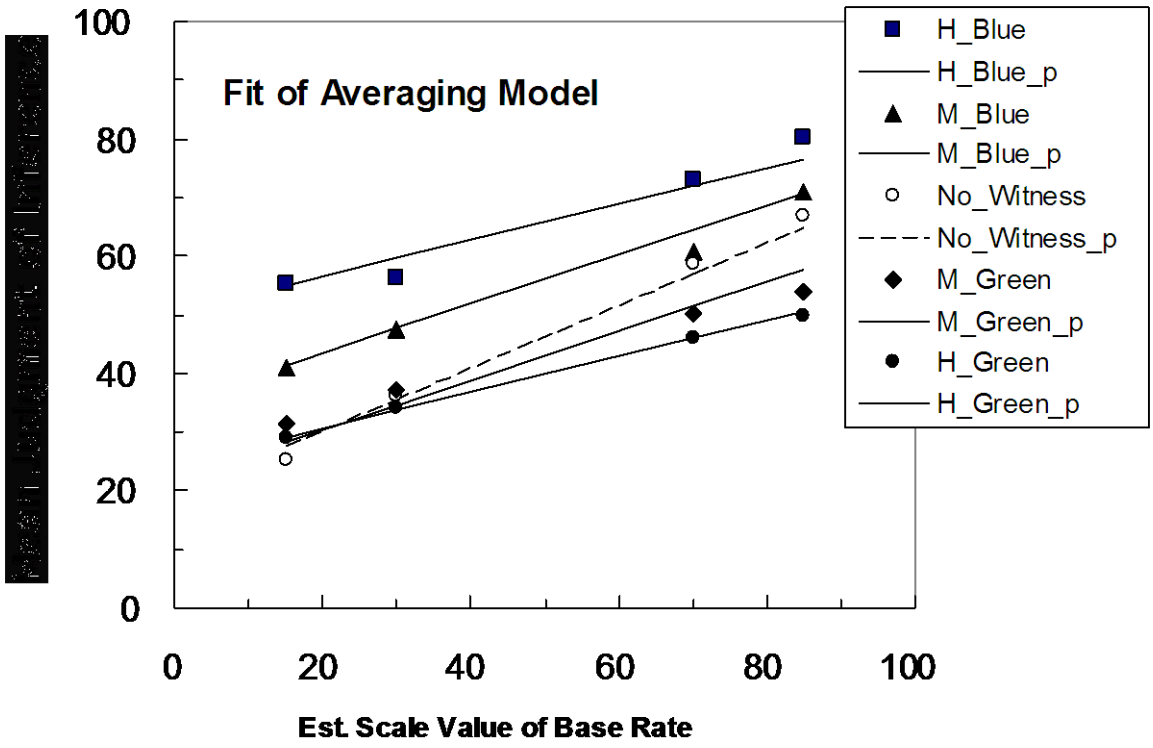

Figure 2. Mean judgments of probability that the cab was Blue, plotted as a function of the estimated scale value of the base rate. Filled squares, triangles, diamonds, and circles show results when a High credibility witness said the cab was "Green", a medium credibility witness said "Green", a medium credibility witness said "Blue" or a high credibility witness said "Blue." The solid lines show corresponding predictions of the averaging model. Open circles show mean judgments when there was no witness, and the dashed line shows corresponding predictions.

Instead, the crossover interaction is consistent with the idea 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 Base Rate of 15%, the mean judgment is 25%. However, when a medium witness also says that the cab was "Green," which should exonerate the Blue cab and thus *lower* the inference that the cab was Blue, the mean judgment actually *increased* from 25% to 31%.

Troutman and Shanteau (1974) reported similar results. They showed that when non-diagnostic evidence (which should have no effect) is presented, it 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. All of these findings are consistent with averaging rather than additive models.

In the old days, it was necessary to write special computer programs to fit models to data (Birnbaum, 1976; Birnbaum & Stegner, 1979; Birnbaum & Mellers, 1983). However, spreadsheet programs such as *Excel* can now be used to fit such models without requiring programming skills. Methods for fitting the

 ISSN 2047-3745

subjective Bayesian model and the averaging models via the Solver in *Excel* are described in detail for this type of study in Birnbaum (2001, Chapter 16). The subjective Bayesian model assumes the Bayesian formulas, but assumes that the effective value of the probabilities differs from the objective values stated in the problem.

Each model has been fit to the data in Table 1, to minimize the sum of squared deviations. The lines shown in Figure 2 are predictions of the averaging model. The estimated parameters are as follows: The weight of the initial impression, w_0, was fixed to 1, the estimated weights of the base rate, medium credibility witness and high credibility witness were 1.11, 0.58. and 1.56, respectively. Note that the weight of base rate was intermediate between that of the two witnesses.

The 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 their objective values. The estimated scale values for testimony that the cab was "Green" or "Blue" were 31.1 and 92.1, respectively. The estimated scale value of the initial impression was 44.5. This model correlates 0.99 with the mean judgments. When the scale values of base rate were fixed to their objective values, the model requires only 6 parameters and still correlates 0.99 with the data.

The sum of squared deviations provides a better index (of badness) of fit of the models than the correlation coefficient. For the null model, which assumes no effect of Base Rate or source validity, this sum of squares is 3027, for the subjective Bayesian (additive) model, it is 188, and for the averaging model, it is 84. For the simpler version of the averaging model with scale values of Base Rate equal to their objective values, the sum of squared deviations was 85.0.

Overview

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 effect, therefore there is no effect. However, one must keep in mind that when results fail to disprove the null hypothesis, they do not prove the null hypothesis.

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 is "bigger" than 221 with this design, I put little weight on such fragile between-subject findings.

In within-subjects designs, even the lawyer-engineer task shows effects of base rate (Novemsky & Kronzon, 1999). Although Novemsky and Kronzon (1999) 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 the appropriate designs, they would have been able to test and reject the additive model. They could have presented additional cases in which there were descriptions but no base rate information, base rate information but no witnesses (as in the dashed curve of Figure 2), different numbers of witnesses, or witnesses with varying amounts of information. In any of these manipulations, the implication of the averaging model is that the effect of any fixed 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 favored averaging over additive models in source credibility studies (e.g., Birnbaum, 1976, Figure 3; Birnbaum & Mellers, 1983, Figure 4C; Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976, Figures 2B and 3).

Edwards (1968), who reviewed early research on Bayesian inference, stated that human inferences might differ from Bayesian inferences for any of three basic reasons-- misperception, misaggregation, and 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

 ISSN 2047-3745

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 good fit in Figure 1.

When the proper experiments are conducted, 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 the other information available, but instead is diminished by the total weight of other information. This is illustrated by the dashed curve in Fig. 2.

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, the reliability of the source, the cue-correlation, or the signal-detection d' of the source. In the case of gambles, weight of a branch is affected by the probability of a consequence. In this experiment, the sources differed in their ability to distinguish Green from Blue cabs. The weight of the base rate was found to be intermediate between the medium and highly diagnostic sources.

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 the base rates were nearly equal to their objective values. In judgments of the value of cars, scale values are determined by estimates provided by sources who drove the car and by the blue book values.

Bias reflects a source's tendency to be rewarded or punished differentially for over as opposed to underestimating the value to be judged. Birnbaum and Stegner (1979) showed that source's bias affected the scale value of the source's testimony. In a court trial, bias would be affected by affiliation with the defense or prosecution. In an economic transaction, it would is affected by association with the buyer or seller.

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 total of hit rate and false alarm rate. Their data were consistent with the scale-adjustment model that bias affects scale value and is thus multiplied by the source's perceived validity.

The judge, who combines the information may also have a type of bias. This concept is known as the judge's *point of view*. The judge might be combining information to determine a buying price, selling price, or a "fair price". Each of these tasks induces a different viewpoint by the judge toward the evidence. An example of a "fair" price is when one person damages another's property and a judge is asked to give a judgment of the value of damages so that her judgment is equally fair to both parties in the dispute. Birnbaum and Stegner (1979) showed that the source's viewpoint affects the configural weight of the higher or lower valued branches. Buyer's 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 judgments 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 judgments of the values of gambles and decisions between them, take the same model, with viewpoint depending on the task to judge buying price, selling price, or "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 depend on identification with the buyer, seller, independent, or decision maker.

References

 ISSN 2047-3745

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 (Eds.), Social attitudes and psychophysical measurement (pp. 401-485). Hillsdale, N. J.: Erlbaum.

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(3), 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.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291.

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 Hillsdale, NJ: Erlbaum.

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 are available via the WWW from the following URL:
http://psych.fullerton.edu/mbirnbaum/bayes/resources.htm

 ISSN 2047-3745