SUBJECTIVE STATISTICAL INFERENCE BASED ON PURE ASSOCIATIONS

Robyn M. Dawes
Carnegie Mellon University, Dept. of Social and Decision Sciences, Pittsburgh PA 15213

Abstract: The "Fourth axiom of probability theory" quickly yields Bayes Theorem – and hence how to deal with inverse probabilities (e.g., $P(h|e)$ versus $P(e|h)$) – and shows in odds form the importance of likelihood ratios as opposed to simple likelihoods. This paper discusses a problem more basic than ignoring Bayes Theorem. The problem is making no comparisons at all! Just associating. Thus, if $P(A|B)$ or $P(B|A)$ is high (low) people tend to think that A and B "go together" (or don't). A distressing number of examples ranging from Nazi ideology to unjustified inferences in the mental health field will be presented, as well as some experimental research. In fact, A and B are often thought to "go together" – simply on the basis that $P(A)$ and $P(B)$ are both high (or low). Criminal behaviors are unusual, minority group membership is unusual, so would you believe? This "Von Rostoff effect" was originally found in paired associates learning of words or nonsense syllables. If a few stimuli were long rather than short and so were a few to be learned responses, people tended to believe that they were paired, even though statistically independent.

Following the conception of Fischhoff and Beyth-Marom (1983) as elaborated by Dawes (1998) the standard cognitive biases and heuristics in the irrational assessment of probability can be understood by considering simple forms of Bayes Theorem. Consider, for example, the relationship between a symptom S and a disease D, and suppose a diagnostician observes this symptom S. If the probability of the disease is assessed on the basis of a pure matching or association between the symptom and the disease, independent of considerations of conditional probability, there is no normative structure to which the judgment corresponds. More often, however, the judgment will be made on the basis of the conditional probabilities—a normatively correct judgment if the conditional is the probability of the disease given the symptom, $P(D|S)$, but a representative judgment if it is the probability of the symptom given the disease, $P(S|D)$. Unfortunately, there is a lot of evidence that the judgment is made on the basis of the latter relationship when conditional probabilities are considered at all.

The relationship between these two probabilities is given by rewriting Bayes theorems as:

$$P(D \mid S) = \frac{P(S \mid D)P(D)}{P(S)} \qquad (1)$$

Bayes theorem can also be written in terms of the "ration rule:

$$\frac{P(D \mid S)}{P(S \mid D)} = \frac{P(D)}{P(S)} \qquad (2)$$

Because Bayes Theorem and the ratio rule follow from the very definition of conditional probability, it is simply incoherent to equate the probability of the disease given the symptom with the probability of the symptom given the disease in the absence of a simultaneous belief that the probability of the disease and the probability of the symptom are identical. (If someone were to make a series of bets based on "fair betting odds" in such a belief, an opponent could make a Dutch book against that person.) Classic representative thinking—e.g., "she gave a typical schizophrenic response, therefore she must be schizophrenic"—embraces such an identity. What is critical of course are the *base rates* of D and of S, yet extensive research has shown that people underutilize such base rates in their intuitive judgments—or at least fail to incorporate them to a sufficient extent into their prior beliefs, given they are often seen as "not relevant." (Why, people often ask, should the *general* probabilities of these symptoms and diseases be relevant to a particular judgment about a particular person with a particular symptom? Such an argument is often followed by bland assertions to the effect that "statistics do not apply to the individual"—a wholly erroneous conclusion, as can be attested to by the heavy cigarette smoker with lung cancer.)

People sometimes do not ignore these base rates *entirely* in the judgments, particularly if the probabilistic nature of the judgment is

made clear (See Gigerenzer, Hell and Blank, 1988.). Nevertheless, the "underutilization" of such base rates is ubiquitous. In the psychological literature it was first decried by Meehl and Rosen (1955), who noted that in case conferences they attended clinical psychology and psychiatry judges tended to ignore base rates completely; later, underutilization has been established in a variety of experimental situations using a wide variety of subjects, especially, by Kahneman and Tversky (1972, 1973). For example, when asked to judge whether a description is of a particular engineer or a particular lawyer, subjects make the same judgment whether the experimenter states that the description was drawn from a pool of 70% engineers and 30% lawyers, or *vice versa*. As Bar-Hillel (1990) points out the 70/30 split does not appear to be "relevant," but it's relevance would most likely be clear if the split were 99/1, and certainly if it were a 100/0. (Again, we can hypothesize that people can recognize probabilistic arguments—although it's possible to be fooled on occasion—but often not make them spontaneously.) Thus, for example, someone who is said to be interested in sailing and carpentry and sharing model train sets with a child is judged to be much more likely to be an engineer than a lawyer, whether this person was drawn from a pool of 70% engineers of 70% lawyers. Making the sampling procedure absolutely "transparent" does, however, reduce that tendency.

The use of Bayes Theorems to understand the relationship between symptoms and disease leads naturally to its use to consider the general relationship between any hypotheses and any bit of evidence. Let e refer to some bit of evidence about whether a hypothesis h is or is not true; for example, a symptom may be regarded as a bit of evidence, and having a particular disease, a hypothesis; or the hypothesis may concern who is going to win the election in the United States in the year 2000 and a particular bit of evidence may be the margin of victory or defeat of the potential candidate in a more local election. Bayes Theorem expressed in terms of evidence and hypotheses is presented as

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)} \qquad (3)$$

This form of Bayes Theorem, while true, often leads to complications when trying to evaluate *P(e)*, the probability of the evidence. There is, however, a way to avoid having to estimate the probability of the evidence. What we can do is to consider *odds* that the hypothesis is true. These odds are the probability the hypothesis is true given the evidence divided by the probability that this hypothesis is false given the evidence—that is, by the ratio *P(h|e)/P(-h|e)*. Once we know the odds, it is trivial to compute the probability. The advantage of considering odds is that the denominator in Bayes Theorem cancels out when we compute these odds. That is,

$$\frac{P(h \mid e)}{P(-h \mid e)} = \frac{P(e \mid h)P(h)}{P(e \mid -h)P(-h)} \qquad (4)$$

Now we are in a position to categorize the most common forms of representative thinking. *Pseudodiagnosticity* refers to making an inference about the validity of hypothesis *h* on the base of evidence *e* without considering alternative hypotheses, particularly without considering hypothesis *-h*. Another way of stating pseudodiagnosticity is that it involves considering only the numerator in equation (3). People do that when they state that a bit of evidence is "consistent with" or "typical of" some hypothesis without concerning themselves about alternative hypotheses, or in particular the negation of the hypothesis. For example, in attempting to diagnose whether a child has been sexually abused, alleged experts in court often refer to symptoms "typical" of such abuse—without concerning themselves with how typical these symptoms are of children who have *not* been abused, or the frequency with which children have been sexually abused. What is happening is that only the first component of the numerator on the right hand side of equation (3) is being assessed without any consideration of the denominator.

Another type of representative thinking involves considering the probability of the evidence given the hypothesis or hypotheses

without looking at the second terms in equation (4), that is, the prior odds (which can be translated into objective "base rates" in situations involving classes of people). The reason that these odds are important is that—as indicated by equation (4)—they yield the *extent* of the evidence and hypotheses considered. For example, being a poor speller may be evidence of dyslexia in the sense that it has a very high probability given dyslexia. However, in order to diagnose dyslexia on the basis of poor spelling we have to know something about the base rates of poor spelling and the base rates of dyslexia in the population from which the person whom we wish to diagnose was chosen.

Dilution effects occur when evidence that does not distinguish between hypotheses in fact influences people to change their mind. the point is similar to that found in pseudodiagnosticity in that people do not realize that evidence that is not very likely given the hypothesis may be equally unlikely given alternative hypotheses, or, again, the negation of that hypothesis, but may in fact believe a hypothesis less as a result of collecting evidence that is unlikely if it is true. Dilution is simply the converse of pseudodiagnosticity.

Finally, it is possible to describe *availability* biases as well by equation (4). People believe that they are sampling evidence given the hypothesis, when in fact they are sampling this evidence given the hypothesis *combined with* the manner in which they are sampling. For example, when clinical psychologists claim that they are sampling, on the basis of their experience, characteristics of people who fall in a certain diagnostic category what they are really sampling is people who fall in that category *and who come to them*. For example, when we sample our beliefs about "what drug addicts are like," most of us are sampling on the basis of how the media presents drug addicts—both in "news" programs and in dramatizations (Dawes 1994b). Doctors often sample on the basis of their contact with addicts when these addicts are ill, perhaps gravely so, while police are often sampling on the basis of their experience with these same addicts during arrests and other types of confrontation. It is not surprising, then, that doctors are in favor of a "medical" approach to drug addiction including such policies as sterile needle exchanges, while police are in favor of much more punitive policies. Both are sampling evidence given evidence *and* their exposure to it.

All these anomalies have been demonstrated experimentally. It is important to point out, however, that the investigation of these anomalies did not spring simply from understanding Bayes Theorem and then from creating very clever experimental situations in which people systematically violate it when making inferences. The motive to study these anomalies of judgment arises from having observed them on a more informal basis outside the experimental setting—and trying to construct an experimental setting, which yields a greater measure of control, that will allow them to be investigated in a coherent manner. The reader is referred, for example, to the essay of Meehl (1977a) or the book of Dawes (1994a) for descriptions of these biases in the clinical judgment of professional psychologists and psychiatrists who rely on their own "experience" rather than (dry, impersonal) "scientific" principles for diagnosis and treatment.

This section will end with a somewhat detailed discussion of the bias that Dawes (2001 has found to be most prevalent in "expert" proclamations encouraging "child sex abuse hysteria." The main problem here is that hypotheses are not compared: instead, single hypotheses are evaluated in terms of the degree to which evidence is "consistent with" them; in addition, evidence is often *sought* in terms of its consistency with or inconsistency with "favorite hypotheses"—rather than in terms of its ability to distinguish *between* hypotheses. This type of pseudodiagnosticity has made its way into the legal system, where experts are allowed to testify that a child's recanting of a report of sexual abuse is "consistent with" the "child sexual abuse accommodation syndrome" (Summit, 1983). The finding is that many children who are *known* to have been sexually abused deny the abuse later (recant); therefore, the probability of the evidence (the child recants) given the hypothesis (actual abuse) is not as low as might be naively assumed; therefore, recanting is "consistent with" having been abused (i.e., can be considered part of a "syndrome" of accommodation to abuse).

The problem with this pseudodiagnostic reasoning is that it compares the probability of the evidence given the hypothesis to the probability of the *negation of the evidence* given the hypothesis, whereas a rational comparison is of the probability of the evidence given the hypothesis to the probability of the evidence given the *negation of the hypothesis.* When considering this latter comparison, we understand immediately that the recanting would be diagnostic of actual abuse only if the probability of recanting given actual abuse were higher than the probability of recanting given such abuse had not occurred —a highly implausible (not to mention paradoxical) conclusion.

Moreover, people actively seek evidence compatible or incompatible with a hypothesis rather than evidence which distinguishes between hypotheses, as has been extensively studied by Doherty and his colleagues (see for example Doherty, Mynatt, Tweney, & Schiavo, 1979). Consider, for example, subjects who are asked to play the role of medical student (and actual medical students prior to specific training, see Wolf, Gruppen, & Billi, 1985) to determine whether patients have one of two conditions. The conditions are explained, and subjects are told that a patient has two symptoms (i.e., bits of evidence for having either). The judges are then presented with the conditional probability of one of these symptoms given one of the diseases (e.g., that it is very likely that the patient has a high fever given the patient has meningitis). They can then choose to find out the probability of the other symptom given the same disease, the probability of the first symptom given the second disease, or the probability of the second symptom given the second disease. While virtually no subjects choose a different symptom and a different disease, the majority chooses to find the probability of the *second* symptom given the first disease (the "one that is focal" as a result of being told the first symptom). But for all these subjects know, of course, the first symptom may more or less typical of the *alternative disease,* as may the second symptom. Finding out that the probability of these two symptoms given only one disease in no way helps to distinguish between that disease and some other. *Of course, in real medical settings doctors may have prior knowledge of these other disease/symptom relationships*, but because the diseases are not even identified in the current setting, such knowledge would be of no help.

Now consider the behavior of a senior SS officer in a concentration camp trying to convince a new guard that the inmates "subhuman." The senior officer finds two starving inmates and tosses a crust of bread half way between them. They fight over it. You see,? The senior officer tells the junior one, "they are not really human. Would any of us fight of a crust of bread simply because we were hungry?" Here there are no comparisons at all. What are specifically lacking is a comparison to how these inmates would behave have they not been starving to death, or how they would have behaved where they not Jewish. Often, however, such hypothetical counterfactual are not readily available to the thought processes of people making inferences, and they can even require time and "cognitive effort" to generate, where the situation would be a guard in a concentration camp facilitates neither time to reflect nor intense taught. (for a fuller discussion, Dawes 2001.)

Finally, people often make judgments on the basis of matching the base rate probabilities of two categories even the absence of any of any information about the relationship between them—or worse yet in the presence of information that are independent. I quote an example from Klaus Fiedler of what he calls such "*pseudo-contingencies*." "When the distributions of TV consumption and aggressiveness within a school class is skewed (e.g., toward high levels of TV consumption and aggressiveness), the teacher may infer a 'positive' pseudo-contingency such that high levels of TV consumption seem to predict high levels of aggressiveness" (Fiedler, 2003). Conversely, teachers who observe high levels of TV consumption but very low levels of aggressiveness (e.g., in a group of students who tend be very lethargic "couch potatoes") may infer a negative relationship. Such an inference may be particularly problematic in social areas involving majority or minority groups. For example, a teacher in a middle-class suburban school, consisting primarily of white students may observe a majority of pro-social behaviors on the part of these students, and therefore associate pro-social behaviors to their Caucasian

background. This effect can be demonstrated experimentally by presenting subjects with instances of desirable and undesirable behavior of members of "group A" and of "group B," where one group is much more common than the other and the instances are primarily of desirable or undesirable behaviors (Hamilton and Gifford, 1976; Rothbart, reference to be provided). While the presentation involves quite carefully constructed lack of contingency between group membership and desirability of behaviors, subject judge there to be one. The researcher is also very careful not to identify the groups, so that no prior prejudice can be involved.

Fiedler primarily discusses situations in which no information about contingency is presented, but here we face a problem. A biased conclusion can be clearly established when the information presented is not supported, but what about making an inference in the absence of any information at? (See discussion in Coombs, Dawes, and Tversky [1972] of the distinction between biases that "fly's in the face" an information, verses judgments in the absence of information.)

Consider the possible contingency between dichotomous variables A and B where their base rates of positive instances may be compatible or discrepant. In the example presented in the figure here, the variable A has a base rate of 80 percent positive and 20 percent negative per instances. Now suppose that variable B matches that base rate; the maximal positive contingency between the two variables yields a phi value of 1.0, while negative contingency reaches only -.25 at the extreme. Conversely, if variable B has base rates of 20 and 80 percent, the maximal positive contingency is only +.25, while the maximal negative contingency is -1.00. Are we not entitled to have some sort of Bayesian ideas about the distribution of possible contingencies? For discussion see Fienberg and Kim, (1999), framed in the context of combining multiple graphical representations to create "larger" ones. But here at least we can say the potential for a large positive or negative correlation between the variables is indeed limited by the base rate match, which provides a (weak) realistic reason for making possible inferences about the match in the absence of the important information.

These situations are presented in Figure 1.

Figure 1A: Matching Base Rates

Figure 1B: Opposing Base Rates

What is striking, of course, is that this contingency judgment is made despite the presence of conflicting information (or is it really unambiguously conflicting?). The information always consists of a sample, which Bayesians maintain should be evaluated in light of a prior belief.) As Tversky and Kahneman point out in an important (but widely ignored) conclusion of their famous 1974 *Science* article on heuristics and biases, there are "usually effective, but they lead to systematic and predicable errors" (page 1131). Here, I have emphasized the errors. Just as there may be, however, a "quasi-rational reason" for looking only at the numerator of the likelihood ratio (if there is nothing else to look to look at-and one is willing to make some prior assumptions about what the denominator may be like), it may be quasi-rational to infer contingency on the basis of base rate matching in the total absence of contingency information.

Bibliography

Bar-Hillel, M. (1990). Back to base-rates. In R. M. Hogarth (Ed.), *Insights in Decision Making A Tribute to Hillel J. Einhorn.* Chicago: University of Chicago Press.

Coombs, C.H., Dawes, R.M. & Tversky, A. (1970). *Mathematical Psychology and Elementary Introduction*. Englewood Cliffs, NJ: Prentice Hall.

Dawes, R.M. (1994a). *House of Cards: Psychology and Psychotherapy Built on Myth*. New York: Free Press.

Dawes, R.M. (1994b). AIDS, sterile needles, and ethnocentrism. In L. Heath, R.S. Tinsdale, J. Edwards, E. Posavac, F.B. Bryant, E. Henderson-King, Y. Suarez- Balcazar & J. Myers (Eds.), *Social Psychological Applications to Social Issues. III: Applications of Heuristics and Biases to Social Issues*, (pp. 31-44). New York: Plenum Press.

Dawes, R.M. Behavioral decision making and judgment. In D. Gilbert, S. Fiske and G. Lindzey (Eds.) *The Handbook of Social Psychology*, *11*, 1998. Boston, MA McGraw-Hill.

Dawes, R.M. *Everyday Irrationality: How Pseudoscientists, Lunatics, and the Rest of Us Fail to Think Rationally*. Boulder, CO: Westview Press, 2001.

Doherty, M.E., Mynatt, C.R., Tweney, R.D. & Schiavo, M.D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111-121.

Fienberg, S.E. & Kim S-H. (1999) Combining conditional log-linear structures. *Journal of the American Statistical Association 94*, 229-239.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review; 90*, 239-260.

Gigerenzer, G., Hell, W. & H. Blank (1988). Presentation and content: The use of base rates as a continuous variable. *The Journal of Experimental Psychology*, *14*, 513-525.

Hamilton, D.L. & Gifford, R.K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12*, 392-407.

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

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

Meehl, P.E. (1977a). Why I do not attend case conferences. In P.E. Meehl (Ed.), *Psychodiagnosis: Selected Papers* (pp. 225-302). New York; W.W. Norton.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability in the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.

Rothbart, M. (1981). Memory processes and social beliefs. In D.L. Hamilton (Ed.), *Cognitive Processes in Stereotyping and Intergroup Behavior*. Hillsdale, NJ, Erlbaum.

Summit, R. C. (1983). The child sexual abuse accommodation syndrome. *Child Abuse and Neglect*, 177-193.

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

Wolf, F. M., Gruppen, L. D., & Billi, J.E. (1985). Differential diagnosis and the competing hypothesis heuristic—a practical approach to judgment under uncertainty and Bayesian probability. *Journal of the American Medical Association, 235*, 2858-2862.