CHAPTER I

The Power of Natural Frequencies: Extending Their Facilitating Effect to Complex Bayesian Situations

SUMMARY

Representing statistical information in terms of natural frequencies rather than probabilities dramatically increases performance in Bayesian inference tasks (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996). This beneficial effect of natural frequencies has already been demonstrated in a variety of applied domains such as medicine, law, and education. All the research on natural frequencies conducted so far has referred only to Bayesian situations where *one binary* (or: *dichotomous*) *cue* can be used to predict *one binary criterion*. Yet, real-life decisions often require dealing with situations where more than one cue is available or where cues have more than one value. This chapter provides empirical evidence that communicating the statistical information in terms of natural frequencies is both possible and beneficial even in such complex situations. The generalization of the natural frequency approach also turns out to be helpful when addressing some current critiques.

INTRODUCTION

The question of how hypotheses should be evaluated in view of empirical evidence is an ancient one. The normative view of classical rationality is that hypotheses are to be evaluated in probabilistic terms. In other words, when evaluating an uncertain claim, one does so by calculating the probability of the claim in the light of given information. The rigorous method for doing this was established during the enlightenment by Thomas Bayes and Pierre Simon de Laplace. One important school of statisticians, known as *Bayesians*, often defends this method as the only valid one.

One of the classical debates of cognitive psychology during the last decades focused on the question: Do unaided humans reason the Bayesian way when updating their belief in a hypothesis in view of new evidence? The mathematical expression for updating of hypotheses in the probabilistic framework is given by Bayes' rule, which expresses the probability of the hypothesis H given the data D as

$$p(\mathrm{H} \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \mathrm{H})p(\mathrm{H})}{p(\mathrm{D} \mid \mathrm{H})p(\mathrm{H}) + p(\mathrm{D} \mid \overline{\mathrm{H}})p(\overline{\mathrm{H}})} \qquad (1.1)$$

Most cognitive psychologists have sought for empirical evidence on the way humans reason by giving their participants the task of finding the probability that a hypothesis is true in light of provided data. Their participants thus had to find the probability of a hypothesis H, given the data D, that is, $p(\mathrm{H} \mid \mathrm{D})$, provided with all the information on the terms appearing on the right side of Equation 1.1, that is, $p(\mathrm{H})$, the a priori probability of the hypothesis H, $p(\mathrm{D} \mid \mathrm{H})$, the probability of the data given that the hypothesis is true, and finally $p(\mathrm{D} \mid \overline{\mathrm{H}})$, the probability of the data given that the hypothesis is not true. Edwards (1968) found that if people have to update their opinions, they change their view in the direction proposed by Bayes' rule. However, he also reported that people are "conservative Bayesians" in the sense that they do not update their prior beliefs as strongly as required by the Bayesian norm.

Fourteen years later Eddy (1982) treated the same question focusing on experts. He found that physicians do not make judgments that follow Bayes' rule when solving the following task (which represents a prototypical Bayesian situation):

The probability of breast cancer is 1% for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

In this case "breast cancer (B)" is the hypothesis H and the positive mammogram outcome (M+) is the datum D. The full information for the task in probabilistic symbols is:

$p(\mathrm{B}) = 1\%$,	the a priori probability or, in medical terms, the *prevalence* of breast cancer is 1%.
$p(\mathrm{M+} \mid \mathrm{B}) = 80\%$,	the hit rate or, in medical terms, the *sensitivity* of mammography, is 80%.
$p(\mathrm{M+} \mid \overline{\mathrm{B}}) = 9.6\%$,	the false alarm rate of mammography or, in medical terms, the complement of the *specificity* of mammography, is 9.6%.

The relevant question now can be expressed as "$p(\mathrm{B} \mid \mathrm{M+}) = ?$", and Equation 1.1 becomes

$$p(\mathrm{B} \mid \mathrm{M+}) = \frac{p(\mathrm{M+} \mid \mathrm{B})\,p(\mathrm{B})}{p(\mathrm{M+} \mid \mathrm{B})\mathrm{p}(\mathrm{B}) + p(\mathrm{M+} \mid \overline{\mathrm{B}})\,p(\overline{\mathrm{B}})} = \frac{(.8)(.01)}{(.8)(.01) + (.096)(.99)} = .078 \qquad (1.2)$$

Bayes' rule yields a probability of breast cancer of 7.8%. However, Eddy (1982) reported that 95 out of 100 physicians estimated this probability to be between 70% and 80%. While Eddy argued that this is due to the confusion of $p(\mathrm{M+} \mid \mathrm{B})$ and $p(\mathrm{B} \mid \mathrm{M+})$, Kahneman and Tversky (1972, p. 450) attributed this phenomenon to people's ignoring the base-rate (which stands synonymously for the a priori probability) and concluded: "In his evaluation of evidence man is apparently not a conservative Bayesian: he is not Bayesian at all." This "base-rate neglect" became one of the famous fallacies investigated in the "heuristics and biases" program (Kahneman, Slovic & Tversky, 1982). After a few years of research on the base-rate neglect, Bar-Hillel (1980, p. 215) stated that "the base-rate fallacy is a matter of established fact".

Later in the 1990s this view changed. Gigerenzer and Hoffrage (1995) focused on the aspect of the *representation* of uncertainty. They established that what makes the task difficult is not Bayesian reasoning per se, but the format of information provided to the participants. In Eddy's (1982) task, quantitative information was provided in probabilities. Gigerenzer and Hoffrage (1995) argued that probabilities make the computation of the Bayesian posterior probability more complex than "natural frequencies", which have been historically the "natural" format of information for the human mind (Figure 1.1). They illustrated this argument by referring to a physician who does not know the sensitivity and the false alarm rate of the test. After testing a large number of patients, she would "naturally" obtain numbers as those depicted in the tree on the right side of Figure 1.1.

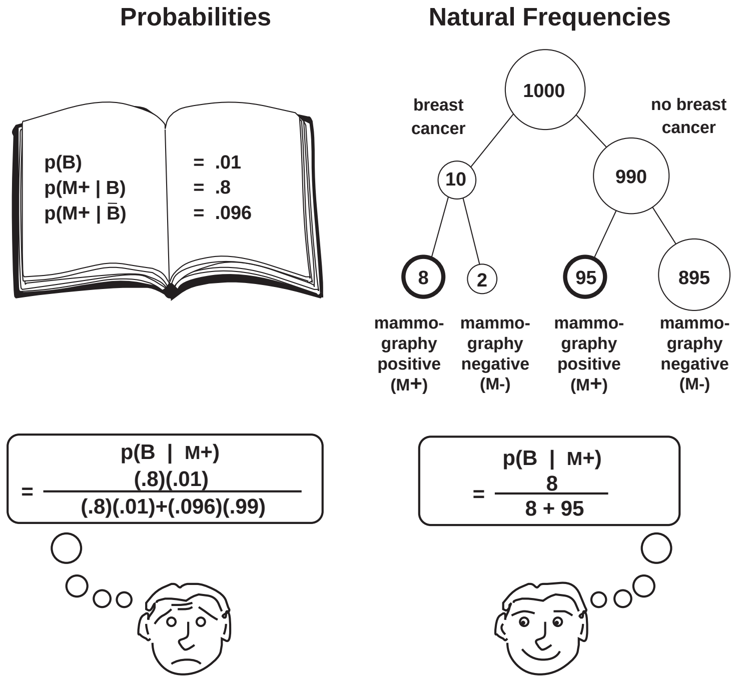
Probabilities
Natural Frequencies
p(B) = .01
p(M+ | B) = .8
p(M+ | B̄) = .096
1000
breast cancer
no breast cancer
10
990
8
2
95
895
mammo-graphy positive (M+)
mammo-graphy negative (M-)
mammo-graphy positive (M+)
mammo-graphy negative (M-)
p(B | M+) = (.8)(.01) / (.8)(.01)+(.096)(.99)
p(B | M+) = 8 / 8 + 95

Figure 1.1: Bayesian algorithm in terms of probabilities (on the left) and in terms of natural frequencies (on the right).

Gigerenzer and Hoffrage (1995) use this tree to describe the *natural sampling* process (p. 687) and call the frequencies that result from natural sampling, *natural frequencies* (Gigerenzer & Hoffrage, 1999, pp. 425-426). Providing the statistical information for Eddy's task in terms of natural frequencies yields the following version of the task:

10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 out of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in a routine screening. How many of these women do you expect to actually have breast cancer?

With this formulation almost half of their participants gave the Bayesian answer, namely, 8 out of 103 women (= 7.8%). This empirical observation is consistent with the fact that the Bayesian computation is simpler (requires fewer operations) for natural frequencies than for probabilities.

At present, there is a lively debate on the facilitating effect of such frequency representations. However, often in this debate either the concept of natural frequencies is misunderstood or the participants' success in solving Bayesian tasks with natural frequencies is attributed to some other, seemingly more general factor. These issues will be discussed in the section "Addressing current critiques of the natural frequency approach".

It is remarkable that so far, almost all research on this topic has only been concerned with a very small part of the world of Bayesian reasoning, namely, with situations having the structure of the mammography problem reported above. The features of such a "basic" situation are the following: The cue that is used to infer the criterion value is *binary* (e.g., a positive or a negative test result). The criterion is also *binary* (e.g., a woman either has breast cancer or not). Furthermore, only *one* cue (e.g., one test result) is used to infer the criterion value (see Table 1.1).

	"Basic" Bayesian situation
Number of cue values	2
Number of criterion values	2
Number of cues	1

Table 1.1: Number of cue values, criterion values, and cues in a "basic" Bayesian situation

Yet, inferences in the real world are not always so simple. Often, more than one cue is available and sometimes cues as well as criteria can have more than two values (i.e., a polychotomous structure). If one or more of the three numbers in a Bayesian situation is larger than in Table 1.1 we call it a "complex Bayesian situation".

Our purpose with the present chapter is threefold. First, we theoretically enhance and generalize the natural frequency approach by extending the basic situation to cases where more than one cue is available, and to cases with non-binary cue or criterion values. Second, we designed two studies to test whether natural frequencies facilitate reasoning in these "complex" situations. Third, we address the most common critiques of the natural frequency approach. In doing so, we take advantage of our extensions to complex Bayesian situations as they shed new light on the current debate on the basic situation.

EXTENSIONS OF THE BASIC SITUATION

Gigerenzer and Hoffrage (1995) had left open whether the beneficial effect of natural frequencies can be generalized to complex situations. Massaro (1998) recently has questioned this possibility. With respect to Bayesian reasoning, and referring to the findings of Gigerenzer and Hoffrage, he claimed that in the case of two cues "a frequency algorithm will not work" (p. 178). However, he did not provide any empirical evidence for this claim. We will now provide a theoretical generalization by manipulating each number in Table 1.1 separately. Each one of these situations is described in the following with the help of a corresponding tree diagram.[3]

[3] Regarding the basic situation there are two ways of organizing the information – both available for frequencies as well as for probabilities: Gigerenzer and Hoffrage (1995) use *frequency trees* (Figure 1.4 on the right), while, for instance, Fiedler et al. (2000) prefer 2 x 2 *frequency tables* to visualize basic Bayesian situations. We use tree diagrams because they are more flexible. Note that although tables can be extended to situations with non-binary information, they cannot display multiple cue situations (such as represented by Figure 1.4).

Number of Cue Values: 3

Cues are often not binary. The outcome of a medical test, for instance, could be positive, negative, or unclear. Natural frequencies describing such a situation with three possible test outcomes ("trichotomous cue") can be depicted in a tree structured as in Figure 1.2.

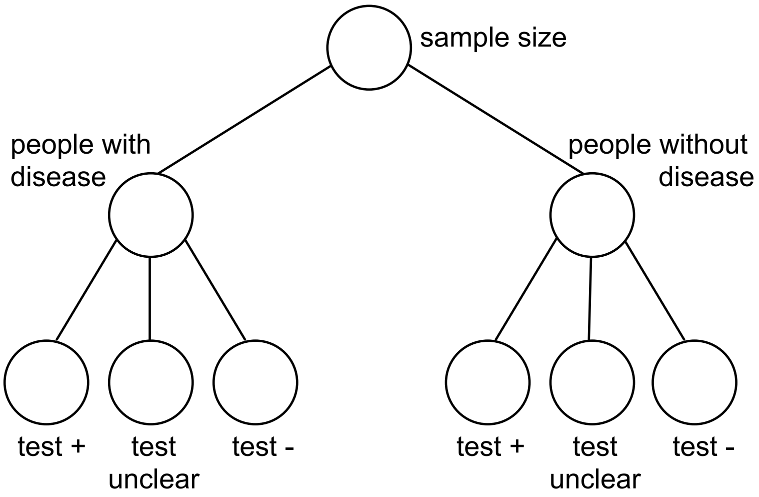

Figure 1.2: Tree structure describing a Bayesian situation containing a trichotomous cue

Clearly cues with any number of values ("polychotomous cue") can be modelled by adding the corresponding nodes at the lowest level. In our studies we only tested tasks with 3 cue values.

Number of Criterion Values: 3

Bayesian inferences are not restricted to situations in which the sample is divided into two complementary groups (e.g., ill and healthy people). Medical tests, for instance, could be sensitive to more than one disease. Let us consider a medical test that is sensitive to two diseases, namely disease 1 and disease 2. If these diseases do not occur simultaneously, the sample has to be divided into three groups corresponding to three different medical hypotheses (healthy, disease 1, disease 2), as depicted in Figure 1.3.

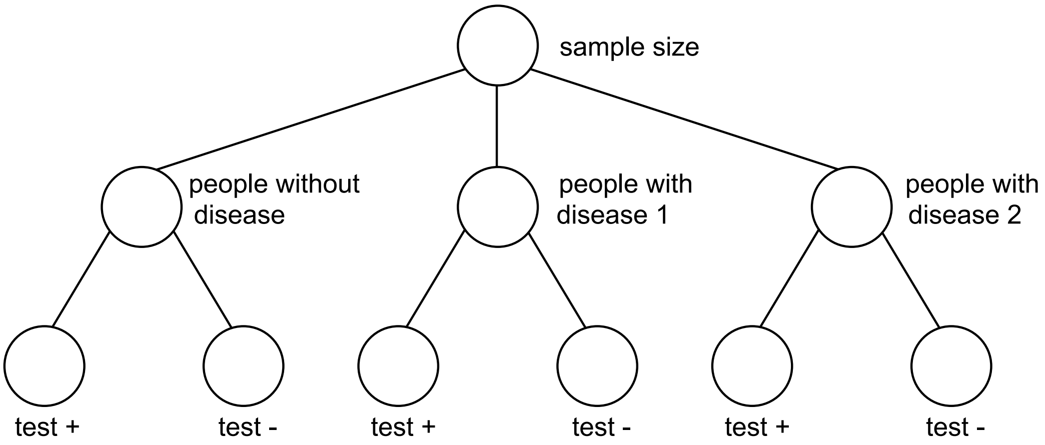

Figure 1.3: Tree structure describing a Bayesian situation containing a trichotomous criterion

The extension to even more than three hypotheses ("polychotomous base-rate") is straightforward. Another hypothesis might be, for instance, that the patient suffers both from disease 1 and 2. In our studies we only tested tasks with 3 criterion values.

Number of Cues: 2 and 3

Even Bayesian decision situations where more than one cue is available can be modelled with the help of natural frequency trees. Consider the case in which a disease is diagnosed based on two medical tests. For binary criterion and cue, the frequency tree looks like the one depicted in Figure 1.4.

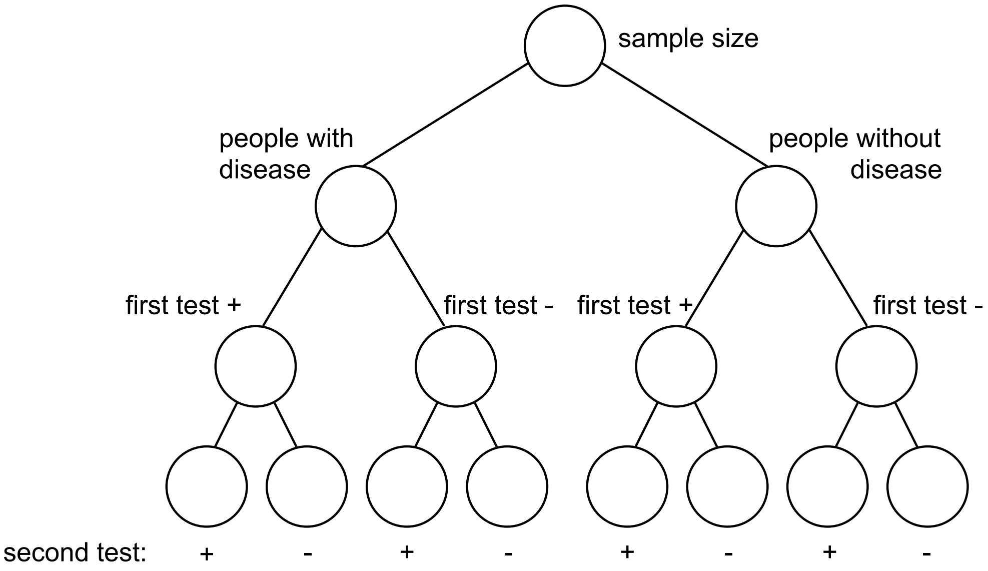

Figure 1.4: Tree structure describing a Bayesian situation containing two cues

If more tests are available we simply have to add branches downward, a new level for each new test. For *n* binary cues we will have 2^{n+1} nodes at the lowest level. In our studies we tested Bayesian tasks with two (Figure 1.4) and with three cues (Figure 1.4 with one extra level).

All the above trees are skeletons that correspond to general Bayesian situations. Specifying concrete complex Bayesian situations requires filling these trees with natural frequencies (such as in Figure 1.1). To investigate the impact of different information formats in such complex situations we conducted two studies. In both studies we compared participants' performance when the information was presented in terms of probabilities with that of participants who received the information in terms of natural frequencies. Whereas in Study 1 untrained participants were tested, Study 2 was devoted to the effect of previous training.

STUDY 1

Method

In Study 1 we presented advanced medical students (*N* = 64) of the Free University of Berlin with four medical diagnostic tasks (summarized in Table 1.2). Each participant worked on all four tasks. Task 1 was a Bayesian task corresponding to Figure 1.2, where we extended Eddy's mammography task by adding unclear test results. Task 2 was a Bayesian task corresponding to Figure 1.3, where a test could detect two diseases, namely Hepatitis A and Hepatitis B. Tasks 3 and 4 were Bayesian tasks with two and three cues, respectively. In Task 3, which corresponds to Figure 1.4, breast cancer had to be diagnosed based on a mammogram and an ultrasound test. In Task 4 an unnamed disease had to be diagnosed on the basis of three medical tests, simply named Test 1, Test 2, and Test 3. The four tasks are summarized in Table 1.2.

Task	*Summary* (*The complete wordings of the tasks are shown in Appendix I.1)*
1	The outcome of one medical test can be positive, negative, or unclear (three cue values: positive, negative, and unclear mammogram result)
2	One medical test can detect two diseases (three criterion values: Hepatitis A, Hepatitis B, and healthy)
3	The outcomes of two medical tests were provided (two binary cues: mammogram and ultrasound)
4	The outcomes of three medical tests were provided (three binary cues: medical test 1, medical test 2, and medical test 3)

Table 1.2: The four tasks of Study 1

Each student received the statistical information for two of the four tasks in probabilities and for the other two in natural frequencies. As an illustration of the tasks consider the two different versions (probability version vs. natural frequency version) of Task 3:

Task 3: Probability version

The probability of breast cancer is 1% for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 9.6% that she will also have a positive mammogram. If a woman has breast cancer, the probability is 95% that she will have a positive ultrasound test. If a woman does not have breast cancer, the probability is 4% that she will also have a positive ultrasound test. What is the probability that a woman at age 40 who participates in routine screening has breast cancer, given that she has a positive mammogram and a positive ultrasound test?

Task 3: Natural frequency version

100 out of every 10,000 women at age 40 who participate in routine screening have breast cancer. 80 out of every 100 women with breast cancer will receive a positive mammogram. 950 out of every 9,900 women without breast cancer will also receive a positive mammogram. 76 out of 80 women who had a positive mammogram and have cancer also have a positive ultrasound test. 38 out of 950 women who had a positive mammogram, although they do not have cancer, also have a positive ultrasound test. How many of the women who receive a positive mammogram and a positive ultrasound test do you expect to actually have breast cancer?

Figure 1.5 illustrates how the tree corresponding to Task 3 would look after filling in the natural frequencies in the "empty" tree of Figure 1.4.

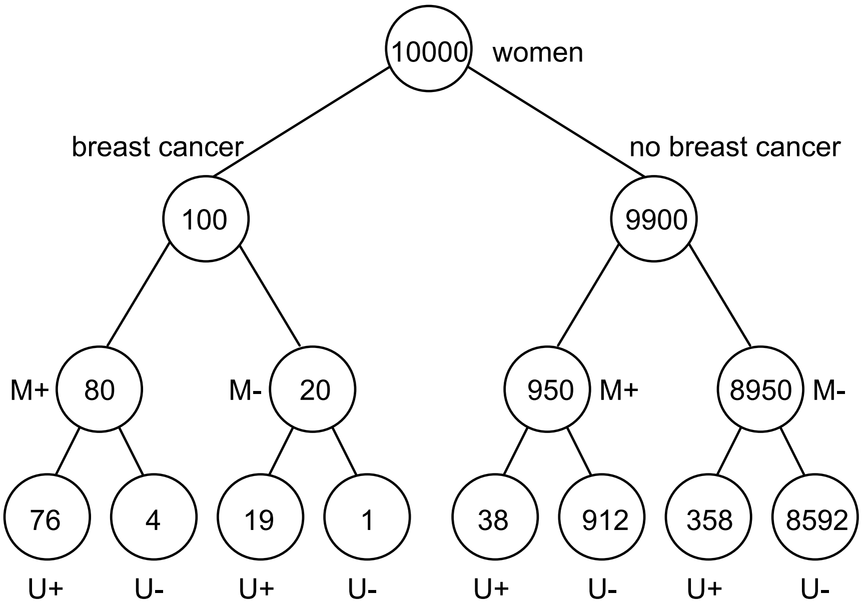

Figure 1.5: Natural frequency tree according to Task 3

It is important to note that – concerning Study 1 – these trees serve only for visualization. The participants in this study were neither presented with trees nor told to construct them; rather, they had to solve the task based only on the wording. Besides requiring a numerical answer, we also asked them to justify their inferences. This allowed us a more detailed view of their reasoning processes.

Results

On average, the medical students worked one hour on all four tasks. Figure 1.6 displays the percentage of Bayesian inferences (correct solutions) for each of the four tasks.

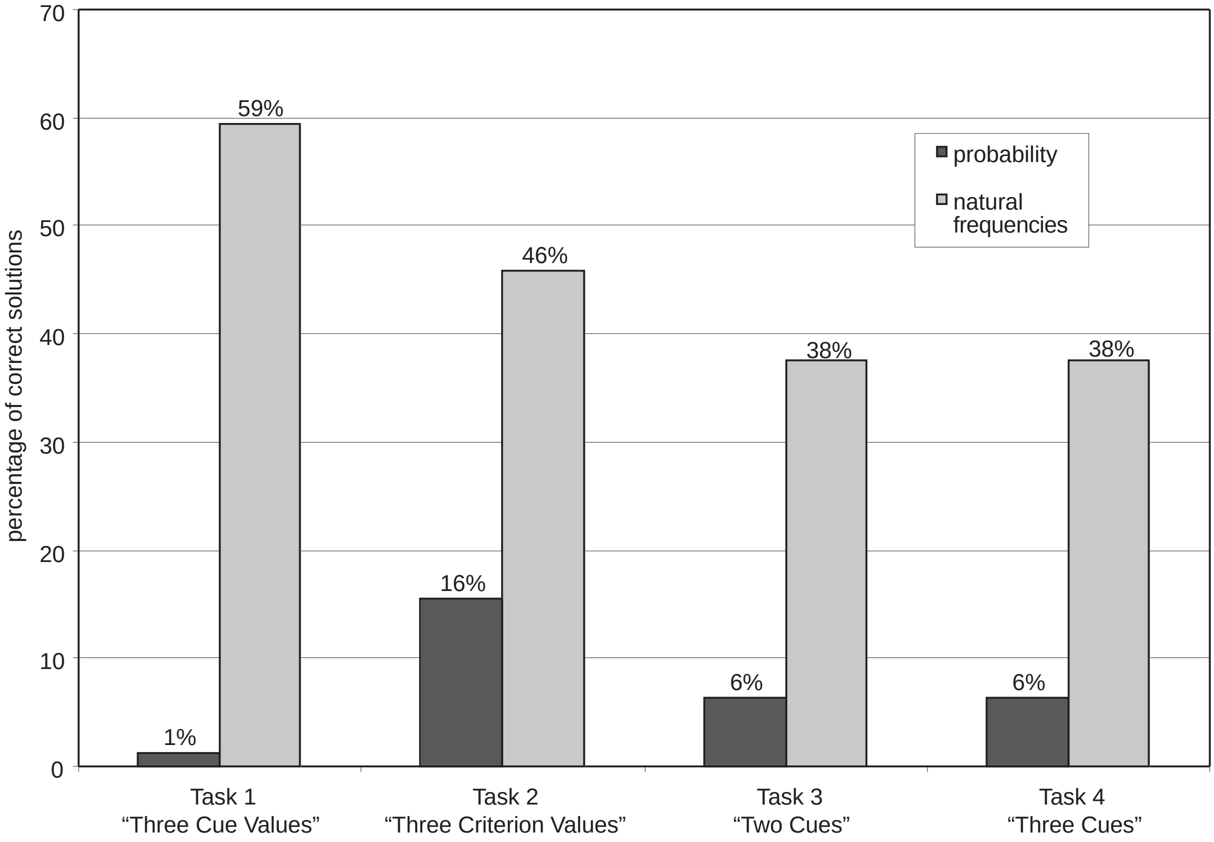

Figure 1.6: Results of Study 1: Percentage of Bayesian inferences for each of the four tasks

In all of the tasks, replacing probabilities with natural frequencies helped the medical students make better inferences. In particular, participants' performance on Task 3 contradicts Massaro's (1998) claim that in the case of two cues "a frequency algorithm will not work" (p. 178).

The percentage of Bayesian inferences averaged across the probability versions of the four tasks was 7% and across the natural frequency versions it was 45%. Natural frequencies were most helpful in Task 1, where the difference in terms of participants' performance between the probability and the frequency version was 59% – 1% = 58%. In the other tasks there was still an increase in participants' performance from the probability versions to the natural frequency versions of about 30%. The comparison between Task 3 and Task 4 suggests that for both the probability and the natural frequency versions, it did not matter whether information was provided on two or on three cues nor whether this information referred to named or unnamed tests (a possible explanation for the unexpected equal performance when provided with two vs. three cues will be given in the general discussion).

Note that the percentages of Bayesian inferences averaged across the four tasks (7% and 45% for probabilities and natural frequencies, respectively) are similar to those obtained in the classical experiments on basic tasks (16% and 46% for probabilities and natural frequencies, respectively) by Gigerenzer and Hoffrage (1995). These results demonstrate that the facilitating effect of natural frequencies on Bayesian reasoning can be extended to complex tasks – at least to cases with either 3 cue values *or* 3 criterion values *or* 3 cues.

Whereas the results of Study 1 were obtained without previously training participants, in Study 2 we examined the effects of different training approaches on complex Bayesian reasoning.

STUDY 2

Sedlmeier and Gigerenzer (in press) and Kurzenhäuser and Hoffrage (2001) have shown that frequencies can also provide a means for training participants for Bayesian situations. In one of their studies Sedlmeier and Gigerenzer gave participants a computerized tutorial in the basic situation – either by teaching them Bayes' rule or by teaching them how to represent the probability information in terms of natural frequencies. Compa-ring the efficiency of both procedures by testing the same participants with basic tasks in which the statistical information was provided in terms of probabilities yielded two results: First, the immediate learning success was twice as high with the representation training (frequency trees). Second, this success was stable. Even five weeks after training the performance of the participants remained a high 90%, whereas with traditional training with probabilities (applying Bayes' rule) the performance dropped to 15%.

In Study 2 we addressed the question of whether a simple written instruction on how to solve a basic task, rather than a computerized training program, could improve participants' ability to solve *complex* tasks.

Method

We recruited advanced medical students (*N* = 78) from Berlin universities (none of them was a participant in Study 1) and divided them randomly into 3 groups. Instead of training, participants received a two-page instruction sheet on the basic mammography task – each group received a different instruction sheet with respect to the solutions presented (all 3 instructions are shown in Appendix I.2). The instructions for Group 1 introduced the basic mammography task by means of probabilities and presented the solution in terms of Bayes' rule. The instructions for Group 2 also expressed the mammography task by means of probabilities but then showed participants how to translate probabilities into natural frequencies and place these into a frequency tree. In the instruction sheet of Group 3, the basic mammography task was already formulated in terms of natural frequencies and participants saw how these frequencies were placed into a frequency tree.

After participants finished studying their instruction sheet, they were exposed to the *complex* Tasks 1 and 3 of Study 1. Participants of Group 1 and 2 got the probability versions of these tasks, whereas participants of Group 3 faced the same tasks expressed in terms of natural frequencies. Table 1.3 summarizes the design of Study 2.

	Basic task used for instruction	Solution presented	Complex tasks tested
Group 1 (*N* = 27)	Mammography task, formulated in terms of probabilities (original task by Eddy, 1982)	Probabilities are inserted into Bayes' rule	Tasks 1 and 3 of Study 1 (both tasks provided in probabilities)
Group 2 (*N* = 25)	Mammography task, formulated in terms of probabilities (original task by Eddy, 1982)	(a) Probabilities are translated into natural frequencies (b) These are placed into a frequency tree and the correct answer is extracted from the tree	Tasks 1 and 3 of Study 1 (both tasks provided in probabilities)
Group 3 (*N* = 26)	Mammography task, formulated in terms of natural frequencies (adaptation of Eddy's task by Gigerenzer & Hoffrage, 1995)	Natural frequencies are placed into a frequency tree and the correct answer is extracted from the tree	Tasks 1 and 3 of Study 1 (both tasks provided in natural frequencies)

Table 1.3: The design of Study 2

Results

Figure 1.7 displays the percentages of Bayesian inferences in Task 1 and Task 3 separately for the three experimental groups.

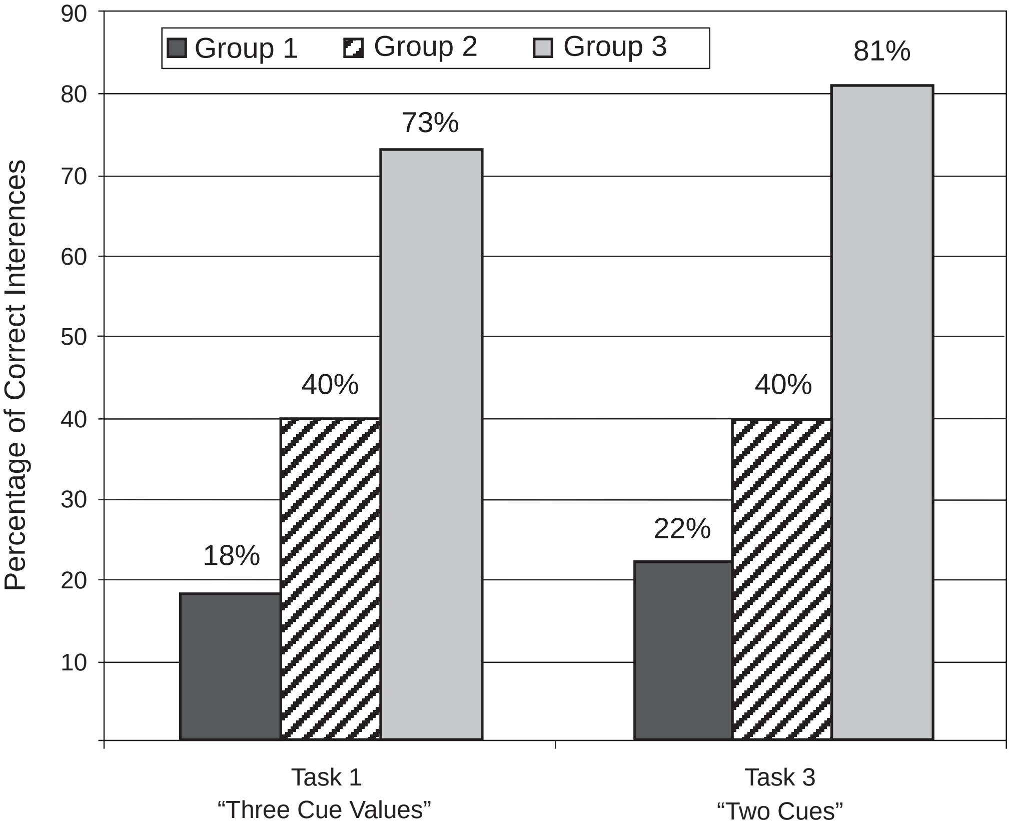

Figure 1.7: The results of Study 2: Percentage of Bayesian inferences for complex Tasks 1 and 3 after instruction

First, note that participants' performance on Task 1 was similar to their performance on Task 3. This suggests that the learning effect of the different walked-through solutions (Groups 1, 2, 3) is basically the same no matter which of the two complex situations is tested afterward. Let us discuss participants' performance for each group:

Group 1: In this group participants had to generalize Bayes' rule (in probabilities) to more complex situations. Figure 1.7 reveals that Group 1 exhibited the worst performance when confronted with complex Bayesian tasks (18% for Task 1 and 22% for Task 3). However, compared with participants' performance for the same tasks in Study 1 (1% for Task 1 and 6% for Task 3, see Figure 1.6) the proportion of

Bayesian inferences was quite high. At least for some of the participants being exposed to Bayes' rule was beneficial: They managed to extend the basic formula to the case involving an unclear test result (which amounts to adding a corresponding term in the denominator) and to the case of two test results (which amounts to applying Bayes' rule twice).

Group 2: Participants of this group had learned from their instruction how to translate probabilities into natural frequencies for the basic situation. In spite of being tested in terms of probabilities just like Group 1, 40% of participants in Group 2 obtained the correct solutions. These participants arrived at the correct solutions by performing the following steps: First they had to translate five probabilities (rather than three as was the case for the basic situation) into natural frequencies appropriately. To construct a corresponding tree they had to add nodes to the tree they had seen in the instruction. In the case of Task 1 they had to add nodes at the lowest level (which describes the mammography outcomes, see Figure 1.2) and in the case of Task 3 they had to add another level (which describes the ultrasound outcomes, see Figure 1.4). From these trees they finally had to extract the frequencies needed for the Bayesian solutions (namely in the form of "Laplacian proportions", i.e., the ratio of successful cases divided by the number of cases that fulfill the condition).

Group 3: Participants of Group 3 were the only ones who were trained and tested with natural frequencies. This instruction method reached a high performance of 73% (Task 1) and 81% (Task 3). Recall that without instruction, performance on the same two tasks was lower, 59% and 38%, respectively (Study 1). Study 2 shows a stronger learning effect with Task 3 (two cues). Analyzing participants' protocols reveals that participants found it easier to add another level to the basic tree than to add nodes within a level. In other words, extending Figure 1.1 to Figure 1.4 seemed to be more intuitive to participants than extending Figure 1.1 to Figure 1.2.

To summarize: Study 2 shows that a simple instruction on how to solve Bayesian tasks in the *basic* situation can amplify performance in *complex* situations. The highest levels where obtained when both the trained and the tested task were consistently formulated in terms of natural frequencies.

In the next section we focus on several critiques and misunderstandings concerning the natural frequency approach regarding the basic situation and we close by providing a *definition of natural frequencies*, generalized to complex situations.

ADRESSING CURRENT CRITIQUES OF THE NATURAL FREQUENCY APPROACH

Until 1995 in cognitive psychology most research on Bayesian reasoning examined the problems people encounter when confronted with probability representations (for an overview see Koehler, 1996). In the following years most research on cognitive aspects in Bayesian reasoning referred either to Gigerenzer and Hoffrage (1995) or to Cosmides and Tooby (1996). Many articles since 1998 comment on or criticize – both theoretically and empirically – the natural frequency approach put forward in these two articles. In the following we address Massaro (1998), Macchi and Mosconi (1998), Mellers and McGraw (1999), Lewis and Keren (1999), Johnson-Laird et al. (1999), Fiedler et al. (2000), Over (2000a, 2000b), Macchi (2000), Evans et al. (2000), Sloman and Slovac (2001) and Girotto and Gonzales (in press).

While we value these contributions for propagating the importance of Bayesian inference, most of the critiques are based on vague definitions (e.g., "nested sets") or on misreadings of the original concepts (e.g., of the term natural frequencies).

THE CONFUSION BETWEEN NATURAL FREQUENCIES AND FREQUENCIES PER SE

The most frequent critique is that not just any kind of frequencies foster insight: additional specific conditions have to be fulfilled (Lewis & Keren, 1999; Macchi, 2000; Evans et al., 2000; Girotto & Gonzales, in press). Lewis and Keren, for instance, tested the following frequency version of Eddy's task:

> *Ten out of every 1,000 women at age forty who participate in routine screening have breast cancer. Eight hundred out of every 1,000 women with breast cancer will receive a positive mammography report. Ninety-six out of every 1,000 women without breast cancer will also receive a positive mammography report. Here is a new representative sample of women at age forty who received a positive mammogram report in routine screening. How many of these women do you expect to actually have breast cancer?*

Their participants performed poorly. This poor performance, however, is not in conflict with the natural frequency approach, as suggested by Lewis and Keren. Rather the opposite is true: it is even predicted by this approach (Gigerenzer & Hoffrage, 1995, 1999). Lewis and Keren's frequencies do not stem from one observed sample and are thus not natural frequencies. Rather, their frequencies have been normalized with respect to several different classes, which makes Bayesian computation more complex than with natural frequencies. For instance, in their task two different numbers are assigned to the state "breast cancer": Observe that the first piece of information (about the prevalence) mentions 10 women with breast cancer, whereas the second (about the sensitivity) mentions 1,000 women with breast cancer. In real-life settings, in contrast, a physician does not observe 10 woman with breast cancer *and* 1,000. Rather, natural frequencies consist of a breakdown of *one* class into subsets, as shown in Figure 1.1.

This property of natural frequencies is due to the process by which they are acquired and is described by the term *natural sampling*. Gigerenzer and Hoffrage (1995) defined the term: "The sequential acquisition of information by updating event frequencies *without* artificially fixing the marginal frequencies (e.g., of disease or no-disease cases) is what we refer to as *natural sampling*. In contrast, [...], an experimenter may want to investigate 100 people with disease and 100 people without disease. This kind of sampling with fixed marginal frequencies is not what we refer to as natural sampling" (Gigerenzer & Hoffrage, 1995, p. 686). Gigerenzer and Hoffrage called frequencies stemming from natural sampling *frequency formats* (1995, p. 687) or later, synonymously, *natural frequencies* (1999, pp. 425-426). This focusing on a special kind of frequencies was not reckoned with by some authors addressing the natural frequency approach.

Johnson-Laird et al. (1999) reported an experiment similar to that of Lewis and Keren and concluded:[4] “In fact, data in form of frequencies by no means guarantee good Bayesian performance” (p. 81). What was apparently not clear to either Lewis and Keren (1999) or to Johnson-Laird et al. (1999) is that the frequencies considered by them did not follow Gigerenzer and Hoffrage’s recipe. Instead of *natural frequencies* both of them used *normalized frequencies* (i.e., fixing the marginal frequencies).

We now illustrate (Figure 1.8) the difference between natural frequencies and normalized frequencies in *complex* Bayesian situations. Krauss, Martignon, and Hoffrage (1999, p.171-172) constructed the following normalized frequencies describing the Bayesian situation of Task 3:

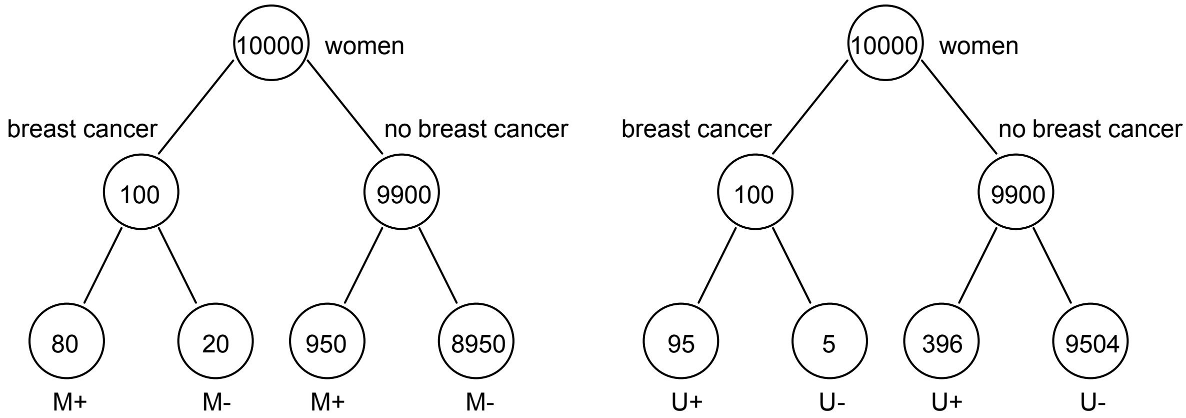

Figure 1.8: Frequencies that are *not* natural frequencies (Task 3; two cues)

In Figure 1.8, each cue is represented by a separate frequency tree. Krauss et al. (1999) tested the following version derived from Figure 1.8:

[4] Johnson-Laird et al. (1999) refer to an experiment – actually conducted by Girotto and Gonzales (2000) – which suffers from the the same problem like Lewis and Keren’s task. Girotto and Gonzales provided the following information (translated from French and abbreviated): „Out of 100 people, 10 are infected. Out of 100 infected people, 90 have a positive test. Out of 100 non-infected, 30 have a positive test.“ In judging *p*(disease | positive test) participants, of course, turned out to be very bad.

> *100 out of 10,000 woman at age 40 who participate in routine screening have breast cancer. 80 out of 100 women with breast cancer will receive a positive mammogram. 950 out of 9,900 women without breast cancer will also receive a positive mammogram. 95 out of 100 women with breast cancer will receive a positive ultrasound test. 396 out of 9,900 women, although they do not have cancer, nevertheless obtain a positive ultrasound test. How many of the women who receive a positive mammogram and a positive ultrasound test do you expect actually to have breast cancer?*

Only 15% of participants could solve this frequency version correctly. Thus, in complex situations the statement remains true that *not just any* frequency representation works.

Why are the numbers of Figure 1.8 not natural frequencies? The answer is: They do not refer to *one* total reference class and thus cannot be placed in *one* tree diagram. The frequencies of Figure 1.8 could only be obtained by testing disjoint groups of patients: one group with mammography test only (the tree on the left), the other group with ultrasound test only (the tree on the right). A physician who had sampled according to Figure 1.8 would have difficulties to answer the question "p(B | M+ & U+) = ?", because she has not sampled information on women with results on both test. Generally, to answer a question regarding *two* cues (e.g., M+ & U+), natural sampling means randomly selecting individuals displaying both cues, such as in Figure 1.5.

Note that furthermore the frequencies of Figure 1.5 represent *tests taken sequentially*. Because a physician usually applies medical tests one after the other the results of Krauss et al. (1999) suggest: if the structure of information matches real-life settings the corresponding inferences will be facilitated. The two-cue case also shows that information should not consist of a juxtaposition of single natural frequency trees (as in Figure 1.8) but rather should be placeable in *one* tree.

As mentioned above, Gigerenzer and Hoffrage (1995) explicitly stated that not just any frequencies help and *again* made this clear in a reply to Lewis and Keren and Mellers and McGraw (Gigerenzer & Hoffrage, 1999). Astonishingly, there is a flow of new articles, which confuse natural frequencies with normalized frequencies. For instance, Macchi and Mosconi (1998, p. 83) and Evans et al. (2000) try to disprove the natural frequency hypothesis by running experiments with normalized frequencies. It is not surprising that Evans et al. (2000) found that the results of their recent experiments "provide little encouragement for the hypothesis advanced by Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995) that frequency formats *per se* are easier than probability formats" (p. 206). Why these authors fail to acknowledge Gigerenzer and Hoffrage's definitions of natural sampling (1995, p. 686) and natural frequencies (1999, pp. 425-426) remains unclear.

DO WE NEED "NESTED SETS", THE "SUBSET PRINCIPLE", OR "PARTITIVE FREQUENCIES"?

The second misunderstanding builds on the first. In this section we address three similar critiques that all share one feature: They each attribute the facilitating effect to a single property of natural frequencies.

Advocates of the so-called "nested sets hypothesis" claim that it is not the frequentistic nature of information that fosters insight, but its nested sets property. They believe that if information is structured in terms of nested sets, the required inference will be simple. This "nested sets hypothesis" was promoted by Sloman and Slovac (2001) and – in a slightly different manner – by Evans et al. (2000). Evans et al. (2000) proposed "that it is the cueing of a set inclusion mental model that facilitates performance" (p. 211) and thus viewed this explanation as an alternative to the natural frequency approach. It is important to note that these authors do not specify their hypotheses precisely. What they probably mean is that their sets arise from successive partitioning *starting from one and only one large sample*. Without this last specification the facilita-ting effect of nested sets would be gone. For instance, in the task corresponding to the nested sets of Figure 1.8 participants performed poorly.

But even if the requirement is fulfilled and the partitioning procedure starts from one and only one set, the nested sets hypothesis fails to reflect a fundamental aspect of natural frequencies, namely that natural sampling means drawing randomly from a population. To illustrate that the nested sets property of information is not sufficient, consider the following counter-example: One could represent Lewis and Keren's (1999) numbers (see above) in terms of "nested sets" by combining all women to reach a grand total of 2,000 women (Figure 1.9).

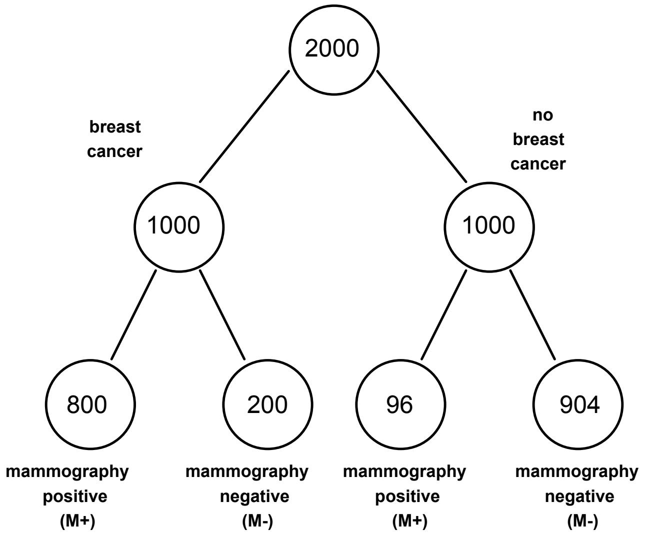
2000
breast cancer
no breast cancer
1000
1000
800
200
96
904
mammography positive (M+)
mammography negative (M-)
mammography positive (M+)
mammography negative (M-)

Figure 1.9: Nested sets that are *not* naturally sampled and thus are not natural frequencies

Although this tree represents nested sets, observe that the partition of the grand total does not reflect the base rate in the population. The given base rate in Lewis and Keren's task – 10 out of 1,000 have breast cancer – cannot be integrated in the tree and the grand total – 2,000 women – does not stem from natural sampling. Although the frequencies of Figure 1.9 consist of nested sets, they do *not* represent what we would call "natural frequencies".

Johnson-Laird et al. (1999) and Girotto and Gonzales (in press) similarly believe that the frequentistic format is not crucial. Instead both attribute the facilitation of Bayesian reasoning to what they call the "subset principle". This principle is (Johnson-Laird et al., 1999, p. 80): "Granted equiprobability, a conditional probability, p(A | B), [...] equals the frequency (chance) of the model of *A and B*, divided by the sum of all the frequencies (chances) of models containing B." Note that this definition matches

Equation 1.2 in Gigerenzer and Hoffrage (1995, p. 687), which gives the quotient that solves a Bayesian task in terms of natural frequencies. Thus, Johnson-Laird et al.'s (1999) subset principle brings nothing new – except their adding the word "model" to the discussion. Their principle is rather a mere redescription of yet another property of natural frequencies. Furthermore – as with the nested sets hypothesis – their principle is not sufficient for explaining facilitation effects. This is again illustrated by Figure 1.9, where the subset principle is fulfilled but insight is not provided.

Macchi (1995, 2000) went along very similar lines with her notion of "partitiveness". In 1995 she had recognized that even in a probabilistic format stressing the relationship between the provided pieces of information can make a difference. She reached an improvement of performance by changing the wording in the cover story slightly. She introduced the term "partitive" (instead of "nested") to describe wordings that clarify the relationship between the percentages in the cover story. The following example is taken from Macchi (1995):

(Non-partitive) "In a population of adolescents, 80% of *suicide attempts* are made by girls and 20% by boys. The percentage of *death by suicide* is three times higher among boys than among girls. What is the probability that an adolescent, selected at random from those who had died by suicide, was a boy?"

(Partitive) "In a population of adolescents, 80% of *suicide attempts* are made by girls and 20% by boys. The percentage of *suicide attempts that result in death* is three times higher among boys than among girls. What is the probability that an adolescent, selected at random from those who had died by suicide, was a boy?"

In Macchis' "partitive" version the percentage of *suicide attempts that result in death* is clearly a part of the set described by the base rate (80% and 20% of *suicide attempts*). In her opinion, "the partitive formulation has the triple effect of identifying the data reference set, eliminating confusion [...] and making it possible to perceive and make use of the relationships between the data." (Macchi, 2000, p. 220). Note that this statement remains true, if "partitive" is replaced by "natural frequency". Indeed, Macchi (2000) also introduced "partitive" frequency versions of Bayesian tasks that are nothing else but natural frequency versions. It is not surprising that partitive probability versions can be solved more easily, because they share crucial properties of natural frequencies. Probabilities per se are normalized and do not refer to one sample size but – when expressed as percentages – to a sample of 100. First when Macchi de-normalized probabilities by expressing them as percentages *and* connecting these percentages to each other by appropriate wordings these "probabilities" could express the nested subset relation bet- ween the information pieces. Similarly, Sloman and Slovac (2001) write: "... the advantage of frequency over single-event probability formats should disappear, when nested-set relations are made transparent in both formats...". Note that, like for Macchi, in the probability format the nested set character has to be *made* transparent. Natural frequencies contain this aspect inherently.

Connecting these reflections with the applications of Bayesian reasoning for risk communication delivers a new argument: If we want to help experts to understand Bayesian situations in medicine and law, natural frequencies are of great help (Hoffrage et al., 2000) and no need is felt for approximations or mimics of this representation format. Why should we teach medical students to represent a Bayesian diagnosis situation with partitive probabilities if we can use natural frequencies?

COMPUTATIONAL COMPLEXITY

Some critiques (e.g., Macchi and Mosconi, 1998) claim that in the basic task natural frequencies eliminate all need for computation, because this format already contains the numbers that compose the answer. Yet, this is exactly the point Gigerenzer and Hoffrage (1995) had made: "Bayesian algorithms are computationally simpler when information is encoded in a frequency format [...]. By 'computationally simpler' we mean that (a) fewer operations (multiplication, addition, or division) need to be performed and (b) the operations can be performed on natural numbers (absolute frequencies) rather than fractions (such as percentages)" (p. 687). Of course, this statement remains true considering natural frequencies in complex Bayesian situations.

Our claim is that the story has to be told the other way around: Probabilities *introduce* the need for computations. By observing samples and assessing subsets' sizes we are naturally *performing and understanding* Bayesian inferences. Things only become cumbersome when the statistical information is expressed in terms of probabilities: From that moment on Bayesian inferences consist of distressing inversions. Side effects of probability formats are base-rate neglect and the confusion of different conditional probabilities. Performing Bayesian inference by means of natural frequencies, instead, requires no inversions and base rates cannot be neglected because every leaf of the tree carries the information about base rates implicitly. From this viewpoint, the base rate fallacy is not "a matter of established fact" (as concluded by Bar-Hillel, 1980, p. 215), but rather a byproduct – or even an artifact – of the normalization of natural frequencies to conditional probabilities. This might also explain the seemingly paradoxical observation that animals are good Bayesians (Gallistel, 1990; Real, 1991) whereas humans appear not to be. Animals in experiments are not faced with artificially constructed conditional probabilities, but in their natural ecology they rather experience natural frequencies. In the studies that documented base-rate neglect humans were provided with conditional probabilities, which cannot be observed directly in nature.

STOCHASTIC DEPENDENCY AND CAUSALITY

Over (2000a) claims that "without higher-level hypotheses about causation or independence, we would be stuck with what can be misleading information from natural sampling" (p. 190). We want to comment on this statement beginning with the issue of stochastic dependency. From a mathematical point of view one could criticize that our Task 3 contains no information on the conditional dependence of the mammogram and the ultrasound test, given the disease. In this task, p(U+ | B) = 95% is given, but there is no information about whether this probability is independent of the outcome of the mammography test (M+ or M-). For instance, when breast cancer is present, a positive mammography test may increase the probability that the ultrasound test is also positive. Imagine that in our Task 3 p(U+ | B & M+) was equal to 70% rather than to 95%. Even in this case, a physician would *automatically* sample the corresponding frequencies that would allow the correct estimate of p(B | U+& M+). Instead of 76 women with the configuration "B, M+, U+" (see Figure 1.5) she would now sample 56 women with this configuration (see Figure 1.10).

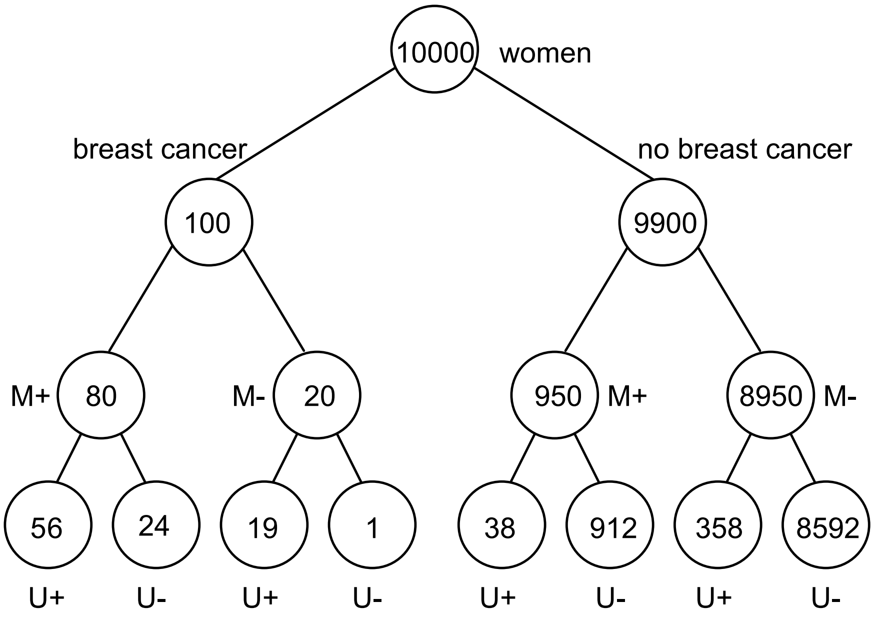

Figure 1.10: Natural frequency tree reflecting two stochastical dependent cues

With the numbers 56 and 38 she can simply compose the answer “56 out of 94 (= 60%)” to the question “p(B | M+ & U+) = ?”, without ever having heard of the concept of stochastic dependency. It is important to note that the physician could ignore any notion of stochastic dependency while relying on natural sampling. The concept of stochastic dependency only acquires importance when probabilities are introduced.

In addition, Over (2000b) stresses the importance of considering higher-level hypotheses about causation. He gives the following example: “Suppose that 36 out of 48 people feel sick and 27 of these 36 have eaten blackberries from some site. We notice this and that 9 out of the 12 people who did not feel sick also ate the blackberries. Now natural sampling will tell us that 27 people feel sick out of the 36 who ate the blackberries, and such a high relative frequency might lead us, if we used only natural sampling, to conclude that eating the blackberries caused the feeling of sickness.”

Yet, to the best of our knowledge, there is no theory claiming that naive reasoning reduces causality assessments to just one proportion, and therefore we do not see why anyone in this situation should come up with such a conclusion. Because the proportion of sick people in the total sample ($^{36}\!/_{48}$ =.75) coincides with the proportion of sick people in the subset of those who ate blackberries ($^{27}\!/_{36}$ =.75) there is no reason to attribute the sickness to the blackberries. Because the predictive accuracy of "having eaten blackberries" is equal to zero, it probably will not be used as a cue at all. If we want to detect causality within naturally sampled information we compare our "Bayesian result" with the base-rate and only if these numbers differ noticeably will we hypothesize a causal relation.

Note that both higher-level concepts are related because detecting causality can be reduced to assessing the direction of a dependency (Cheng, 1997). Over is right in stressing the importance of forming higher-level hypotheses and he is also right that only considering the *result* of Bayesian inferences is not sufficient to form such hypotheses. Yet, a simple comparison of the base-rate with the actual Bayesian inference (of course we do not need to compare exact percentages – rough proportions are totally sufficient) allows us to form hypotheses on causality.

THE AMBIGOUS USE OF THE TERM "PROBABILITY VERSION"

Some authors attempt to show that under certain circumstances participants can also handle probabilities. Yet, there is a problem shared by the probability versions of Fiedler et al. (2000), Macchi (2000), and Evans et al. (2000): Their so-called "probability versions" contain information in terms of absolute numbers. For instance, the statistical information of the "probability version" of Fiedler et al. (2000), is: [5]

> *The study contains data from 1,000 women. 99% of the women did not have breast cancer and 1% had breast cancer. Of the women without breast cancer 10% had a positive mammogram and 90% had a negative mammogramm. Of the women with breast cancer 80% had a positive mammogram and 20% had a negative mammogram. Task: What´s the probability of breast cancer, if a women has a positive mammogram result?*

It is not surprising that participants can cope with these "probability versions". Providing the total sample ("the study contains data from 1,000 women") describes

[5] Fiedler et al.'s (2000) main focus is the impact of the different ways information can be sampled. Since in the classical text problem paradigm information is already sampled and provided by the experimenter, we will not address this issue in this chapter. A dispute with Fiedler et al.'s sampling paradigm can be found in Kurzenhäuser and Krauss (2001).

precisely drawing a random sample. Therefore it mimics the procedure of natural sampling and facilitates computational demands a great deal: Computing 1% of 1,000 women is just a simple division *that leads automatically to natural frequencies* (namely, "10 out of 1,000 women have breast cancer"). The following statement "80% of the women with breast cancer had a positive mammogram" now directly leads to "8 out of these 10 women have a positive mammogram", etc.). The correct answer now easily can be derived – with no danger of confusing conditional probabilities, committing the base-rate fallacy or struggling with any inversions. Moreover, one should note that Fiedler et al. (2000) did not provide *probabilities*, but *relative frequencies* (pure percentages without mentioning the term "probability"). As Gigerenzer and Hoffrage (1995) showed, participants performance is poor with relative frequencies when the grand total is *not* provided, because it is not clear what the percentages should be related to. We already saw that "starting from one grand total" is one of the keys for understanding Bayesian situations. Of course it is of scientific interest to provide versions like Fiedler et al.'s (2000), but instead of "probability version" such versions should rather be called "relative frequency version with providing the grand total and asking a probability question". In our view, a *probability version* should only contain one format, namely probabilities. Evans et al. (2000), who wrote under the title "Frequency vs. Probability Formats" were imprecise when adopting even the two formats mentioned in their title: Instead of "probability versions" in their experiments they used "relative frequency versions" and in addition they misinterpreted the term "frequency formats" as "just take any frequencies".

Girotto and Gonzales (in press) also strained the term "probability version" and thus claimed that "probabilities" can easily be handled. They introduced a representation in terms of "numbers of chances", which actually are the same as natural frequencies if one replaces cases with chances, such as replacing "in 4 cases out of 100" with "4 chances out of 100". If these "numbers of chances" corresponded to normalized frequencies, the effect would be gone. Numbers of chances thus just represent a clever way to translate natural frequencies into a language that looks like a single-event statement.

DEFINING "NATURAL FREQUENCIES"

Since many of the critiques concerning the natural frequency approach centered on the concept of "natural frequencies" we close this chapter by providing a definition of this concept, which ought to clarify all misunderstandings. This definition, for the first time, is formulated in the general case. For our comprehensive definition of the term "natural frequencies" we make use of the notion of frequency trees:

Imagine a tree diagram that starts from one root-node and is split into *N* levels. Assume that the number of branches starting at each node of the *n*th level is constant for *n* ($n = 1, \dots, N$). Furthermore, assume that the number *a* ($a \in \boldsymbol{N}$) assigned to a node is equal to the sum $\sum_{i=1}^{m} b_i$, where the b_i ($b_i \in \boldsymbol{N}_0$) are the numbers assigned to nodes branching from it.

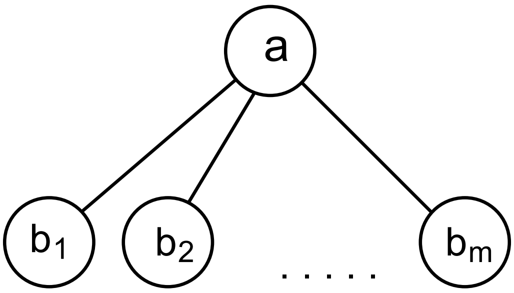

Figure 1.11: In a categorization tree for each node $a = \sum_{i=1}^{m} b_i$

If this holds for every node of a tree, this tree is a *categorization tree*.

Definition:
A set of frequencies is a set of natural frequencies if and only if

(a) a *single* categorization tree can display the frequencies
(b) the frequencies can be considered the result of natural sampling, that is, the frequencies are obtained by a random sample, which is represented by the root-node

Frequencies that follow this definition refer to one randomly drawn reference set and automatically consist of nested sets. Note that it is *not* required that the set of frequencies corresponds to a Bayesian task. Let us see how this definition provides straightforward answers to typical questions:

Question 1: Can the simple expression “35 out of 42” be considered a natural frequency, even if no other information is presented and no task is posed?

Answer: Yes, “35 out of 42” can be displayed in a tree consisting of two nodes and it can be considered the result of a natural sampling process.

Question 2: If we reconfigure Figure 1.1 by first partitioning the sample with respect to the mammogram result and then with respect to the state of illness (tree on the right in Figure 1.12), do we still have natural frequencies?

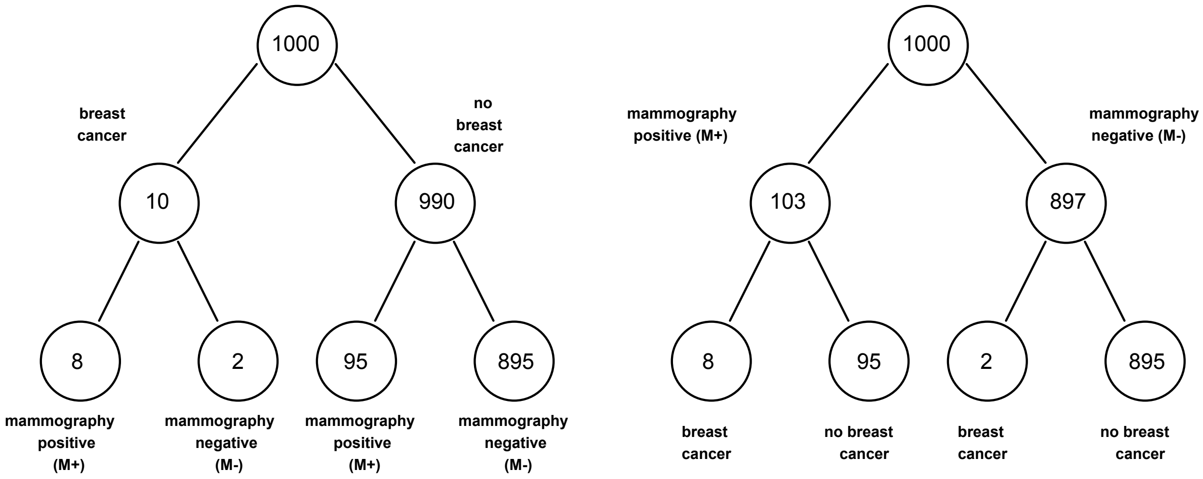
1000
breast cancer
no breast cancer
10
990
8
2
95
895
mammography positive (M+)
mammography negative (M-)
mammography positive (M+)
mammography negative (M-)
1000
mammography positive (M+)
mammography negative (M-)
103
897
8
95
2
895
breast cancer
no breast cancer
breast cancer
no breast cancer

Figure 1.12: Natural frequency trees that describe the same naturally sampled reference set and that only differ with respect to the order of subsetting

Answer: Yes, the natural frequency approach does not fix the order of subsetting. Although in a classical Bayesian text problem the provided information is *inverse* to the question, this is not a requirement for natural frequencies.

Question 3: Imagine a simulator that draws randomly from a given population. If the randomly drawn numbers – for instance, the base rate – deviate from the expected values, do we still have natural frequencies?

Answer: Yes, these numbers are natural frequencies because both conditions of the definition are fulfilled.

Question 4: Can we judge whether an unlabeled tree containing just frequencies reflects natural sampling?

Answer: No, because not all frequencies that fit a tree are automatically natural frequencies. Natural frequencies cannot be seen content free. For instance, whether the set "1,000 out of 2,000", "800 out of 1,000", and "96 out of 1000" is a set of natural frequencies depends on what these numbers stand for. Figure 1.9 reflects a natural frequency tree if it describes a population where the base-rate is 50% and the two relevant conditional probabilities are 80% and 9.6% respectively.

CONCLUSION

The possibility of communicating statistical information in terms of natural frequencies is not restricted to what we called the basic situation with one binary predictor for inferring a binary criterion. In situations where more than one cue is provided, or where either cues or the criterion have more than two possible values, the statistical information can still be represented in terms of natural frequencies. In two studies we have shown that in complex Bayesian situations natural frequencies have the same beneficial effect as could be demonstrated for basic situations in previous research (Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2000). Considering such extensions is important as in many real-life situations they are the rule rather than the exception. To reach a medical diagnosis, for instance, usually more than one test is applied, or in court trials usually more than one piece of evidence is available.

However, extending the natural frequency approach certainly has its limitations. Natural frequencies are no doubt helpful in facilitating inferences in the case of two or three cues, where the amount of information is still cognitively manageable. In a Bayesian decision situation with 10 binary cues things may change. In this case, the corresponding frequency tree would amount to more than 2,000 natural frequencies. Although Krauss, Martignon and Hoffrage (1999) suggested a way of how to reduce complexity in such situations (by only considering the so-called *Markov frequencies*), we doubt that people are Bayesians regardless of the degree of complexity. Rather, it is our claim that the human mind is equipped with an adaptive toolbox containing simple heuristics that allow "fast and frugal" decisions – even in highly complex environments (Gigerenzer, Todd, & the ABC Research Group, 1999). These simple heuristics are helpful in making inferences in situations under limited time, with limited knowledge, and within our cognitive and computational constraints. For instance, when memory limitations keep us from making use of natural frequencies, we could base our decision just on the best cue (e.g., Gigerenzer and Goldstein, 1999; see also general discussion).

The common denominator between these fast and frugal heuristics and the natural frequency approach is ecological rationality. While the fast and frugal heuristics are ecologically rational, as they are adapted to the structure of information in the environment (Martignon & Hoffrage, 1999), it is ecologically rational to represent the statistical information required for a Bayesian inference task in terms of natural frequencies, as the human mind is adapted to this format. Future research has to reveal the crucial variables (e.g., number of cues) that trigger switching from being a Bayesian to being fast and frugal. In Martignon and Krauss (in press) we have undertaken the first tentative steps in this direction.

Egon Brunswik once referred to the organism and its environment as "equal partners". This was not meant to say that they are equal in all aspects of structural detail; rather, Brunswik suggested the simile of a married couple: "Perhaps the organism could be seen as playing the role of the husband and the environment that of the wife, or the reverse may be argued as well" (1957, p. 5). The advocates of the heuristics and biases program divorced natural reasoning processes (Brunswik's organism) from the naturally available information format (Brunswik's environment). Our studies have shown that changing the experimental situation so that it better reflects essential features of the environment (by providing natural frequencies and thus maintaining relevant base-rate information), reasoning will not only become more accurate, but also more consistent with the relevant statistical norm, namely Bayes' rule. Thus, if Brunswik's Mr. Cognition is reunified with his wife, Mrs. Environment, they will give birth to sound reasoning.