Chapter 1: Introduction: What are natural frequencies, and why are they relevant for physicians and patients?

It just takes a look at the daily newspaper to remind us that statistics are an important part of our everyday lives: We read about the voter turnout in political polls, gains and losses on the stock market, food risks, unemployment rates, the probability of rain tomorrow – and this is still the front page. In medicine, too, statistics have become indispensable: How frequently does a specific disease occur? How often does it come with a certain symptom? How often does a drug cause side effects? What is the risk of treatment A, compared to treatment B? Physicians have to deal with these questions all the time, and the statistics that form the answers to these questions guide decisions on the assignment of diagnostic tests and treatments. For instance, assumptions about the probability with which a suspicious x-ray picture of the breast actually predicts breast cancer can influence the decision about subsequent diagnostic tests, that is, if an invasive test (a biopsy) or a non-invasive test (ultrasound) should be performed (Eddy, 1982).

Not only physicians but also patients have to deal with statistical information that concerns their health status. Physicians are legally obliged to obtain consent from the patient for treatment, except in the rare cases in which the patient is not mentally capable to make decisions or there is extreme time pressure. This guideline follows directly from the constitutional right of individual autonomy (Bundesärztekammer, 1990). The respect for individual autonomy is not only legally prescribed, but is also included in the three basic principles of biomedical ethics (the other two principles being “do no harm” and “do good”; Marshall, 1996). Both legal and ethical principles imply that not only consent, but “informed consent” should be obtained: Ideally, patients should be informed about both benefits and risks of a treatment and its alternatives before the decision to enter treatment is made (Bundesärztekammer, 1990; Doyal, 2001; General Medical Council, 1998; Gigerenzer, 2002; Marshall, 1996; Ubel & Loewenstein, 1997). Only when patients have sufficient and understandable information about pros and cons can they evaluate their options according to their personal values and decide on the course of treatment that is best for them – either alone or together with the physician (Charles, Gafni, & Whelan, 1999; Coulter, 1997a).

Information about benefits and risks includes statistical information (e.g., the probability of a certain complication associated with an operation). Given the omnipresence of statistics in everyday life and their importance in medical practice, one could assume that medical professionals and medical laymen alike have learned to understand and use such statistical information without problems. Unfortunately, this is not the conclusion that can be drawn from the empirical evidence. For instance, medical professionals have repeatedly been shown to misjudge the positive predictive value of a diagnostic test, that is, the probability with which it can be predicted that a person has a disease, given a positive test result. In one study, medical professionals from Harvard Medical School were given the following text problem: “If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 percent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?” (Casscells, Schoenberger, & Graboys, 1978, p. 999). Only 11 of the 60 participating medical professionals inferred the correct positive predictive value that is 2%. Answers ranged from 0.095% to 99%, the most common answer was 95%. Similarly, in an informal sample taken by Eddy (1982), most physicians overestimated the positive predictive value of a mammogram by about one order of magnitude (75% instead of 8%). In another study, half of 20 AIDS counselors said that a positive HIV test for a low-risk client predicts an HIV infection with absolute certainty, that is, 100% (Gigerenzer, Hoffrage, & Ebert, 1998). However, the positive predictive value of HIV tests for low-risk clients can be as low as 50%.

Also patients have repeatedly been shown to have problems in understanding clinical information containing statistical information (e.g. Coulter, 1997b; Doyal, 2001; Lloyd, 2001; Schwartz, Woloshin, Black, & Welch, 1997; Weinstein, 1999). For example, in a sample of 56 patients who were counseled on their risk of having a stroke with or without a certain operation (the operation lowered the stroke risk, but operation-induced stroke could occur in rare cases as a complication), only one patient was able to recall the two risk figures he was told one month later. The risk estimates of the others were very variable and generally much too high; some had even forgotten that there was a stroke risk associated with the operation (Lloyd, Hayes, London, Bell, & Naylor, 1999). In another study, only 56% of 633 women were able to answer correctly the question of which is greater, a risk of 1 in 112 or 1 in 384 (Grimes & Snively, 1999; see also Yamagishi, 1997). Given such difficulties in obtaining a basic understanding of statistical information, some physicians ask themselves if they should not “recognize the utopian nature of the goal of properly informed consent and return to the more honest and realistic paternalism of the past” (Doyal, 2001, p. i29) – even if there actually is a broad legal and professional consensus about the clinical duty to obtain informed consent (Bundesärztekammer, 1990; Coulter, 1997a, 1997b; Doyal, 2001; General Medical Council, 1998; Marteau, 1995; Ubel & Loewenstein, 1997).

Do psychologists share the pessimism of medical professionals concerning statistical thinking in lay people? For a large part of the last thirty years, the answer was "yes". Take, for example, the diagnostic inference problem that was briefly sketched above. Here, the prior probability of a hypothesis (e.g. that a woman might have breast cancer) is supplemented with new evidence (e.g. a positive mammogram) and has to be updated accordingly. Psychologists have studied this type of problem, also called a "Bayesian" inference problem, since the 1960s (e.g. Edwards, 1968). Numerous experiments had revealed that not only lay people, but also experts have substantial trouble solving them, either overweighting the prior evidence ("conservatism", Edwards, 1968) or neglecting it ("base-rate neglect", e.g. Bar-Hillel, 1980; see review by Koehler, 1996a). By 1980 it was concluded that humans are not equipped to solve these problems and that therefore "the genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (Bar-Hillel, 1980, p. 215).

But in the 1990s, some authors offered a new interpretation of the results. Whereas the conclusion that humans simply cannot reason the Bayesian way was based on the assumption of processing errors *inherent* in our minds (e. g., Kahneman, Slovic, & Tversky, 1982), they made the ecological argument that people are indeed able to solve Bayesian inference problems when given an *external* representation of the data that facilitates rather than complicates human reasoning (Cosmides & Tooby, 1996; Gigerenzer and Hoffrage, 1995). Gigerenzer and Hoffrage (1995) reported that their participants performed much better when the statistical information was represented in a so-called "natural frequency format" (will be described in detail later in the text), rather than in other formats such as probabilities or percentages. The facilitating effect of natural frequencies on Bayesian inference problems has been replicated several times (Cosmides & Tooby, 1996; Hoffrage, Lindsey, Hertwig & Gigerenzer, 2000; Hoffrage & Gigerenzer, 1998; Hoffrage & Gigerenzer, in press; Koehler, 1996b). These findings allow for a much more positive appraisal of our ability to understand statistical information, also because frequency representations were shown to reduce or eliminate other well-known "cognitive illusions" such as the conjunction fallacy (Hertwig & Gigerenzer, 1999) or the overconfidence bias (Gigerenzer, Hoffrage, & Kleinbölting, 1991).

In the present dissertation, I would like to explore how natural frequencies can be used as a tool to improve statistical thinking in physicians and patients. Before I elaborate on this question, I first describe the tool in more detail.

What are natural frequencies?

Natural frequencies are the result of an information-sampling process called "natural sampling" (Gigerenzer & Hoffrage, 1995; Kleiter, 1994). Natural sampling is described as a sequential process of observing and counting events. To illustrate this process, think of a physician who learns from direct experience rather than from books with statistics. She observes, case by case, whether her patients have a disease and whether the outcome of a test is positive or negative; thus she counts individuals according to their features (e.g., disease versus no disease, positive test result vs. negative test result; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). For example, the physician could observe whether the women in her practice received a positive or negative result in a routine screening for breast cancer, and whether they have breast cancer or not. The result of such a natural sampling process can be displayed in a natural frequency tree (Figure 1.1).

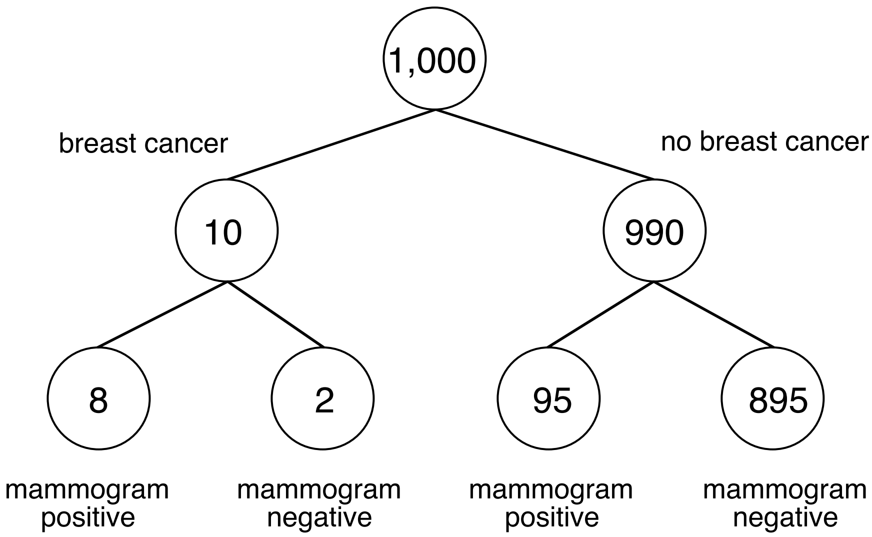

Figure 1.1 Natural frequency tree. The figure shows the sequential partitioning of one total sample into subsamples. The result of this sampling process are natural frequencies.

Natural frequencies are the result of the sequential partitioning of one total sample (here: 1,000 women) into subsamples. The base rates in the sample and the subsamples (e. g., the number of women who have breast cancer) correspond to the *natural* base rates as they could be obtained by observation of all cases or a representative draw. This is opposed to systematic sampling in scientific research, where often base rates are *artificially* fixed a priori to compare, for example, 100 people with disease to 100 without (Gigerenzer & Hoffrage, 1995; Kleiter, 1994). Please also note that an isolated number, such as 95, is not by itself a natural frequency; it only becomes one in its relation to other numbers in the tree. Without this relation, it is simply an absolute frequency.

Gigerenzer and Hoffrage (1995) introduced natural frequencies as an alternative way to represent statistical information in Bayesian inference problems. Prior to that, the information was typically represented as single-event probabilities or, as a variant, relative frequencies[1]. Table 1.1 shows the text of the same Bayesian inference problem for these three statistical formats (adapted from Gigerenzer & Hoffrage, 1995).

Table 1.1
Text of the mammography problem for three statistical formats

Format	Problem text – Standard menu [a]
Natural frequencies	Ten of every 1,000 women at age forty who participate in routine screening have breast cancer. Eight of every 10 women with breast cancer will have a positive mammogram. Ninety-five of every 990 women without breast cancer will also have a positive mammogram. Here is a new representative sample of women in this age group who had a positive mammogram in a routine screening. How many of these women do you expect actually to have breast cancer? ___ out of ___
Relative frequencies	One percent of women at age forty who participate in routine screening have breast cancer. Eighty percent of the women with breast cancer will have a positive mammogram. Of the women without breast cancer, 9.6% will also have a positive mammogram. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? ___%
Single-event probabilities	The probability of breast cancer is 1% for women at age forty who participate 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. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer? ____%

Note. [a] The problems are shown in the standard information menu, i.e. they display three pieces of information (Gigerenzer & Hoffrage, 1995). See text for further information.

The mammography problem displayed in Table 1.1 is a basic version of a Bayesian inference problem: The situation consists of a binary hypothesis H or $\neg H$ (here: breast cancer or no breast cancer; the "¬" stands for absence) and one binary cue D or $\neg D$ (D stands for data, here: test positive; $\neg D$ is a negative test result in this example). Three pieces of information are specified (this is also called a *standard information menu*; for other ways to segment the statistical information, see Gigerenzer & Hoffrage, 1995, and Chapter 3):

- The prior probability or base rate $p(\text{H})$, here the probability of having breast cancer.

[1] In the terminology of Gigerenzer and Hoffrage (1995), the term "relative frequencies" denotes percentages that refer to multiple events ("1% of *the women* ..."). The probability format can also contain percentages, as in the example above, but these refer to single cases ("The probability is 1% that *a woman* ...").

- The sensitivity of the data $p(\text{D}|\text{H})$, here the proportion of positive mammograms among women with breast cancer.
- The false-alarm rate of the test $p(\text{D}|\neg\text{H})$, here the proportion of positive mammograms among women without breast cancer.

The task is to find the posterior probability, or positive predictive value $p(\text{H}|\text{D})$, namely, the probability of the hypothesis, given the data: Of all the women with a positive mammogram, how many women do also have breast cancer? Or, for the other versions, respectively: What is the probability that a woman who tests positive actually has breast cancer?

The normatively correct solution to the mammography problem is 8 out of 103, or 7.7% (see below). Gigerenzer and Hoffrage (1995, Study 1) found that with natural frequencies, considerably more Bayesian inference problems were solved correctly (46%) than with single-event probabilities (18%). The performance rate with relative frequencies was comparable to the low performance with probabilities. The two latter results were consistent with previous findings on low performance in Bayesian inference problems (Bar-Hillel, 1980). The former result was new. What is the explanation for the advantage of natural frequencies over probabilities and relative frequencies? The main argument for the facilitative effect of natural frequencies is computational: Bayesian computations are simpler when the information is represented in natural frequencies rather than with probabilities or relative frequencies (Gigerenzer & Hoffrage, 1995; see also Kleiter, 1994). Equation 1 shows the computational steps for the natural frequency format.

$$p(H \mid D) = \frac{H \,\&\, D}{D} = \frac{H \,\&\, D}{H \,\&\, D + \neg H \,\&\, D} = \frac{8}{8 + 95} \qquad (1)$$

The number of *H & D* cases (i.e. breast cancer and positive mammogram) has to be divided by the total number of *D* cases (i.e. all positive mammograms). D consists of the number of *H & D* cases plus the number of *¬H & D* cases (i.e. no breast cancer and positive mammogram). In the mammography problem, *H & D* and *¬H & D* are 8 and 95, respectively. With natural frequencies, these two numbers can be derived directly from the problem text. Having identified these two numbers, only two simple computational steps remain: *H & D* and *¬H & D* have to be added to obtain the total number of D cases (103), and then *H & D* has to be divided by this sum (i.e. divide 8 by 103).

Computations are more complex for the probability and the relative frequency version. This is because probabilities and relative frequencies no longer contain information about the actual base rates, because each of the three pieces of information given in the text problem is normalized to 100%. For example, the sensitivity tells us that 80% of women with breast cancer receive a positive test result, but we cannot see from the 80% whether having breast cancer is a frequent or a rare event in the population (the sensitivity in terms of natural frequencies, i.e. "8 out of 10", includes this information). The benefit of normalization is that the resulting values fall within the uniform range of 0 and 1, and thus can be easily compared to each other (for this argument, it does not make a difference whether the probability is stated as a decimal in the interval [0, 1] or as a percentage in the interval [0, 100]). As quality criteria of diagnostic tests, information about the error rates of diagnostic tests in the medical literature is typically stated as percentages to facilitate comparisons of the quality of different diagnostic tests. However, if the task is not comparison of several tests, but the interpretation of one test as in a Bayesian inference, normalization does have a cost: The base rate information has to be put back in by multiplying the conditional probabilities by their respective base rates, which makes computation more complex.

$$p(H \mid D) = \frac{p(H)p(D \mid H)}{p(H)p(D \mid H) + p(\neg H)p(D \mid \neg H)} = \frac{(.01)(.80)}{(.01)(.80) + (.99)(.096)} \qquad (2)$$

Equation 2 is Bayes' rule for probabilities and percentages. Bayes' rule is named after the English reverend Thomas Bayes (1702-1761) who is credited with having discovered it (Stigler, 1983). The general idea is the same in both Equation 1 and 2, that is, the proportion of correct positives (numerator) is divided by all positives (denominator). The terms $p(\mathrm{H})p(\mathrm{D}|\mathrm{H})$ and $p(\neg\mathrm{H})p(\mathrm{D}|\neg\mathrm{H})$ correspond to *H & D* and *¬H & D* in Equation 1. The two additional computational steps in (2) result from the multiplication of conditional probabilities $p(\mathrm{D}|\mathrm{H})$ and $p(\mathrm{D}|\neg\mathrm{H})$ with the base rates $p(\mathrm{H})$ and $p(\neg\mathrm{H})$, respectively.

Computational simplification is one of two explanations that Gigerenzer and Hoffrage (1995) originally offered for the natural frequency effect. Note that the computational argument consists of two parts: "By 'computationally simpler', we mean that (a) fewer operations (multiplication, addition, or division) need to be performed in Equation 2 than in Equation 1 [here: in Equation 1 than in Equation 2], and (b) the operations can be performed on natural numbers (absolute frequencies) rather than fractions (such as percentages)." (Gigerenzer & Hoffrage, 1995, p.687).

The second explanation brings in an evolutionary perspective. It is argued that the human mind appears to be "tuned" to make inferences from natural frequencies rather than from probabilities and percentages: For most of their existence, humans and animals have made inferences from information encoded sequentially through direct experience, and natural frequencies can be seen as the final tally of such a process (hence the term "natural" frequencies; see Cosmides & Tooby, 1996; Kleiter, 1994). In contrast, mathematical probability did not emerge until the mid-17th century; in other words, probabilities and percentages are much more "recent" in evolutionary terms (Gigerenzer & Hoffrage, 1995). Therefore, it is assumed that minds have evolved to deal with natural frequencies rather than with probabilities.

Although the beneficial effect of natural frequencies on Bayesian reasoning was replicated several times for different groups of lay people and experts (Hoffrage & Gigerenzer, 1998; Hoffrage & Gigerenzer, in press; Hoffrage, Lindsey, et al., 2000), there is still considerable theoretical controversy in the literature about *why* this effect can be observed (see the discussions in Gigerenzer & Hoffrage, 1999; Hoffrage et al., 2002). Both explanations offered by Gigerenzer and Hoffrage (1995) have been heavily disputed, and especially the evolutionary argument has been met with skepticism (Fiedler, Brinkmann, Betsch, & Wild, 2000; Girotto & Gonzalez, 2001). It should be noted that, strictly speaking, the evolutionary argument has yet to be tested, because it is still not clear how the effects of the computational and evolutionary explanations can be disentangled (Hoffrage et al., 2002). Furthermore, many authors have argued that it is not the use of frequency formats per se, but rather some third factor that could be the explanation for the obtained results. Although this debate is highly interesting, I will address it only partially in this dissertation. The reason is that only part of it is relevant for the main question of the dissertation, namely, how natural frequencies can be used to improve statistical thinking in physicians and patients. I will not address the studies showing that natural frequencies are not the only tool to improve Bayesian reasoning (e.g., Evans, Handley, Perham, Over, & Thompson, 2000; Girotto & Gonzalez, 2001; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Macchi, 2000), because I do not want to argue here that natural frequencies are the only tool to improve medical risk communication. I will also not address studies that were based on an incorrect interpretation of natural frequencies (e.g., Evans et al., 2000; Johnson-Laird et al., 1999; Lewis & Keren, 1999; Macchi & Mosconi, 1998; Macchi, 2000), because most of the conclusions drawn there do not apply to natural frequencies as defined here (the misinterpretations have been addressed extensively in Gigerenzer & Hoffrage, 1999; Hoffrage et al., 2002). I will only address those studies that are relevant for predictions made in this dissertation (see Chapter 3).

Natural frequencies in the medical context

I said above that the main question of this dissertation is how natural frequencies can be used as a tool to improve statistical thinking in physicians and patients. The effect of natural frequencies on statistical thinking has been typically studied with text problems such as the mammography problem introduced above. I will subsequently refer to this method of studying Bayesian inferences as the “text problem paradigm”. Within the text problem paradigm, there have already been studies that explored if not only students, but also medical experts could benefit from natural frequencies when solving Bayesian inference tasks (Hoffrage & Gigerenzer 1998, in press). In these studies, 48 experienced physicians and 96 medical students were given four diagnostic inference problems. The results were basically the same as those described above: Performance was low with probability formats (18% medical students, 10% physicians), and clearly higher with natural frequency formats (57% medical students, 46% physicians). It can therefore be concluded that not only lay people, but also medical experts do benefit from the use of natural frequency formats in Bayesian inference tasks.

In their daily work, physicians will only rarely encounter written Bayesian inference problems. But they nevertheless make diagnostic inferences. As mentioned above, the information to base these inferences on (e.g., the sensitivity of a diagnostic test that is mentioned in a medical journal) is typically represented in terms of percentages or probabilities, and not in terms of natural frequencies. How can the insights on intuitive information representation be applied to this situation? One idea was to use natural frequencies in a tutorial on diagnostic inferences for medical students. The question tested in Chapter 2 was whether teaching medical students how to translate probabilities into natural frequencies would help them to make diagnostic inferences based on these probabilities.

Using natural frequencies to teach (future) medical experts to draw correct diagnostic inferences is one way of applying this tool in the medical context. A second application was explored in the following chapters of this dissertation: How can natural frequencies be used to educate medical lay people about the uncertainties and risks associated with diagnostic tests? To explore this question, I chose one specific example of a diagnostic test: the screening mammography. In a screening mammography, women who do not show any symptoms of breast cancer get an x-ray picture of their breasts (a mammogram). The goal is to detect breast cancer in its early stages to reduce mortality. As mentioned at the beginning of this chapter, to make a truly informed decision about participation in mammography screening, women have to understand the risks, benefits and the efficiency of this diagnostic test. In Chapter 4, I analyzed how currently available German mammography pamphlets inform women about these issues and identified a number of factors that could influence understanding. In Chapter 5, I tested if a mammography pamphlet that includes natural frequencies would be better understood than one that includes percentages, and what kind of information members of the main audience of mammography pamphlets actually request.

To summarize, out of the many potential applications of the tool of natural frequencies to the medical domain, in this dissertation I looked at two specific applications: One, using natural frequencies to teach medical students how to interpret diagnostic test results. And two, using natural frequencies to help medical lay people to understand statistical information given in health information pamphlets.