9

Bayesian Versus Frequentist Inference

Eric-Jan Wagenmakers[1], Michael Lee[2], Tom Lodewyckx[3], and Geoffrey J. Iverson[2]

[1] Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands ej.wagenmakers@gmail.com
[2] Department of Cognitive Sciences, University of California at Irvine, 3151 Social Science Plaza, Irvine CA 92697, USA mdlee@uci.edu and giverson@uci.edu
[3] Department of Quantitative and Personality Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium tom.lodewyckx@student.kuleuven.be

9.1 Goals and Outline

Throughout this book, the topic of order restricted inference is dealt with almost exclusively from a Bayesian perspective. Some readers may wonder why the other main school for statistical inference – *frequentist* inference – has received so little attention here. Isn't it true that in the field of psychology, almost all inference is frequentist inference?

The first goal of this chapter is to highlight why frequentist inference is a less-than-ideal method for statistical inference. The most fundamental limitation of standard frequentist inference is that it does not condition on the observed data. The resulting paradoxes have sparked a philosophical debate that statistical practitioners have conveniently ignored. What cannot be so easily ignored are the practical limitations of frequentist inference, such as its restriction to nested model comparisons.

The second goal of this chapter is to highlight the theoretical and practical advantages of a Bayesian analysis. From a theoretical perspective, Bayesian inference is principled and prescriptive and – in contrast to frequentist inference – a method that does condition on the observed data. From a practical perspective, Bayesian inference is becoming more and more attractive, mainly because of recent advances in computational methodology (e.g., Markov chain Monte Carlo and the WinBUGS program [95]). To illustrate, one of our frequentist colleagues had been working with the WinBUGS program and commented "I don't agree with the Bayesian philosophy, but the WinBUGS program does allow me to implement complicated models with surprisingly little effort." This response to Bayesian inference is diametrically opposed to the one that was in vogue until the 1980s, when statisticians often sympathized with the Bayesian philosophy but lacked the computational tools to implement models with a moderate degree of complexity.

The outline of this chapter is as follows: Section 9.2 introduces the Fisherian and the Neyman-Pearson flavors of frequentist inference and goes on to list a number of limitations associated with these procedures. Section 9.3 introduces Bayesian inference and goes on to lists a number of its advantages. Section 9.4 briefly presents our conclusions.

9.2 Frequentist Inference and Its Problems

Frequentist inference is based on the idea that probability is a limiting frequency. This means that a frequentist feels comfortable assigning probability to a repeatable event in which the uncertainty is due to randomness, such as getting a full house in poker (i.e., aleatory uncertainty [78]). When n hands are played and a full house is obtained in s cases, then, with n very large, the probability of a full house is just s/n. But a frequentist must refuse to assign probability to an event where uncertainty is also due to lack of knowledge, such as the event of Alexander Grischuk ever winning a major poker championship (i.e., epistemic uncertainty [34, 78]).

Because uncertainty about parameters is epistemic, frequentist inference does not allow probability statements about the parameters of a statistical process. For instance, the fact that a frequentist 95% confidence interval for the normal mean μ is $[-0.5, 1.0]$ does not mean that there is a 95% probability that μ is in $[-0.5, 1.0]$. Instead, what it means is that if the same procedure to construct confidence intervals was repeated very many times, for all kinds of different datasets, then in 95% of the cases would the true μ lie in the 95% confidence interval (cf. the example presented in Section 9.2.1).

Discussion of frequentist inference is complicated by the fact that current practice has become an unacknowledged amalgamation of the p-value approach advocated by Fisher [32] and the α-level approach advocated by Neyman and Pearson [76]. Hubbard and Bayarri [49, p. 176] summarized and contrasted the paradigms as follows:

> The level of significance shown by a p value in a Fisherian significance test refers to the probability of observing data this extreme (or more so) under a null hypothesis. This data-dependent p value plays an epistemic role by providing a measure of inductive evidence against H_0 in single experiments. This is very different from the significance level denoted by α in a Neyman-Pearson hypothesis test. With Neyman-Pearson, the focus is on minimizing type II, or β, errors (i.e., false acceptance of a null hypothesis) subject to a bound on type I, or α, errors (i.e., false rejections of a null hypothesis). Moreover, this error minimization applies only to long-run repeated sampling situations, not to individual experiments, and is a prescription for behaviors, not a means of collecting evidence.

Clearly then, Fisher's approach is very different from that of Neyman and Pearson. Yet, most researchers believe the paradigms have somehow merged and interpret the p-value both as a measure of evidence and as a repetitive error rate. It appears that the confusion between the two different procedures is now close to total, and it has been argued that this mass confusion "has rendered applications of classical statistical testing all but meaningless among applied researchers." [49, p. 171]. Additional references include [9, 18, 37, 38, 39, 40, 43, 91].

We now discuss several general problems of both the Fisherian and the Neyman-Pearson procedure (cf. [14, 26, 48, 53, 91, 97]). Although only the Neyman-Pearson procedure is truly frequentist (i.e., it requires knowledge about performance in long-run sampling situations), we will perpetuate the confusion and refer to both the Fisherian and the Neyman-Pearson procedure as "frequentist."

9.2.1 Frequentist Inference Generally Does Not Condition on the Observed Data

As argued by Berger and Wolpert [14], frequentist evidence is often pre-experimental or unconditional.[4] This means that "a particular procedure is decided upon for use, and the accuracy of the evidence from an experiment is identified with the long run behavior of the procedure, were the experiment repeatedly performed." [14, p. 5]. We illustrate the problem with this unconditional approach by an example that highlights the pathological properties of frequentist confidence intervals (cf. [15, p. 468]).

Consider a uniform distribution with mean μ and width 1. Draw two values randomly from this distribution, label the smallest one s and the largest one l, and check whether the mean μ lies in between s and l. If this procedure is repeated very many times, the mean μ will lie in between s and l in half of the cases. Thus, (s, l) gives a 50% frequentist confidence interval for μ. But suppose that for a particular draw, $s = 9.8$ and $l = 10.7$. The difference between these values is 0.9, and this covers 9/10th of the range of the distribution. Hence, for these particular values of s and l we can be 100% confident that $s < \mu < l$, even though the frequentist confidence interval would have you believe you should only be 50% confident.

This example shows why it is important to condition on the data that have actually been observed. The key problem is that frequentist methods do not do this, so that for data x, "(...) a procedure which looks great pre-experimentally could be terrible for particular x(...)" [14, p. 9]. Other examples of pathological behavior of frequentist confidence intervals can be found in [15, pp. 466–469], [14], and, in particular, [52].

[4] Frequentist' procedures sometimes do condition on important aspects of the data. Conditioning is always partial, however, and there exist situations in which it is unclear what aspects of the data on which to condition.

Table 9.1. Two different sampling distributions, $f(y)$ and $g(y)$ that lead to two different p-values for $y = 5$

	Data y					
Distribution	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$f(y)\|H_0$	.04	.30	.31	.31	.03	.01
$g(y)\|H_0$	.04	.30	.30	.30	.03	.03

9.2.2 Frequentist Inference Depends on Data That Were Never Observed

The p-value is the probability under the null hypothesis of observing data *at least as extreme* as the data that were actually observed. This means that the p-value is partly determined by data that were never observed, as is illustrated in the following example (cf. [4, 14, 19, 97]).

Assume the data y can take on six integer values, $y \in \{1, 2, ..., 6\}$, according to one of the sampling distributions $f(y)$ or $g(y)$. Further assume that what is observed is $y = 5$. As can be seen from Table 9.1, the observed datum is equally likely under $f(y)$ and $g(y)$. Yet, a one-sided p-value is $.03 + .01 = .04$ under $f(y)$ and $.03 + .03 = .06$ under $g(y)$. This is solely due to the fact that the more extreme observation $y = 6$, which was never observed, is less likely under $f(y)$ than it is under $g(y)$. Jeffreys famously summarized the situation: "*What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure" [55, p. 385, italics in original].

9.2.3 Frequentist Inference Depends on the Intention With Which the Data Were Collected

Because p-values are calculated over the sample space, changes in the sample space can greatly affect the p-value. For instance, assume that a participant answers a series of 17 test questions of equal difficulty; 13 answers are correct, 4 are incorrect, and the last question was answered incorrectly. Under the standard binomial sampling plan (i.e., "ask 17 questions"), the two-sided p-value is .049. The data are, however, also consistent with a negative binomial sampling plan (i.e., "keep on asking questions until the fourth error occurs"). Under this alternative sampling plan, the experiment could have been finished after four questions, or after a million. For this sampling plan, the p-value is .021.

What this simple example shows is that the intention of the researcher affects statistical inference – the data are consistent with both sampling plans, yet the p-value differs. Berger and Wolpert [14, pp. 30–33] discussed the result-

ing counterintuitive consequences through a story involving a naive scientist and a frequentist statistician.

In the story, a naive scientist has obtained 100 independent observations that are assumed to originate from a normal distribution with mean θ and standard deviation 1. In order to test the null hypothesis that $\theta = 0$, the scientist consults a frequentist statistician. The mean of the observations is 0.2, and hence the p-value is a little smaller than .05, which leads to a rejection of the null hypothesis. However, the statistician decides to probe deeper into the problem and asks the scientist what he would have done in the fictional case that the experiment had *not* yielded a significant result after 100 observations. The scientist replies that he would have collected another 100 observations. Thus, it may be hypothesized that the implicit sampling plan was not to collect 100 observation and stop; instead, the implicit sampling plan was to first take 100 observations and check whether $p < .05$. When the check is successful, the experiment stops, but when the check fails, another 100 observations are collected and added to the first 100, after which the experiment stops.

The statistician then succeeds in convincing the scientist that use of the implicit sampling plan requires a correction in order to keep the type I error rate at $\alpha = .05$ [81]. Unfortunately, this correction for planning multiple tests now leads to a p-value that is no longer significant. Therefore, the puzzled scientist is forced to continue the experiment and collect an additional 100 observations. Note that the interpretation of the data (i.e., significant or not significant) depends on what the scientist was planning to do in a situation that did not actually occur. If the very same data had been collected by a scientist who had answered the statistician's question by saying, whether truthfully or not, "I would not have collected any more observations," then the data would have been judged to be significant: Same data, different inference.

But the story becomes even more peculiar. Assume that the scientist collects the next 100 observations and sets up another meeting with the statistician. The data are now significant. The statistician, however, persists and asks what the scientist would have done in case the experiment had not yielded a significant result after 200 observations. Suppose that the scientist now answers "This would have depended on the status of my grant renewal. If my grant is renewed, I would have had enough funds to test another 100 observations. If my grant is not renewed, I would have had to stop the experiment. Not that this matters, of course, because the data were significant anyway."

The frequentist statistician then explains that the inference depends on the grant renewal; if the grant is not renewed, the sampling plan stands and no correction is necessary. But if the grant is renewed, the scientist could have collected more data, in the fictional case that the data would not have been significant after 200 observations. This calls for a correction for planning multiple tests, similar to the first one. Berger and Wolpert [14, p. 33] end their story: "The up-to-now honest scientist has had enough, and he sends in a request to have the grant renewal denied, vowing never again to tell the statistician what he could have done under alternative scenarios."

We believe that most researchers find it awkward that the conclusions from frequentist statistics depend critically on events that have yet to happen – events that, moreover, seem to be irrelevant with respect to the data that have actually been obtained.

9.2.4 Frequentist Inference Does Not Prescribe Which Estimator Is Best

Frequentist inference is not derived from a set of simple axioms that describe rational behavior. This means that any statistical problem potentially affords more than one frequentist solution, and it may be unclear which one is best. For instance, many different estimators may be proposed for a particular parameter θ. Which estimator for θ should we prefer? The common strategy is to narrow down the set of admissible estimators by considering only estimators that are *unbiased*. An estimator $t(\cdot)$ based on data y is unbiased when

$$\int_Y t(y)p(y|\theta)\ dy = \theta, \tag{9.1}$$

for all θ, where Y indicates the sample space (cf. [65]); that is, the only estimators taken into consideration are those that, averaged over the data that could arise, do not systematically overestimate or underestimate θ.

Although the criterion of unbiasedness has intuitive appeal, it is in fact highly contentious. First, the criterion is based on all possible datasets that could be observed (i.e., the sample space Y). This means that the intention of the researcher affects which estimators are unbiased and which are not. For instance, for the binomial sampling plan the unbiased estimator is s/n, where s is the number of correct responses out of a total of n questions, but for the negative binomial sampling plan the unbiased estimator is $(s-1)/(n-1)$. Second, an estimator that is unbiased for θ may well be biased for some nonlinear transformation of θ such as $\sqrt{\theta}$.

Finally, unbiased estimators may perform uniformly worse than biased estimators. Consider, for instance, the datum y distributed as $\mathcal{N}(\sqrt{\theta}, 1)$ with $\theta > 0$. The unbiased estimator for θ is $t(y) = y^2 - 1$. But when $|y| < 1$, $t(y)$ is negative, which conflicts with the knowledge that $\theta > 0$. A new estimator $t^{new}(y)$ may be proposed that is given by $t^{new}(y) = y^2 - 1$ when $|y| \geq 1$ and $t^{new}(y) = 0$ otherwise. The new estimator $t^{new}(y)$ is biased but does uniformly better than the unbiased estimator $t(y)$, this means that $t(y)$ is *inadmissible* (cf. [79]).

It should also be noted that in the above example, $t(y)$ is biased downward when $|y| < 1$, but biased upward when $|y| \geq 1$. Thus, an estimator may be unbiased for all possible datasets taken together, but it may – at the same time – be biased for every single dataset considered in isolation [79, p. 122].

Frequentist statisticians are aware of this problem, in the sense that they acknowledge that "(...) an overly rigid insistence upon unbiasedness may lead to difficulties" [96, p.432]. This statement highlights an important problem:

Frequentist inference does not specify a unique solution for every statistical problem. When unfortunate consequences of, say, "an overly rigid insistence upon unbiasedness" become apparent, adhoc estimators may be proposed that remedy the immediate problem – but this clearly is not a satisfactory state of affairs.

9.2.5 Frequentist Inference Does Not Quantify Statistical Evidence

According to the Fisherian tradition, p-values reflect the strength of evidence against the null hypothesis. General guidelines associate specific ranges of p-values with varying levels of evidence: A p-value greater than .1 yields "little or no real evidence against the null hypothesis," a p-value less than .1 but greater than .05 implies "suggestive evidence against the null hypothesis," a p-value less than .05 but greater than .01 yields "moderate evidence against the null hypothesis," and a p-value less than .01 constitutes "very strong evidence against the null hypothesis" [17, p. 9]; see also [101, p. 157].

If p-values truly reflect evidence, a minimum requirement is that equal p-values provide equal evidence against the null hypothesis (i.e., the p-postulate [97]). According to the p-postulate, $p = .05$ with 10 observations constitutes just as much evidence against the null hypothesis as does $p = .05$ after 50 observations.

It may not come as a surprise that Sir Ronald Fisher himself was of the opinion that the p-postulate is correct: "It is not true...that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability [the p-value], we thereby make full allowance for the size of the sample, and should be influenced in our judgement only by the value of probability indicated" [31, p. 182], as cited in [91, p. 70].

Nevertheless, some researchers believe that the p-postulate is false and that $p = .05$ after 50 observations is more reliable than $p = .05$ after 10 observations. For instance, Rosenthal and Gaito [84] found that the confidence with which a group of psychologists were willing to reject the null hypothesis increased with sample size (cf. [75]). Consistent with the psychologists' intuition, an article co-authored by 10 reputable statisticians stated that "a given p-value in a large trial is usually stronger evidence that the treatments really differ than the same p-value in a small trial of the same treatments would be" [80, p. 593], as cited in [91, p. 71].

Finally, several researchers have argued that when the p-values are the same, studies with small sample size actually provide *more* evidence against the null hypothesis than studies with large sample size (e.g., [1, 3, 69, 75]). For a summary of the debate, see [90]. Abelson considered the very question of whether a researcher would be happier with a $p = .05$ after testing 10 cases per group or after testing 50 cases per group, and then firmly concluded "Undergraduates inevitably give the wrong reply: Fifty cases per group, because a bigger sample is more reliable." The appropriate answer is "ten cases per

group, because if the p values are the same, the observed effect size has to be bigger with a smaller n" [1, p. 12].

In order to draw a firm conclusion about the veracity of the p-postulate, we first need to define what "evidence" is. The details of a rational (i.e., coherent or Bayesian) definition of evidence are presented in Section 9.3. Here it suffices to say that such an analysis must always reject the p-postulate (e.g., [26, 64, 98]): From a rational perspective, $p = .05$ after only 10 observations is more impressive than $p = .05$ after 1000 observations. In fact, it may happen that that for a large dataset, a frequentist analysis will suggest that the null hypothesis should be rejected, whereas a rational analysis will suggest that the null hypothesis is strongly supported.

9.2.6 Frequentist Inference Does Not Apply to Non-nested Models

Consider the study on Dissociative Identity Disorder (DID), introduced in Chapter 2 and discussed throughout this book. In this study, Huntjens et al. [50] set out to study memory processes in DID-patients. These patients often report *inter-identity amnesia* (i.e., impaired memory for events experienced by identities that are not currently present). For instance, the identity "lonely girl" may have limited or no knowledge of the events experienced by the identity "femme fatale." To test whether DID-patients were really affected by interidentity amnesia or whether they were simulating their amnesia, the authors assessed the performance of four groups of subjects on a multiple-choice recognition test. The dependent measure was the number of correct responses. The first group were the DID-patients, the second group were Controls, the third group were controls instructed to simulate interidentity amnesia (Simulators), and the fourth group were controls who had never seen the study list and were therefore True amnesiacs.

From the psychological theorizing that guided the design of the experiment, one can extract several hypotheses concerning the relative performance of the different groups. One hypothesis, H_{1a}, states that the mean recognition scores μ for DID-patients and True amnesiacs are the same and that their scores are higher than those of the Simulators: $\mu_{con} > \{\mu_{amn} = \mu_{pat}\} > \mu_{sim}$. Another hypothesis, H_{1b}, states that the mean recognition scores μ for DID-patients and Simulators are the same and that their scores are lower than those of the True amnesiacs: $\mu_{con} > \mu_{amn} > \{\mu_{pat} = \mu_{sim}\}$.

The hypotheses H_{1a} and H_{1b} are non-nested, and frequentist inference is not well suited for the comparison of such models (e.g., [60]). The main problem is that it is not clear whether H_{1a} or H_{1b} should be considered the null hypothesis. One might try both possibilities, but this runs the danger of simultaneously rejecting (or accepting) both H_{1a} and H_{1b}. Moreover, it is not clear how to interpret the hypothetical result of $p = .04$ when H_{1a} serves as the null hypothesis, and $p = .06$ when H_{1b} serves as the null hypothesis – even though H_{1a} is rejected and H_{1b} is not, this does not mean that H_{1b} is much better than H_{1a}.

9.2.7 Interim Conclusion

In the preceding analyses we have argued that frequentist procedures suffer from fundamental philosophical and practical problems. These problems are not some kind of well-kept secret, as statisticians have written about these frequentist flaws for many decades; the website `http://biology.uark.edu/coop/Courses/thompson5.html` documents some of their efforts by listing 402 articles and books that criticize the use of frequentist null hypothesis testing.

Indeed, the selection of problems mentioned in Sections 9.2.1 to 9.2.6 was certainly not exhaustive. Other problems include the fact that α-levels are arbitrary "surely, God loves the .06 nearly as much as the .05" [85, p. 1277], the fact that inference in sequential designs is overly complicated and conservative "Sequential analysis is a hoax" [2, p. 381], the fact that p-values are often misinterpreted, even by those teaching statistics [44], the fact that Fisherian frequentist inference does not allow one to obtain evidence in support of the null hypothesis, and the fact that frequentist inference is internally inconsistent or *incoherent*. The latter means that when statistical conclusions need to be backed up by betting on them, the frequentist will be a sure loser (for details, see Section 9.3.1).

All of this makes one may wonder why – despite the harsh criticism – the flogged horse of frequentist inference is still alive and well, at least in the field of psychology [1]. We believe the reason for this is most likely an unfortunate combination of several factors. Among the most important of these are ease of application, presumed lack of an appealing alternative, limited statistical knowledge among practitioners, faulty and one-sided teaching of statistics at universities, historical precedent, and – for a few special cases – exact numerical correspondence of frequentist "flogged horse" inference with rational inference to be discussed below.

This concludes our summary of frequentist inference and its problems. We now turn to a discussion of the other major statistical paradigms for statistical inference, which differs from frequentist inference in a few key assumptions. We will argue that, both philosophically and practically, this paradigm constitutes a superior alternative to frequentist inference.

9.3 Bayesian Inference and Its Advantages

In Bayesian inference, parameters are random variables. Uncertainty or degree of belief with respect to the parameters is quantified by probability distributions. For a given model, say H_1, the prior distribution $p(\theta|H_1)$ for a parameter θ is updated after encountering data y to yield a posterior distribution $p(\theta|y, H_1)$. The posterior information contains all of the relevant information about θ. Note that the posterior distribution is conditional on the data y that

have been observed; data that could have been observed, but were not, do not affect Bayesian inference.

Specifically, Bayes' rule states that the posterior distribution $p(\theta|y, H_1)$ is proportional to the product of the prior $p(\theta|H_1)$ and the likelihood $f(y|\theta, H_1)$:

$$p(\theta|y, H_1) = p(\theta|H_1)f(y|\theta, H_1)/m(y|H_1). \tag{9.2}$$

In this equation, $m(y|H_1)$ is the marginal probability of the data; it is computed by integrating out the model parameters using the law of total probability:

$$m(y|H_1) = \int p(y, \theta|H_1)\ d\theta = \int p(\theta|H_1)f(y|\theta, H_1)\ d\theta. \tag{9.3}$$

This shows that $m(y|H_1)$ can also be interpreted as a *weighted average likelihood* where the weights are provided by the prior distribution $p(\theta|H_1)$. Because $m(y|H_1)$ is a number that does not depend on θ, $m(y|H_1)$ can be conveniently ignored when the goal is to estimate θ. However, when the goal is Bayesian hypothesis testing, $m(y|H_1)$ becomes critically important.

For concreteness, consider the choice between two possibly non-nested models, H_1 and H_2. The extension to more than two models is entirely possible and follows the same recipe. Bayes' rule dictates how the prior probability of H_1, $p(H_1)$, is updated through the data to give the posterior probability of H_1, $p(H_1|y)$:

$$p(H_1|y) = p(H_1)m(y|H_1)/\sum_t p(H_t)m(y|H_t). \tag{9.4}$$

In the same way, one can calculate the posterior probability of H_2, $p(H_2|y)$. The ratio of these posterior probabilities is given by

$$\frac{p(H_1|y)}{p(H_2|y)} = \frac{p(H_1)}{p(H_2)}\frac{m(y|H_1)}{m(y|H_2)}, \tag{9.5}$$

which shows that the posterior odds $p(H_1|y)/p(H_2|y)$ is equal to the product of the prior odds $p(H_1)/p(H_2)$ and the ratio of marginal probabilities $m(y|H_1)/m(y|H_2)$. Thus, the ratio of marginal probabilities – henceforth referred to as the Bayes factor [55] – quantifies the change from prior to posterior odds brought about by the data. The Bayes factor, or the log of the Bayes factor, is often interpreted as the *weight of evidence* coming from the data [42]. Thus, a Bayes factor hypothesis test prefers the model under which the observed data are most likely. For details see [13], [15, Chapter 6], [41, Chapter 7], [56], and [77]; for an introduction in Bayesian inference, see Chapters 3 and 4 of the present book.

Jeffreys [55] proposed labeling the evidence provided by the Bayes factor according to a classification scheme that was subsequently revised by Raftery [82, Table 6]. Table 9.2 shows the Raftery classification scheme. The first column shows the Bayes factor, and the second column shows the associated

Table 9.2. Interpretation of the Bayes factor in terms of evidence

Bayes factor BF_{12}	$p(H_1\|y)$	Evidence
1–3	.50–.75	Weak
3–20	.75–.95	Positive
20–150	.95–.99	Strong
> 150	$> .99$	Very strong

posterior probability when it is assumed that both H_1 and H_2 are a priori equally plausible. The third column shows the verbal labels for the evidence at hand, in the case of a comparison between two models. Note that these verbal labels are associated with the level of evidence that is provided by the Bayes factor (i.e., a comparison between two models). These verbal labels should not be associated with posterior model probabilities (PMPs) when the set of candidate models is larger than two, for example, consider the problem of finding the best set of predictors for a regression equation. By considering all possible combinations of predictors, the model space can easily comprise as many as 100,000 candidate models. When a single model out of such a large set has a posterior probability of, say, .50, this would constitute a dramatic increase over its prior probability of .000001, and hence the data provide "very strong" rather than "weak" evidence in its favor.

Bayesian procedures of parameter estimation and hypothesis testing have many advantages over their frequentist counterparts. Below is a selective list of 10 specific advantages that the Bayesian paradigm affords.

9.3.1 Coherence

Bayesian inference is *prescriptive*; given the specification of a model, there exists only one way to obtain the appropriate answer. Bayesian inference does not require adhoc solutions to remedy procedures that yield internally inconsistent results. Bayesian inference is immune from such inconsistencies because it is founded on a small set of axioms for rational decision making. Several axiom systems have been proposed, but they all lead to the same conclusion: Reasoning under uncertainty can only be coherent if it obeys the laws of probability theory (e.g., [15, 20, 22, 23, 30, 53, 66, 68, 83, 92]).

One of the famous methods to prove this far-reaching conclusion is due to Bruno de Finetti and involves a betting scenario [22]. Assume there exists a legally binding ticket that guarantees to pay 1 euro should a proposition turn out to be true. For instance, the proposition could be "In 2010, the Dutch national soccer team will win the world cup." Now you have to determine the price you are willing to pay for this ticket. This price is the "operational subjective probability" that you assign to the proposition.

The complication is that this scenario also features an opponent. The opponent can decide, based on the price that you determined, to either buy this ticket from you or to make you buy the ticket from him. This is similar to the "I cut, you choose" rule where the person who cuts a cake gets to choose the last piece; it is then in that person's own interest to make a fair judgment.

In the example of the ticket, it is obviously irrational to set the price higher than 1 euro, because the opponent will make you buy this ticket from him and he is guaranteed to make a profit. It is also irrational to set the price lower than 0 euro, because the opponent will "buy" the ticket from you at a negative price (i.e., gaining money) and is again guaranteed to make a profit.

Now suppose you have to determine the price of three individual tickets. Ticket A states "In 2010, the Dutch national soccer team will win the world cup"; ticket B states "In 2010, the French national soccer team will win the world cup"; and ticket C states "In 2010, either the Dutch or the French national soccer team will win the world cup." You can set the prices any way you want. In particular, there is nothing to keep you from setting the prices such that price(ticket A) + price(ticket B) $\neq$ price(ticket C). However, when you set the prices this way, you are guaranteed to lose money compared to your opponent; for instance, suppose you set price(ticket A) = 0.5 euro, price(ticket B) = 0.3 euro, and price(ticket C) = 0.6 euro. Then the opponent will buy ticket C from you, sell you tickets A and B, and he is guaranteed to come out ahead. A set of wagers that ensures that somebody will make a profit, regardless of what happens, is called a *Dutch book*.

Using betting scenarios such as the above, de Finetti showed that the only way to determine subjective values and avoid a certain loss is to make these values obey the rules of probability theory (i.e., the rule that probabilities lie between 0 and 1, the rule that mutually exclusive events are additive, and the rule of conditional probability); that is, the only way to avoid a Dutch book is to make your prices for the separate tickets *cohere* according to the laws of probability calculus.

The concept of coherence refers not just to the betting scenario, but more generally to the combination of information in a way that is internally consistent. For example, consider Bayesian inference in the case that the data arrive in two batches, y_1 and y_2 [79, pp. 64-65]. Following the adage "today's posterior is tomorrow's prior" [65, p. 2], we can update from the initial prior $p(\theta)$ to a posterior $p(\theta|y_1)$ and then update this posterior again, effectively treating $p(\theta|y_1)$ as a prior, to finally obtain $p(\theta|y_1, y_2)$. The crucial aspect is that when the data are conditionally independent, it does not matter whether we observe the dataset batch-by-batch, all at once, or in reverse order.

As an additional example of coherence, consider a set of three models: H_1 postulates that $\mu_a = \mu_b = \mu_c$, H_2 postulates that $\{\mu_a = \mu_b\} > \mu_c$, and H_3 postulates that $\mu_a > \{\mu_b = \mu_c\}$. Then, denoting the Bayes factor for model H_i over model H_j by BF_{ij}, we can deduce from the identity

$$\frac{m(y|H_1)}{m(y|H_2)} = \frac{m(y|H_1)}{m(y|H_3)} \frac{m(y|H_3)}{m(y|H_2)}, \tag{9.6}$$

that $BF_{12} = BF_{32} \times BF_{13}$. This means that if the data are twice as likely under H_1 than under H_3 and thrice as likely under H_3 than under H_2, we know that the data are six times more likely under H_1 than under H_2. If the Bayes factors would not commute like this, one could construct a situation in which one could hold intransitive beliefs – a situation that would violate the axioms of rational decision making upon which Bayesian inference rests.

9.3.2 Automatic Parsimony

In statistical hypothesis testing, the ideal model captures all of the replicable structure and ignores all of the idiosyncratic noise. Such an ideal model yields the best predictions for unseen data coming from the same source. When a model is too complex, it is said to *overfit* the data; the model mistakenly treats idiosyncratic noise as if it were replicable structure. When a model is too simple, it is said to *underfit* the data, which means that the model fails to capture all of the replicable structure in the data. Models that underfit or overfit the data provide suboptimal predictions and are said to generalize poorly (e.g., [73, 100]).

The main challenge of hypothesis testing or model selection is to identify the model with the best predictive performance. However, it is not immediately obvious how this should be done; complex models will generally provide a better fit to the observed data than simple models, and therefore one cannot simply prefer the model with the best "goodness-of-fit" – such a strategy would lead to massive overfitting. Intuition suggests that this tendency for overfitting should be counteracted by putting a premium on simplicity. This intuition is consistent with the law of parsimony or "Occam's razor" (cf. `http://en.wikipedia.org/wiki/Occam's_Razor`), which states that when everything else is equal, simple models are to be preferred over complex models [53, Chapter 20].

Formal model selection methods try to quantify the trade-off between goodness-of-fit and parsimony. Many of these methods measure a model's overall performance by the sum of two components: One that measures descriptive accuracy and one that places a premium on parsimony. The latter component is also known as the Occam factor [70, Chapter 28]. For many model selection methods, the crucial issue is how to determine the Occam factor. One of the attractive features of Bayesian hypothesis testing is that it automatically determines the model with the best predictive performance – Bayesian hypothesis testing therefore incorporates what is known as an automatic Occam's razor. In order to see why this is the case, we explore two lines of reasoning.

First, recall that Bayesian model selection is based on the marginal probability of the data given model t, $m(y|H_t)$. Now denote a sequence of n data

points by $y^n = (y_1, ..., y_n)$; for example, y_{i-1} denotes the $(i-1)^{th}$ individual data point, whereas y^{i-1} denotes the entire sequence of observations ranging from y_1 up to and including y_{i-1}. Quantify predictive performance for a single data point by the logarithmic loss function $-\ln \widehat{p}_i(y_i)$: the larger the probability that $\widehat{p}_i$ (determined based on the previous observations y^{i-1}) assigns to the observed outcome y_i, the smaller the loss. From the definition of conditional probability (i.e., $p(y_i|y^{i-1}) = p(y^i)/p(y^{i-1})$), it then follows that the marginal probability of the data may be decomposed as a series of sequential, "one-step-ahead" probabilistic predictions (e.g., [21, 99]):

$$\begin{aligned} m(y^n|H_t) &= p(y_1, ..., y_n|H_t) \\ &= p(y_n|y^{n-1}, H_t)p(y_{n-1}|y^{n-2}, H_t)...p(y_2|y_1, H_t)p(y_1|H_t). \end{aligned} \quad (9.7)$$

Thus, (9.7) shows that the model with the highest marginal probability will also have the smallest sum of one-step-ahead prediction errors, as $-\ln m(y^n|H_t) = \sum_{i=1}^{n} -\ln p(y_i|y^{i-1}, H_t)$.

According to the second line of reasoning, every statistical model makes a priori predictions. Complex models have a relatively large parameter space and are therefore able to make many more predictions and cover many more eventualities than simple models. However, the drawback for complex models is that they need to spread out their prior probability across their entire parameter space. In the limit, a model that predicts almost everything has to spread out its prior probability so thinly that the occurrence of any particular event will not greatly add to that model's credibility. Formally, the marginal probability of the data is calculated by averaging the likelihood $f(y|\theta, H_t)$ over the prior $p(\theta|H_t)$. When the prior is very spread out, it will occupy a relatively large part of the parameter space in which the likelihood is almost zero, and this greatly decreases the average or marginal likelihood.

As a more concrete example, consider two people, Bart and Lisa, who each get 100 euros to bet on the winner of the 2010 world cup soccer. Bart decides to divide his money evenly over 10 candidate teams, including those from Brazil and Germany. Lisa divides her money over just two teams, betting 60 euros on the team from Brazil and 40 euros on the team from Germany. Now if either Brazil or Germany turn out to win the 2010 world cup, Lisa wins more money than Bart. By betting all her money on just two teams, Lisa was willing to take a risk, whereas Bart was just trying to keep his options open. For Bart, this means that even if his prediction of Brazil winning turns out to be correct, he will still lose the 90 euros he bet on the other countries to win. The point of the story is that, both at the betting office and in Bayesian inference, hedging your bets is not necessarily the best option, because this requires you to spread your resources – be it money or prior probability mass – thinly over the alternative options.

9.3.3 Extension to Non-nested Models

Bayesian hypothesis testing is based on the marginal probability of the data given model t, $m(y|H_t)$, and therefore it does not make a fundamental distinction between nested and non-nested models. This means that Bayesian hypothesis testing can be applied in many more situations than frequentist hypothesis testing. In cognitive psychology, for instance, important substantive questions concern the extent to which the law of practice follows a power function versus an exponential function or the extent to which category learning is best described by an exemplar model or a prototype model. For Bayesian inference, the substantive questions can be statistically tested in exactly the same way, whether the competing models are nested or not. For frequentist inference, however, the fact that the models are non-nested causes grave complications.

Another class of possibly non-nested models that are of great relevance for psychologists are those that incorporate order restrictions. For instance, consider again the case of the Huntjens et al. study on DID discussed in Section 9.2.6 and throughout this book. For the data from the study, hypothesis H_{1a} states that the mean recognition scores μ for DID-patients and True amnesiacs are the same and that their scores are higher than those of the Simulators: $\mu_{con} > \{\mu_{amn} = \mu_{pat}\} > \mu_{sim}$, whereas hypothesis H_{1b} states that the mean recognition scores μ for DID-patients and Simulators are the same and that their scores are lower than those of the True amnesiacs: $\mu_{con} > \mu_{amn} > \{\mu_{pat} = \mu_{sim}\}$. Within the frequentist paradigm, a comparison of these models is problematical. Within the Bayesian paradigm, however, the comparison is natural and elegant (e.g., [35, 47, 59, 60, 61, 62, 94]).

The general recipe, outlined in O'Hagan and Forster [79, pp. 70-71] is to carry out order restricted inference by first considering the posterior distribution of the unconstrained model and then restricting one's attention to the part of the posterior distribution that obeys the parameter constraints. In a Markov chain Monte Carlo (MCMC) simulation, for instance, this can be accomplished automatically by retaining only those samples that are in line with the constraints. The work reported in this book attests to the ability of Bayesian inference to address substantive psychological questions that involve order restrictions in a manner that is unattainable by frequentist means.

9.3.4 Flexibility

Bayesian inference allows for the flexible implementation of relatively complicated statistical techniques such as those that involve hierarchical nonlinear models (e.g., [71, 72, 74, 86, 87, 88, 89]). In hierarchical models, parameters for individual people are assumed to be drawn from a group-level distribution. Such multilevel structures naturally incorporate both the differences and the commonalities between people and therefore provide experimental psychology

with the means to settle the age-old problem of how to deal with individual differences.

Historically, the field of experimental psychology has tried to ignore individual differences, pretending instead that each new participant is a replicate of the previous one [6]. As Bill Estes and others have shown, however, individual differences that are ignored can lead to averaging artifacts in which the data that are averaged over participants are no longer representative for any of the participants (e.g., [28, 29, 45]). One way to address this issue, popular in psychophysics, is to measure each individual participant extensively and deal with the data on a participant-by-participant basis.

In between the two extremes of assuming that participants are completely the same and that they are completely different lies the compromise of hierarchical modeling (cf. [63]). The theoretical advantages and practical relevance of a Bayesian hierarchical analysis for common experimental designs has been repeatedly demonstrated by Jeff Rouder and colleagues (e.g., [86, 88, 89]). Although hierarchical analyses can be carried out using orthodox methodology [46], there are strong philosophical and practical reasons to prefer the Bayesian methodology (e.g., [36, 68]).

9.3.5 Marginalization

Bayesian statistics makes it easy to focus on the relevant variables by integrating out so-called nuisance variables (e.g., [5, 12]). Consider, for instance, the case of the normal distribution, for which the likelihood function is given by

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{9.8}$$

For this example, we follow [70, Chapter 24] and propose *conjugate improper priors* for μ and σ. A prior is said to be *conjugate* when it is in the same distributional family as the posterior distribution. For instance, when the prior for μ is normal, the posterior for μ is also normal. Conjugate priors are often the only ones that allow analytical derivation of the posterior. A prior is said to be *improper* when it does not integrate to a finite number. For instance, when the prior for μ is a normal distribution with mean $\mu_0 = 0$ and standard deviation $\sigma_\mu \to \infty$, this yields a prior that is flat across the entire real line. For the present example, we use conjugate improper priors on μ and σ because they lead to elegant analytical results that correspond to results from frequentist inference.

In particular, we assume here that the prior on μ is normal with mean $\mu_0 = 0$ and standard deviation $\sigma_\mu \to \infty$. This flat prior simply states that all values of μ are equally likely a priori. Because σ is always greater than 0, but $\log \sigma$ covers the entire real line, a standard "uninformative" prior is flat on the log scale, which transforms to the prior $p(\sigma) = 1/\sigma$. Using these priors, one can analytically derive the joint posterior distribution of μ and σ given the data (i.e., $p(\mu, \sigma|y)$) (e.g., [70, Chapter 24]).

Now that we have defined the priors and know the joint posterior distribution of μ and σ, consider two scenarios in which one needs to eliminate a nuisance parameter. In the first scenario, we want to learn about the mean μ of a normal distribution with unknown standard deviation σ. Thus, μ is the parameter of interest, whereas σ is a parameter that one would like to ignore (i.e., a nuisance parameter).

Using the law of total probability, it is straightforward to marginalize over, or integrate out, σ, as $p(\mu|y) = \int p(\mu, \sigma|y)\ d\sigma$. The fact that this equation can be rewritten as $p(\mu|y) = \int p(\mu|\sigma, y)p(\sigma)\ d\sigma$ highlights the fact that the nuisance parameter σ can only be integrated out once it has been assigned a prior distribution. After integrating out σ, the resulting posterior marginal distribution for $p(\mu|y)$ turns out to be the Student-t distribution, the famous frequentist distribution for a test statistic that involves the mean of a normal distribution with unknown variance [54].

In the second situation, we want to learn about the standard deviation σ of a normal distribution with unknown mean μ. This means that σ is the parameter of interest, whereas μ is now the nuisance parameter. From the joint posterior distribution of μ and σ, we can again apply the law of total probability, this time to integrate out μ, as follows: $p(\sigma|y) = \int p(\sigma, \mu|y)\ d\mu = \int p(\sigma|\mu, y)p(\mu)\ d\mu$. As before, this equation shows that the nuisance parameter μ can only be integrated out when it has been assigned a prior distribution. After computing the marginal posterior distribution $p(\sigma|y)$, the Most Probable value for σ (given the data y) turns out to be $\sigma_{MP} = \sqrt{S^2/(n-1)}$, where n equals the number of observations and $S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$. The factor $n-1$ (instead of n) also occurs in frequentist inference, where $S^2/(n-1)$ is the unbiased estimator for the variance of a normal distribution with unknown mean.

In sum, Bayesian inference allows the user to focus on parameters of interest by integrating out nuisance parameters according to the law of total probability. The resulting marginal posterior distributions may have matching frequentist counterparts, but this only holds in a few special cases.

9.3.6 Validity

Bayesian inference yields results that connect closely to what researchers want to know. To clarify this claim by analogy, Gerd Gigerenzer has suggested that for many researchers statistical inference involves an internal Freudian struggle amond the Superego, the Ego, and the Id (e.g., [37, 39]). In Gigerenzer's analogy, the Superego promotes Neyman-Pearson hypothesis testing, in which an α-level is determined in advance of the experiment. The Ego promotes Fisherian hypothesis testing, in which the precise value of p supposedly measures the strength evidence against the null hypothesis. Finally, the Id desires that the hypotheses under consideration are assigned probabilities, something that the Superego and Ego are unable and unwilling to do. As a result of this unconscious internal conflict, researchers often report results from frequentist

procedures, but often believe – implicitly or even explicitly – that they have learned something about the probability of the hypotheses under consideration.

We agree with Gigerenzer that, deep down inside, what researchers really want is to draw Bayesian conclusions. Or, in the words of Dennis Lindley, "Inside every Non-Bayesian, there is a Bayesian struggling to get out," as cited in [53]. This assertion is supported by the fact that researchers often misinterpret frequentist concepts – and misinterpret them in a manner that is decidedly Bayesian (i.e., the interpretation would have been correct if the method of inference had been Bayesian) [44].

To illustrate the foregoing with a concrete example, consider a frequentist confidence interval for the normal mean μ: $\mu \in [-0.5, 1.0]$. As we have seen in Section 9.2.1, the correct but counterintuitive interpretation of this result is that when the frequentist procedure is applied very many times to all kinds of possible datasets, the different intervals cover the true value of μ in 95% of the cases. But why would this be relevant for the researcher who wants to learn about μ for his or her data? In contrast, consider the same $[-0.5, 1.0]$ interval for μ, but now assume it is a Bayesian 95% credible interval. Consistent with intuition and consistent with what researchers want to know, this Bayesian interval conveys that there is a .95 probability that μ lies in $[-0.5, 1.0]$. From the viewpoint of "operation subjective probability" discussed in Section 9.3.1, this confidence interval means that when a coherent researcher is asked to set a fair price for a ticket that promises to pay 1 euro should the assertion "μ is in $[-0.5, 1.0]$" turn out to be true, that researcher will set the price of the ticket at exactly 0.95 euro.

9.3.7 Subjectivity That Is Open to Inspection

A common objection to Bayesian inference is that it is subjective and therefore has no place in scientific communication. For instance, in an article entitled "Why Isn't Everyone a Bayesian?" Bradly Efron argued that "Strict objectivity is one of the crucial factors separating scientific thinking from wishful thinking" and concluded that "The high ground of scientific objectivity has been seized by the frequentists" [27, p. 4].

Efron's claims need to be amended for several reasons. First, from a subjective Bayesian perspective, there is no such thing as "strict objectivity," as reasoning under uncertainty is always relative to some sort of background knowledge. In this view, the search for "strict objectivity" is a quixotic ideal. Thus, subjective Bayesians might want to change Efron's claim to "The high ground of scientific objectivity is a concept that cannot be seized by anyone, because it does not exist."

Second, there exists a school of *objective* Bayesians, who specify priors according to certain predetermined rules [58]. Given a specific rule, the outcome of statistical inference is independent of the person who performs the analysis.

Examples of objective priors include the unit information priors, that is, priors that carry as much information as a single observation [57], priors that are invariant under transformations [55], and priors that maximize entropy [51]. Objective priors are generally vague or uninformative (i.e., thinly spread out over the range for which they are defined). Thus, objective Bayesians might want to change Efron's claim to "Although the high ground of scientific objectivity may *appear* to be seized by the frequentists, objective Bayesians have a legitimate claim to scientific objectivity also."

Third, frequentist inference is not as objective as one may (wishfully) think. As illustrated in Section 9.2.3, the intention with which an experiment is carried out can have a profound impact on frequentist inference. The undisclosed ideas and thoughts that guided experimentation are crucial for calculating frequentist measures of evidence. Berger and Berry concluded that the perceived objectivity of frequentist inference is largely illusionary [11]. Thus, critics of frequentist inference might want to change Efron's claim to "Although the high ground of scientific objectivity may *appear* to be seized by the frequentists, upon closer inspection this objectivity is only make-believe, as in reality frequentists have to rely on the honesty and introspective ability of the researchers who collected the data."

In contrast to frequentist inference, Bayesian inference generally does not depend on subjective intentions (cf. Section 9.2.3), or on data that were never observed (cf. Section 9.2.2) [67]. The posterior distribution of parameters θ is written $p(\theta|y)$, and the marginal probability of a model, say H_0, is given by $m(y|H_0)$; in both cases, y is the observed data, and it is irrelevant what other data could have been observed but were not.

In Bayesian inference, the subjectivity that Efron alluded to comes in through the specification of the prior distribution for the model parameters. Regardless of whether this specification occurs automatically, as in the case of objective priors, or whether it occurs through the incorporation of prior knowledge, as in the case of subjective priors, the crucial point is that the prior distribution is formally specified and available for all other researchers to inspect and criticize. This also means that Bayesian subjectivity can be analyzed by formal methods that quantify robustness to the prior (e.g., [8, 25]). Note how different the notion of subjectivity is for the two paradigms: Bayesian subjectivity is open to inspection, whereas frequentist subjectivity is hidden from view, carefully locked up in the minds of the researchers that collected the data. Therefore, a final adjustment of Efron's statement might read "Scientific objectivity is illusionary, and both Bayesian inference and frequentist inference have subjective elements; the difference is that Bayesian subjectivity is open to inspection, whereas frequentist subjectivity is not."

9.3.8 Possibility of Collecting Evidence in Favor of the Null Hypothesis

Bayesian hypothesis testing allows one to obtain evidence in favor of the null hypothesis. In the Fisherian paradigm, p-values can only be used to reject the null hypothesis. The APA task force on statistical inference stressed this point by issuing the warning "Never use the unfortunate expression *accept the null hypothesis*" [102, p. 599]. Of course, what is unfortunate here is not so much the expression, but rather the fact that Fisherian p-values are incapable of providing support for the null hypothesis. This limitation hinders scientific progress, because theories and models often predict the absence of a difference. In the field of visual word recognition, for instance, the entry-opening theory [33] predicts that masked priming is absent for items that do not have a lexical representation. Another example from that literature concerns the work by Bowers et al. [16], who have argued that priming effects are equally large for words that look the same in lowercase and uppercase (e.g., kiss/KISS) or that look different (e.g., edge/EDGE), a finding supportive of the hypothesis that priming depends on abstract letter identities.

A final example comes from the field of recognition memory, where Dennis and Humphreys' "bind cue decide model of episodic memory" (BCDMEM) predicts the absence of a list-length effect and the absence of a list-strength effect [24]. This radical prediction of a null effect allows researchers to distinguish between context-noise and item-noise theories of inference in memory. Within the Fisherian paradigm, support for such informative predictions can only be indirect.

In contrast to the Fisherian hypothesis test, the Bayesian hypothesis test quantifies evidence by comparing the marginal probability of the data given one hypothesis, say $m(y|H_A)$, to the the marginal probability of the data given another hypothesis, say $m(y|H_B)$. The null hypothesis has no special status in Bayesian inference, and evidence for it is quantified just as it is for any other hypothesis, in a way that automatically strikes a balance between goodness-of-fit and parsimony (cf. Section 9.3.2).

9.3.9 Opportunity to Monitor Evidence as It Accumulates

Bayesian hypothesis testing allows one to monitor the evidence as the data come in [10]. In contrast to frequentist inference, Bayesian inference does not require special corrections for "optional stopping" [97].

Consider, for instance, a hypothetical experiment on the neural substrate of dissociative identity disorder. In this experiment, the researcher Marge has decided in advance to use functional magnetic resonance imaging (fMRI) to test 30 patients and 90 normal controls in a total of 4 between-subjects conditions, using the same design as Huntjens et al. [50]. Marge inspects the data after 15 participants in each condition have been tested and finds that the

results quite convincingly demonstrate the pattern she hoped to find. Unfortunately for Marge, she cannot stop the experiment and claim a significant result, as she would be changing the sampling plan halfway through and be guilty of "optional stopping." She has to continue the experiment, wasting not just her time and money, but also the time and efforts of the people who undergo needless testing.

Within the frequentist paradigm, it is possible to adopt special sampling plans that take into account the need or desire to monitor the data as they accumulate; however, these sampling plans yield conclusions that are much more conservative than the one that assumes a fixed sample size. Thus, the very same data may lead to a clearly significant result under a fixed sample size scheme, but to a clearly nonsignificant result under a variable sample size scheme; the difference is due to the fact that the variable sample size scheme incorporates a correction for the eventuality that the experiment could have ended at a different point in time than it actually did.

In contrast, for Bayesian hypothesis testing there is nothing wrong with gathering more data, examining these data, and then deciding whether or not to stop collecting new data – no special corrections are needed. As stated by Edwards et al., "(...) the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." [26, p. 193].

9.3.10 Possibility of Incorporating Prior Knowledge

Bayesian inference allows prior knowledge to influence conclusions [67]. Priors are not only tremendously useful for incorporating existing knowledge, they are also a prerequisite for rational inference: "If one fails to specify the prior information, a problem of inference is just as ill-posed as if one had failed to specify the data." [53, p. 373]. Another perspective on priors was put forward by Berger, who argued that "(...) when different reasonable priors yield substantially different answers, can it be right to state that there *is* a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?" [7, p. 125]. Thus, rather than considering priors a nuisance, we believe they are useful [67], necessary [53], and informative with respect to the robustness of one's conclusions [7]. Priors are an integral part of rational inference; one can only enjoy the Bayesian omelet when one is prepared to break the Bayesian eggs [93, p. 578].

9.4 Concluding Comments

In experimental psychology, the dominance of frequentist inference is almost complete. The first goal of this chapter was to demonstrate that the frequentist framework, despite its popularity, has several serious deficiencies. The second

goal of this chapter was to show how the Bayesian framework is both flexible and principled. Our conclusion is that the field of psychology can gain a lot by moving toward the Bayesian framework for statistical inference and by moving away from the frequentist framework.

Perhaps frequentist inference has survived for so long because researchers translate the frequentist statistical outcomes to informal Bayesian conclusions. For instance, most experienced experimental psychologists would take seriously a priming effect of 25 msec ($p = .03$, $N = 30$ subjects, $k = 20$ items per condition), whereas they would be skeptical of a priming effect of 4 msec ($p = .03$, $N = 257$ subjects, $k = 20$ items per condition). Such an informal Bayesian interpretation of frequentist results is another indication of the internal conflict between the frequentist Superego and Ego versus the Bayesian Id; see Section 9.3.6 and [37].

It is our hope that more and more psychologists will start to move away from frequentist inference and turn instead to formal Bayesian inference. It may take therapy, medication, or perhaps even surgery, but in the end, researchers will be happier people once they allow their inner Bayesian to come out.

References

[1] Abelson, R.P.: On the surprising longevity of flogged horses: Why there is a case for the significance test. Psychological Science, **8**, 12–15 (1997)

[2] Anscombe, F.J.: Sequential medical trials. Journal of the American Statistical Association, **58**, 365–383 (1963)

[3] Bakan, D.: The test of significance in psychological research. Psychological Bulletin, **66**, 423–437 (1966)

[4] Barnard, G.A.: The meaning of a significance level. Biometrika, **34**, 179–182 (1947)

[5] Basu, D.: On the elimination of nuisance parameters. Journal of the American Statistical Association, **72**, 355–366 (1977)

[6] Batchelder, W.H.: Cognitive psychometrics: Combining two psychological traditions. CSCA Lecture, Amsterdam, The Netherlands, October 2007.

[7] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis (2nd ed.). New York, Springer (1985)

[8] Berger, J.O.: Robust Bayesian analysis: Sensitivity to the prior. Journal of Statistical Planning and Inference, **25**, 303–328 (1990)

[9] Berger, J.O.: Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science, **18**, 1–32 (2003)

[10] Berger, J.O., Berry, D.A.: The relevance of stopping rules in statistical inference. In: Gupta, S.S., Berger, J.O. (eds) Statistical Decision Theory and Related Topics: Vol. 1. New York, Springer (1988)

[11] Berger, J.O., Berry, D.A.: Statistical analysis and the illusion of objectivity. American Scientist, **76**, 159–165 (1988)

[12] Berger, J.O., Liseo, B., Wolpert, R.L.: Integrated likelihood methods for eliminating nuisance parameters. Statistical Science, **14**, 1–28 (1999)

[13] Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association, **91**, 109–122 (1996)
[14] Berger, J.O., Wolpert, R.L.: The Likelihood Principle. Institute of Mathematical Statistics (2nd ed.), Hayward, CA (1988)
[15] Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. New York, Wiley (1994)
[16] Bowers, J.S., Vigliocco, G., Haan, R.: Orthographic, phonological, and articulatory contributions to masked letter and word priming. Journal of Experimental Psychology: Human Perception and Performance, **24**, 1705–1719 (1998)
[17] Burdette, W.J., Gehan, E.A.: Planning and Analysis of Clinical Studies. Charles C. Springfield, IL, Thomas (1970)
[18] Christensen, R.: Testing Fisher, Neyman, Pearson, and Bayes. The American Statistician, **59**, 121–126 (2005)
[19] Cox, D.R.: Some problems connected with statistical inference. The Annals of Mathematical Statistics, **29**, 357–372 (1958)
[20] Cox, R.T.: Probability, frequency and reasonable expectation. American Journal of Physics, **14**, 1–13 (1946)
[21] Dawid, A.P.: Statistical theory: The prequential approach. Journal of the Royal Statistical Society, Series A, **147**, 278–292 (1984)
[22] De Finetti, B.: Theory of Probability, Vols. 1 and 2. New York, Wiley (1974)
[23] DeGroot, M.-H.: Optimal Statistical Decisions. New York, McGraw-Hill (1970)
[24] Dennis, S., Humphreys, M.S.: A context noise model of episodic word recognition. Psychological Review, **108**, 452–477 (2001)
[25] Dickey, J.M.: Scientific reporting and personal probabilities: Student's hypothesis. Journal of the Royal Statistical Society, Series B, **35**, 285–305 (1973)
[26] Edwards, W., Lindman, H., Savage, L.J.: Bayesian statistical inference for psychological research. Psychological Review, **70**, 193–242 (1963)
[27] Efron, B.: Why isn't everyone a Bayesian? The American Statistician, **40**, 1–5 (1986)
[28] Estes, W.K.: The problem of inference from curves based on group data. Psychological Bulletin, **53**, 134–140 (1956)
[29] Estes, W.K.: Traps in the route to models of memory and decision. Psychonomic Bulletin & Review, **9**, 3–25 (2002)
[30] Fishburn, P.C.: The axioms of subjective probability. Statistical Science, **1**, 335–345 (1986)
[31] Fisher, R.A.: Statistical Methods for Research Workers (5th ed.). London, Oliver and Boyd (1934)
[32] Fisher, R.A.: Statistical Methods for Research Workers (13th ed.). New York, Hafner (1958)
[33] Forster, K.I., Mohan, K., Hector, J.: The mechanics of masked priming. In: Kinoshita, S., Lupker, S.J. (eds) Masked Priming: The State of the Art. New York, Psychology Press (2003)
[34] Galavotti, M.C.: A Philosophical Introduction to Probability. Stanford, CA, CSLI Publications (2005)

[35] Gelfand, A.E., Smith, A.F.M., Lee, T. M.: Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. Journal of the American Statistical Association, **87**, 523–532 (1992)

[36] Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, Cambridge University Press (2007)

[37] Gigerenzer, G.: The superego, the ego, and the id in statistical reasoning. In: Keren, G., Lewis, C. (eds) A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Hillsdale, NJ, Erlbaum (1993)

[38] Gigerenzer, G.: We need statistical thinking, not statistical rituals. Behavioral and Brain Sciences, **21**, 199–200 (1998)

[39] Gigerenzer, G.: Mindless statistics. The Journal of Socio–Economics, **33**, 587–606 (2004)

[40] Gigerenzer, G., Krauss, S., Vitouch, O.: The null ritual: What you always wanted to know about significance testing but were afraid to ask. In: Kaplan, D. (ed) The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks, CA, Sage (2004)

[41] Gill, J.: Bayesian Methods: A Social and Behavioral Sciences Approach. Boca Raton, FL, CRC Press (2002).

[42] Good, I.J.: Weight of evidence: A brief survey. In: Bernardo, J.M., DeGroot, M.-H., Lindley, D.V., Smith, A.F.M. (eds) Bayesian Statistics 2. New York, Elsevier (1985)

[43] Goodman, S.N.: P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. American Journal of Epidemiology, **137**, 485–496 (1993)

[44] Haller, H., Krauss, S.: Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research, **7**, 1–20 (2002)

[45] Heathcote, A., Brown, S., Mewhort, D.J.K.: The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin & Review, **7**, 185–207 (2000)

[46] Hoffman, L., Rovine, M.J.: Multilevel models for the experimental psychologist: Foundations and illustrative examples. Behavior Research Methods, **39**, 101–117 (2007)

[47] Hoijtink, H.: Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. Multivariate Behavioral Research, **36**, 563–588 (2001)

[48] Howson, C., Urbach, P.: Scientific Reasoning: The Bayesian Approach (3rd ed.). Chicago, Open Court (2006)

[49] Hubbard, R., Bayarri, M.J.: Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. The American Statistician, **57**, 171–182 (2003)

[50] Huntjens, R.J.C., Peters, M.L., Woertman, L., Bovenschen, L.M., Martin, R.C., Postma, A.: Inter-identity amnesia in dissociative identity disorder: A simulated memory impairment? Psychological Medicine, **36**, 857–863 (2006)

[51] Jaynes, E.T.: Prior probabilities. IEEE Transactions on Systems Science and Cybernetics, **4**, 227–241 (1968)

[52] Jaynes, E.T.: Confidence intervals vs Bayesian intervals. In: Harper, W.L., Hooker, C.A. (eds) Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. 2. Dordrecht, Reidel (1976)

[53] Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge, Cambridge University Press (2003)
[54] Jeffreys, H.: On the relation between direct and inverse methods in statistics. Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences, **160**, 325–348 (1937)
[55] Jeffreys, H.: Theory of Probability. Oxford, Oxford University Press (1961)
[56] Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association, **90**, 377–395 (1995)
[57] Kass, R.E., Wasserman, L.: A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, **90**, 928–934 (1995)
[58] Kass, R.E., Wasserman, L.: The selection of prior distributions by formal rules. Journal of the American Statistical Association, **91**, 1343–1370 (1996)
[59] Klugkist, I., Kato, B., Hoijtink, H.: Bayesian model selection using encompassing priors. Statistica Neerlandica, **59**, 57–69 (2005)
[60] Klugkist, I., Laudy, O., Hoijtink, H.: Inequality constrained analysis of variance: A Bayesian approach. Psychological Methods, **10**, 477–493 (2005)
[61] Klugkist, I., Laudy, O., Hoijtink, H.: Bayesian eggs and Bayesian omelettes: Reply to Stern (2005). Psychological Methods, **10**, 500–503 (2005)
[62] Laudy, O., Zoccolillo, M., Baillargeon, R.H., Boom, J., Tremblay, R.E., Hoijtink, H.: Applications of confirmatory latent class analysis in developmental psychology. European Journal of Developmental Psychology, **2**, 1–15 (2005)
[63] Lee, M.D., Webb, M.R.: Modeling individual differences in cognition. Psychonomic Bulletin & Review, **12**, 605–621 (2005)
[64] Lindley, D. V.: A statistical paradox. Biometrika, **44**, 187–192 (1957)
[65] Lindley, D.V.: Bayesian Statistics, a Review. Philadelphia, PA, SIAM (1972)
[66] Lindley, D.V.: Scoring rules and the inevitability of probability. International Statistical Review, **50**, 1–26 (1982)
[67] Lindley, D.V.: The analysis of experimental data: The appreciation of tea and wine. Teaching Statistics, **15**, 22–25 (1993)
[68] Lindley, D.V.: The philosophy of statistics. The Statistician, **49**, 293–337 (2000)
[69] Lindley, D.V., Scott, W.F.: New Cambridge Elementary Statistical Tables. London, Cambridge University Press (1984)
[70] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge, Cambridge University Press (2003)
[71] Morey, R.D., Pratte, M.S., Rouder, J.N.: Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. Journal of Mathematical Psychology (in press)
[72] Morey, R.D., Rouder, J.N., Speckman, P.L.: A statistical model for discriminating between subliminal and near-liminal performance. Journal of Mathematical Psychology, **52**, 21–36 (2008)
[73] Myung, I.J., Forster, M.R., Browne, M.W.: Model selection [Special issue]. Journal of Mathematical Psychology, **44**(1–2) (2000)

[74] Navarro, D.J., Griffiths, T.L., Steyvers, M., Lee, M.D.: Modeling individual differences using Dirichlet processes. Journal of Mathematical Psychology, **50**, 101–122 (2006)

[75] Nelson, N., Rosenthal, R., Rosnow, R.L.: Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, **41**, 1299–1301 (1986)

[76] Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A, **231**, 289–337 (1933)

[77] O'Hagan, A.: Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society, Series B, **57**, 99–138 (1997)

[78] O'Hagan, A.: Dicing with the unknown. Significance, **1**, 132–133 (2004)

[79] O'Hagan, A., Forster, J.: Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference (2nd ed.). London, Arnold (2004)

[80] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G.: Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: Introduction and design. British Journal of Cancer, **34**, 585–612 (1976)

[81] Pocock, S.J.: Group sequential methods in the design and analysis of clinical trials. Biometrika, **64**, 191–199 (1977)

[82] Raftery, A.E.: Bayesian model selection in social research. In: Marsden, P.V. (ed) Sociological Methodology. Cambridge, Blackwells (1995)

[83] Ramsey, F.P.: Truth and probability. In: Braithwaite, R.B. (ed) The Foundations of Mathematics and Other Logical Essays. London, Kegan Paul (1926)

[84] Rosenthal, R., Gaito, J.: The interpretation of levels of significance by psychological researchers. The Journal of Psychology, **55**, 33–38 (1963)

[85] Rosnow, R.L., Rosenthal, R.: Statistical procedures and the justification of knowledge in psychological science. American Psychologist, **44**, 1276–1284 (1989)

[86] Rouder, J.N., Lu, J.: An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychonomic Bulletin & Review, **12**, 573–604 (2005)

[87] Rouder, J.N., Lu, J., Morey, R.D., Sun, D., Speckman, P.L.: A hierarchical process dissociation model. Journal of Experimental Psychology: General (in press)

[88] Rouder, J.N., Lu, J., Speckman, P.L., Sun, D., Jiang, Y.: A hierarchical model for estimating response time distributions. Psychonomic Bulletin & Review, **12**, 195–223 (2005)

[89] Rouder, J.N., Lu, J., Sun, D., Speckman, P., Morey, R., Naveh-Benjamin, M.: Signal detection models with random participant and item effects. Psychometrika (in press)

[90] Royall, R.: The effect of sample size on the meaning of significance tests. The American Statistician, **40**, 313–315 (1986)

[91] Royall, R.M.: Statistical Evidence: A Likelihood Paradigm. London, Chapman & Hall (1997)

[92] Savage, L.J.: The Foundations of Statistics. New York, Wiley (1954)

[93] Savage, L.J.: The foundations of statistics reconsidered. In: Neyman, J. (ed) Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Berkely, CA, University of California Press (1961)

[94] Smith, A.F.M., Roberts, G.O.: Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B, **55**, 3–23 (1993)

[95] Spiegelhalter, D.J., Thomas, A., Best, N., Lunn, D.: WinBUGS Version 1.4 User Manual. Medical Research Council Biostatistics Unit, Cambridge (2003)

[96] Stuart, A., Ord, J.K., Arnold, S.: Kendall's Advanced Theory of Statistics Vol. 2A: Classical Inference & the Linear Model (6th ed.). London, Arnold (1999)

[97] Wagenmakers, E.-J.: A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review, **14**, 779–804 (2007)

[98] Wagenmakers, E.-J., Grünwald, P.: A Bayesian perspective on hypothesis testing. Psychological Science, **17**, 641–642 (2006)

[99] Wagenmakers, E.-J., Grünwald, P., Steyvers, M.: Accumulative prediction error and the selection of time series models. Journal of Mathematical Psychology, **50**, 149–166 (2006)

[100] Wagenmakers, E.-J., Waldorp, L.: Model selection: Theoretical developments and applications [Special issue]. Journal of Mathematical Psychology, **50**, 99–214 (2006)

[101] Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. New York, Springer (2004)

[102] Wilkinson, L., the Task Force on Statistical Inference: Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, **54**, 594–604 (1999)