# How Strong is the Confirmation of a Hypothesis by Significant Data?

Thomas Bartelborth*

Preprint: 2013-9-11

Abstract: The aim of the article is to determine how much a hypothesis *H* is actually confirmed if it has successfully passed a classical significance test. Bayesians have already raised many serious objections against significance testing, but in doing so they have always had to rely on epistemic probabilities and a further Bayesian analysis, which are rejected by classical statisticians. Therefore, I will suggest a purely frequentist evaluation procedure for significance tests, that should also be accepted by a classical statistician. This procedure likewise indicates some additional problems of significance tests. Such tests generally offer only incremental support of a hypothesis, although an absolute confirmation is necessary, and they overestimate positive results for small effects, since the confirmation of *H* in these cases is often rather marginal. This phenomenon leads in specific cases, for example, in cases of ESP-hypotheses, such as precognition, too easily to a significant confirmation. I will propose a method of how to evaluate and supplement significance tests so that we can avoid their epistemic deficits.

Keywords: significance tests, Bayesianism, confirmation, likelihoodism, falsification, inference to the best explanation.

## 1 Epistemic Uncertainty and the Logic of Significance Testing

Scientists of most empirical disciplines present the results of their research as soon as their theories have been successfully confirmed by a significance test. Many scientific journals only publish an article when its hypotheses have successfully passed a significance test. These tests are carried out to decisively reduce our epistemic uncertainty. However, despite the great importance this test procedure has in science, it remains unclear how strong the confirmation by a successful test of significance is. Does it provide us with a good reason to accept the hypotheses which have been proven to be

* University of Leipzig, Institute of Philosophy, Beethovenstraße 15, 04107 Leipzig, email: <bartelbo@uni-leipzig.de>.

significant? To answer this question, I will examine how significance testing works in simple examples and will develop a special method to evaluate the reasons we have to believe in significantly confirmed hypotheses.

For the justification of a hypothesis $H$ by data $E$, we first have to distinguish between different types of confirmation that justify different degrees of confidence in a hypothesis. In most cases the confirmation will be an *incremental confirmation* which makes the hypothesis more plausible, but does not confirm it to an extent that we should accept it. In order to accept a hypothesis, we need an *absolute confirmation*. Incremental confirmations will seldom change the epistemic status of the hypotheses significantly. In any case, it would be helpful if we could determine the exact degree of support, and if we would have a method of adding or balancing different confirmations of a hypothesis by different data. If we find a sufficient number of incremental confirmations of a hypothesis, we will eventually assume that it has become rational to accept the hypothesis for now. In this case, we would speak of an absolute confirmation and take the hypothesis to be true, at least, until further data appear.

The first problem relates to how we can represent different degrees of plausibility or support in the classical framework. For classical epistemology as well as for classical statistics the only representation of epistemic uncertainty that is allowed is given by a distribution of all statements into three categories: We first have the *belief set* $G$ of all accepted propositions, and second the dual set $R$ of rejected propositions, and third the neutral set $N$ of statements for which we have neither sufficiently good reasons for its acceptance nor for its rejection. In particular, in classical epistemology we do not work with degrees of belief as the Bayesians, who make finer epistemological distinctions and thereby can add up different incremental confirmations.

In the classical case, the reduction of uncertainty consists in shifting statements $H$ from the neutral set $N$ (by accepting them) to the belief set $G$ when we achieve an absolute confirmation of $H$. In classical significance testing (as advocated by Sir Ronald Fisher, e.g., Fisher, 1935a, 1935b) two hypotheses are involved: the so-called null hypothesis $H_0$ and an alternative hypothesis $H_1$, which I will call the *research hypothesis* (or *target hypothesis*), because it is the hypothesis we want to justify in the end. At the beginning, both hypotheses have to be placed in the neutral set $N$, and we aim at moving $H_1$ to the belief set $G$ (and accordingly the null hypothesis has to be clas-

sified as rejected). To do this, we need an absolute confirmation that statistical tests of significance therefore should provide. If, on the other hand, the results of these tests were insignificant both hypotheses should remain in the neutral set.

The logic of significance testing is thus a special case of *eliminative induction*. In order to confirm the target hypothesis, at first the target or research hypothesis $H_1$ is proposed as a sort of antithesis to a certain null hypothesis $H_0$. If, for example, our research question is whether a new teaching method has really improved the performance of students (which we want to measure with a certain standardized school achievement test) our null hypothesis may be that the new performance $\theta$ (as measured by the school test) has remained the same as before (measured in earlier tests): $\theta = a$. The value $a$ is our *null value*. If we then get some measurement results represented by a value of a random variable $T$ (and perhaps we find $T > a$), then we have to decide if the difference to our null value is only a chance variation (or measurement error) or is due to a real effect ($\theta > a$) of the new teaching method. Our research hypothesis $H_1$ in this situation may be that $\theta > a$ (one-sided test) or $\theta \neq a$ (two-sided test). For greater simplicity I will always work with one-sided tests $\theta > a$ here. In most cases, the null hypothesis represents the skeptical view that the observed deviation from our null value (here: $T > a$) is only a random fluctuation and represents no real effect (i.e. $\theta = a$).

In order to justify the research hypothesis we must, in the second step, falsify (probabilistically) the null hypothesis. If this is successful, we are then rationally justified in accepting the last remaining possibility: our research hypothesis $H_1$. If the observed data or the value of our test variable $T$ is too far away from the null value (to be interpreted as being due to pure chance), we reject the null hypothesis and speak of a *significant result* vindicating our research hypothesis. Typically, we use the so-called p-value to decide if we can reject the null hypotheses. If $T$ takes the value $x$ in a certain empirical study, we construct as our evidence set all values $T \geq x$ (including in the set all values of $T$ that speak against $H_0$ at least as strong as the value $x$ in a one-sided test), and then determine the p-value $p := \mathrm{P}(T \geq \mathrm{x} \mid H_0)$. Statistics tells us how we can determine this p-value by modeling the test situation in such a way that we obtain a probability distribution for $T$ (the sampling distribution) if we assume $H_0$ to be true.

Typically, if $p < 0.05$ or $p < 0.01$, we conclude that the result is too improbable (given $H_0$) such that we would remain with $H_0$. These values are called the *significance levels* of our test. They determine in which cases we regard the null hypothesis as effectively probabilistically falsified. Intuitively, we reject the null hypothesis on the basis of data which seem very improbable if the null hypothesis is true. Of course, this is no strict deductive falsification and this may be a first reason for some of the problems with significance testing that will be discussed in the following. The main question is how strong the confirmation given by significant data is for our research hypothesis. In many cases we can admit that we have a (weak) incremental confirmation, but it nevertheless seems to be much weaker than the absolute confirmation we need.

## 2 Significance Testing and Feeling the Future

A recent debate shows some of the problems of this method. The psychologist Daryl Bem examined in nine experiments phenomena of precognition (Bem 2011). For instance in the first experiment, the participants were supposed to predict on which side of a screen (left or right) certain pre-determined pictures would appear. In other experiments well known psychological effects (like priming) were tested to see if they also work reverse chronologically. These experiments were designed as significance tests, and in eight experiments the results confirmed significantly (and sometimes even highly significantly) the precognition hypothesis. In the first experiment of predicting the side of the picture the 100 test subjects (50 women and 50 men) with 36 trials each succeeded in 53.1% of the cases (p-value 0.011), and if we restrict the test to erotic pictures, they even succeeded in 57.6% of the cases (p-value 0.00008) against a null hypothesis that predicted only a hit rate of 50%. These results were eventually published in a prominent psychology journal.

The critics of these results (e.g. Alcock 2011) have raised many "minor" methodological objections against it, but the criticized methods of Bem are within the range of normal practices of psychological research—and also of that of other sciences. We will presumably have to admit that the test results speak weakly for the target hypothesis $H_1$ that precognition exists. But it remains the crucial question whether the confirmation of $H_1$ is strong enough that we should now accept $H_1$, or at least should regard it provi-

sionally as a scientifically acceptable hypothesis (as long as we have no other evidence). Since most scientists do not want to accept the precognition hypothesis on the basis of Bem's results, the method of significance testing was once again the target for criticism. Thus, the example of Bem (2011) seems to be a good illustration of how difficult it is to assess the strength of the confirmation provided by significance tests. In any case, some critics (cf. Wagenmakers et al. 2011) argued that significance testing makes it too easy to confirm a hypothesis, or that the test overstates the meaning of the data for the hypothesis.

My aim here is not to give a review of psi phenomena. There actually are some significant results speaking in favor of psi phenomena, which Bem (2011) mentions. However, they mostly have not been positively replicated; and the next difficult question is how we are to incorporate the results of replication studies in our assessment of hypotheses. There is unfortunately no easy rule for doing this. Furthermore, many renowned journals show no willingness to publish the results of replication studies at all; but due to the intense discussion of Bem's article the same journal published a great replication study (Galak et al. 2012), in which no evidence for the precognition hypothesis was found. In any case, we need, as a primary matter, highly dependable standard tests that can themselves provide us absolute confirmations.

Some of the direct criticisms of Bem's article (2011) are the following: Alcock (2011), for example, has argued, among other things, that Bem has actually done several t-tests at once, and hence the significance level should have to be corrected accordingly (perhaps on the basis of the Bonfferroni correction); and Alcock also argued against one-sided significance testing, because the error probability is then focused on one side, whereby it is easier to gain significant results. This corresponds to a deviation in two-sided tests with a significant result on each side at the 2.5% level. Both objections are primarily arguments for operating with a smaller level of significance, but Bem's procedure is a common practice in scientific research. In particular, these criticisms do not put into question the general method of significance testing.

We can find other criticisms of Bem's results in Wagenmakers et al. (2011). First, the authors complain that Bem has done an *exploratory study* and mixed it with a confirmatory study. Hence, they plead for a better separation of these types of research.

> Instead of presenting exploratory findings as confirmatory, one should ideally use a two-step procedure. First, in the absence of strong theory, one can explore the data until one discovers an interesting new hypothesis. But this phase of exploration and discovery needs to be followed by a second phase, one in which the new hypothesis is tested against new data in a confirmatory fashion. (Wagenmakers et al. 2011, 427)

This distinction between exploratory and confirmatory data is, however, rather doubtful from an epistemological point of view and in practice often difficult to check.

Although the mentioned objections to Bem's results seem to be thus rather marginal, we do not want to accept Bem's conclusion. Most scientists will probably see no sufficient reasons in the data of Bem to believe in precognition and will continue to assume that we until now have no absolute confirmation for the hypothesis of precognition. This certainly has to do with the fact that the hypothesis is in conflict with our basic views about the functioning of the world. Precognition is a form of reverse causality for which there seems to be no plausible mechanism. So we have the particular situation that the results of Bem in our example can speak more against the methodology used than in favor of precognition. This is, at least, how some psychologists as Wagenmakers et al. (2011) interpret the results.

## 3 Some Bayesian Objections against Significance Tests

Before I present my own approach for assessing confirmation by significance testing, I want to address some other serious criticisms, mostly put forward by Bayesians. Some Bayesians (e.g. Wagenmakers et al., 2011 Wetzels et al. 2011) argue that we have to look for a better method of hypotheses testing since significance testing makes it too easy to confirm a theory and the example of Bem (2011) was only one such case. Of course, they plea for the application of Bayesian methods in evaluating scientific theories and present us some of the classical problems of significance testing.

A reasonably comprehensive collection of these problems with many instructive examples can be found in Wagenmakers (2007). Wagenmakers divides the objections into three types. The first type of problems shows that the p-value is a function not of actual but of *hypothetical* data. To determine how strongly certain data confirm a specific research hypothesis (or speak against the null hypothesis) one uses the p-value $PV_T(E)$ of the data $E$ (represented by $T = x$) with respect to a certain test statistic $T$, which is a sum of

probabilities summed up over all values of $T$ that are at least as extreme as the one actually observed ($T = x$). That means, instead of working with $P(T{=}x \mid H_0)$ we use the likelihood $P(T{\in}D \mid H_0)$ with D: = $\{y \mid P(T{=}y \mid H_0) \leq P(T{=}x \mid H_0)\}$. For a one-sided null hypothesis we can normally simplify: $D = \{y \mid y{\geq}x\}$. Greco (2011) calls this step from $P(T{=}x \mid H_0)$ to $P(T{\in}D \mid H_0)$ "weakening the evidence". As a consequence of this procedure it may happen that evidence $E$ occurs and $E$ has the same likelihood in both theories $T_1$ and $T_2$ ($P(E \mid T_1) = P(E \mid T_2)$), but the theories suggest different p-values for $E$ and, therefore, perhaps $T_1$ is confirmed more strongly by E than $T_2$. The Bayesian regards this as implausible. In Bayesian epistemology often only $P(E \mid T)$ is taken as indicator of the confirmation of $T$ by $E$.

Bayesian methodology is, in this case, based on the so-called *law of likelihood* that looks quite plausible at first sight. It says that the likelihood $P(E \mid T)$ should determine the strength of the confirmation of T provided by E, and especially when comparing two theories $T_1$ and $T_2$ that the likelihood ratio $P(E \mid T_1)/P(E \mid T_2)$ expresses the relative confirmation E provides to $T_1$ relative to $T_2$. Likelihoodists (as Royall 1997) often argue in this context that confirmation is in every case (only) *comparative*. Modern Bayesians also often use the corresponding Bayes factor $BF_{01}$ as a measure of comparative confirmation that can decide between the null hypothesis $H_0$ and our research hypothesis $H_1$. The Bayes factor coincides with the likelihood ratio when the likelihoods are objective. For the Bayesian the importance of the Bayes factor $BF_{10}$ results from its role in the updating of two hypotheses by data $E$:

$$\frac{P(H_0 \mid E)}{P(H_1 \mid E)} = \frac{P(H_0)}{P(H_1)} \cdot \frac{P(E \mid H_0)}{P(E \mid H_1)} = \frac{P(H_0)}{P(H_1)} \cdot \mathrm{BF}_{01}$$

The Bayes factor $BF_{01}$ is the *update factor* for Bayesians if we consider the ratio of the posterior probabilities of two hypotheses. This establishes its epistemic significance. Two advantages of this factor with respect to significance testing are firstly that it is *symmetrical* in both hypotheses and thus can also constitute a confirmation of the null hypothesis[1] and secondly that it allows us to quantify the *degree* of relative confirmation. Furthermore,

[1] Rouder et al. (2009) argues that any test procedure that only allows confirmation of the target hypothesis but has no means to confirm the rival null hypothesis is always biased by this fact alone to the target hypothesis. This seems to me an interesting general point for further epistemological discussions.

new data (e.g. replication studies) can be taken quickly into account by again multiplying with the new Bayes factor. In addition, the Bayes factor seems not to be dependent on the prior probabilities of the hypotheses and thus seems to be an objective magnitude.

However, we still have the problem that the value $P(E|H_1)$ for our target hypothesis cannot be easily determined because it typically still contains a (nuisance) parameter $\theta$ (in our above example for the mean school achievement) which we do not know. The hypothesis $H_1$ (which for example says that $\theta > a$) does not directly determine probabilities for the occurrence of certain data E, since it is no point hypothesis ($\theta = b$) but a "long disjunction" of point hypotheses. Therefore, we need a density distribution $f(\theta|H_1)$ for the parameter $\theta$ of $H_1$ in order to integrate it out: $P(D|H_1) = \int P(D|\theta, H_1) \cdot f(\theta|H_1)\, d\theta$. These calculations will often be somewhat complicated (and we regularly need computer assistance by special Bayesian programs), and, in particular, the dependence on a prior density $f(\theta|H_1)$ may lead to some objections to this procedure (s. below). Furthermore, Bayesians provide us with certain approximations as the Savage-Dickey Density Ratio or the Bayesian Information Criterion (BIC) for an easier application of their methodology (cf. Rouder et al. 2009).

Bayesians, in addition, supply a classification in which case the Bayes factor $BF_{10}$ indicates a more or less strong comparative confirmation of $H_1$ relative to $H_0$. Dating back to Jeffreys they interpret 1–3 as insignificant confirmation of $H_1$, 3–10 as weak confirmation and speak of above 10 as strong confirmation. While these assessments are not entirely uniform by various Bayesians, we do have an intuitively comprehensible characterization of this magnitude: If we have only two hypotheses and start with the same prior probability $P(H_1) = 0.5 = P(H_0)$ for both, we can use the Bayes factor to calculate the posterior probability of the hypotheses: $P(H_1|D) = BF_{10} / (1 + BF_{10})$. Thus we obtain for $BF_{10} = 3$ just $P(H_1|D) = 0.75$ (for a Bayes factor of 10 we get the posterior probability 0.91, for $Bf_{10} = 20$ we get 0.95 and for Bayes factor of 100 we get 0.99). We can thereby at least interpret a Bayes factor of 10 as indicating an absolute confirmation if we work with a threshold conception of theory choice and regard a probability above 90% as sufficient to accept a theory. In order to facilitate comparisons between classical (three-valued) epistemology and Bayesianism, I will take such a threshold conception of theory acceptance for granted in the rest of the article. If the posterior probability of a theory $T$ arrives at 90% or above, I will thus

regard $T$ as absolutely confirmed by our background knowledge and will accept it. If we start with two hypotheses and are at first neutral with respect to them, we should, therefore, assess confirmations with Bayes factors below 10 only as weak incremental confirmations.

If we apply these ideas to the experiments of Bem (2011), we obtain for the research hypothesis $H_1$ only three values below 3 and one of 5.9 while the other experiments speak weakly for the null hypothesis, and we even get values up to 7.6 in favor of the null hypothesis (see Wagenmakers et al. 2011). According to this analysis we have in no experiment convincing data in support of Bem's precognition hypothesis. The question thus is if the interpretation of the data as absolutely confirmatory for his hypothesis should mainly be attributed to the Fisherian methodology of null hypothesis testing and not to the power of the data. If this is the case, it would perhaps also explain (at least in part) the great popularity of significance testing among the practitioners of the empirical sciences.[2]

In their reply Bem et al. (2011) complain that the Bayes factor in Wagenmakers et al. (2011) is not really objective since it depends on a prior density $f(\theta|H_1)$, which according to Bem is ill chosen since it is too uninformative and does not use our prior knowledge about effect sizes in psychology. Hence, it assigns too much weight to implausibly high values of effect size and thereby favors the null hypothesis. If we instead take the typical effect sizes and other prior information into account, we get higher Bayes factors that in most studies indicate substantial confirmation for the precognition hypothesis. If we, furthermore, combine the evidence across the nine studies by multiplying the nine Bayes factors we even get extreme evidence in favor of the precognition hypothesis. Of course, Wagenmakers et al. (2011a) do not agree with this diagnosis and present several objections to this procedure, but they are far from being as convincing as their original points against Bem. For my evaluation of significance testing the lesson is to com-

[2] In a similar way Wetzels et al. (2011) argue for the same conclusion. They have analyzed 855 examples of t-tests (with p-values between 1% and 5%), which were published recently in psychology, using the corresponding Bayes factors, and came to the conclusion that according to a Bayesian analysis we can speak in 70% of these cases only of "anecdotal evidence" with a Bayes factor below three. They claim that significance testing would systematically overestimate the strength of the confirmation of the target hypotheses by the data.

pare only point hypotheses with other point hypotheses in order to avoid these problems of choosing a subjective prior density.

In the context of those cases in which a second point hypothesis exists that specifies in the same way as the null hypothesis a particular model and thus determines the likelihoods for certain data, particularly likelihoodists as Royall (1997) argue for a strictly comparative evaluation of the hypothesis, and suggest the likelihood ratio (which corresponds to the Bayes factor) as a measure for the relative confirmation of the two hypotheses. However, likelihoodists do not accept epistemic probabilities in the form of prior probabilities or densities. Only in the case of objective likelihoods do they rely on the law of likelihood, which seems to be intuitively very plausible at the first sight. It says: If $P(E|H_2) > P(E|H_1)$ then the evidence $E$ favors (prima facie) hypothesis $H_2$ to hypothesis $H_1$.[3]

Unfortunately, Fitelson (2011) has argued against the law of likelihood with certain counter-examples: Consider a well-mixed typical deck of cards and draw randomly a card $K$ for which we have two competing hypotheses: ($H_1$) "$K$ is a black card" and ($H_2$ ) "$K$ is an ace of spades". Our information is: ($E$) "$K$ is a spade". Now, indeed it is $P(E|H_2) = 1 > P(E|H_1) = 0.5$. Thus, according to the law of likelihood $E$ supports $H_2$ more than $H_1$, but on the other hand, $H_1$ follows deductively from our information E, and logical derivability seems to be stronger than the purely probabilistic support of the likelihood comparison. Of course, we have here a special situation, which is due to the logical relations that obtain: $H_2 \Rightarrow E \Rightarrow H_1$.

This is obviously not a common situation for hypotheses testing and, thus, we should not immediately reject the law of likelihood altogether. But it remains an open question whether it should be mandatory. Consider a further simple example: Suppose for two diseases $X$ and $Y$ we find that in both cases for 20% of the ill persons their maximum temperature is 39° yet for $Y$ additionally in 10 % of the cases temperatures may rise up to 40°. ($E$) John has now one of the diseases and has a maximum temperature of 39°. Then we get: $P(E|X) = 0.2 = P(E|Y)$. Are we thereby committed to the opinion that both hypotheses are supported alike by $E$? One can argue in this direction but one need not. We can also argue that such a high fever is a pretty extreme value for disease $X$ but is not quite as rare for disease $Y$ and,

[3] If, e.g., $H_1$ and $H_2$ both assert a certain disease $X_1$ or $X_2$ respectively to be present and for $X_1$ certain symptoms $E$ can be expected, while this is not the case for $X_2$, then the occurrence of symptoms $E$ speaks in favor of $H_1$ relative to $H_2$.

hence, speaks a little more for *Y*. Our intuitions are not entirely clear in such cases. The law of likelihood therefore remains plausible, but it is not so well established that it could decide the debate in favor of Bayesianism.

As a consequence, in comparing significance testing with Bayesianism I will determine the posterior probability of the target hypothesis, which we can calculate in simple examples and which gives us presumably a better intuitive assessment of the strength of the confirmation than the Bayes factor. To facilitate such a comparison I'm again assuming that a probability of about 90% corresponds to the acceptance of a hypothesis. Of course, we could also choose other thresholds, but for an intuitive evaluation of the classical approach this setting turns out to be helpful.

A second objection of Wagenmakers (2007) against significance testing goes in a similar direction as the first one. The calculation of p-values depends not only on the data but also on our *sample plans* or *stop rules*. Hence, the data *D* do not determine the confirmation of the hypotheses, but it is essential for the test which intentions the experimenter had when he planned the test. It may be that two scientists have exactly the same data *D*, but they support a hypothesis *H* with different strength for the two scientists just because they had different sample plans in mind. According to the Bayesians this demonstrates an unnecessary subjective influence on theory choice, which for Bayesians should be only based on the data and perhaps further objective background knowledge but not on the goals of the scientists collecting the data. Bayesians, therefore, have developed some examples in which these influences actually appear to be entirely strange.

Consider such a simple example: We toss a coin 20 times and it lands heads 15 times. If we designed the experiment for exactly 20 throws, then this result is a significant deviation from the null hypothesis that the coin is fair (calculated using the standard binomial distribution). But we will get another p-value if we use another sample plan as for example: *toss the coin until you get 15 times heads* (calculated by the negative binomial distribution) or even: *toss the coin until you get a significant result*. For the last case, William Feller (1968) has already shown that we arrive at a significant result with probability one. These cases show why it's necessary for significance testing to take into account the intentions of the experimenter in planning the experiment. This subjective influence on the test results is of course irritating, but certainly not a knock-out criterion for the Fisherian approach. The classical statistician can point out that Bayesians also have to

use subjective elements in their theory evaluations as the prior probabilities of hypotheses and perhaps some likelihoods (especially the so called *catch-all likelihood*).

The third type of criticism of significance testing from a Bayesian perspective is that the p-value is often regarded as a measure of the strength of the confirmation. Several classical statisticians assumed (s. Wagenmakers 2007, 787 ff.) that the smaller the p-value and the greater the number $n$ of tested subjects the better is the support of the target hypothesis, since the evidence against the null hypothesis is measured by the p-value and large trials give us the most reliable evidence against it. Wagenmakers mentions for example Fisher for whom the p-value gives "the strength of the evidence against the hypothesis" (Fisher, 1958, p 80). But this assumption (called the *p-postulate* by Wagenmakers) can not be maintained if we take into account certain Bayesian considerations presented firstly in 1957 by Lindley, which are now known as Lindley's paradox. It shows, at least, that a smaller p-value need not be accompanied by a higher posterior probability of the target hypothesis. More specifically, it can be shown that for every constant p-value $\alpha$ and an increasing number $n$ the posterior probability of the null hypothesis converges (slowly) to one for all plausible prior probabilities. Bayesians considered this phenomenon a fatal flaw of significance testing and for the p-postulate, since in this case the same p-value leads to different confirmations of the null hypothesis, and we recognize even a decreasing plausibility of the target hypothesis in contrast to the p-postulate for greater values of n.

Berger & Delampady (1987) suggested a possible answer which the classical statistician can give: The point-null hypothesis $\theta = a$ is only an idealization actually representing the more complex case where the value lies in a small interval around $a$, and it would even be unusual for a Bayesian to give a strictly positive probability to such a point hypothesis. In the Bayesian account we eventually compare two models which do not fit well together, namely a point hypothesis and an interval hypothesis. This idealization produces only small deviations for the classical statistician but is more problematic for a Bayesian (cf. Berger & Delampady 1987, 321). In any case, for the more realistic representation of a null hypothesis as small interval, the convergence which the results in Lindsey's paradox suggest cannot be reproduced in the Bayesian framework, but the probability of the null hypothesis converges in many cases to $\alpha$ (see Berger & Delampady 1987, 322).

Nonetheless, important differences also remained in the article of Berger & Delampady between measures of confirmation based on p-values and the corresponding Bayesian calculations that culminate for the authors in the conclusion:

> The overwhelming conclusion is that P-values are typically at least an order of magnitude smaller than Bayes factors or posterior probabilities for $H_0$. This would indicate that say, claiming that a P-value of .05 is significant evidence against a precise hypothesis is sheer folly; the actual Bayes factor may well be near 1, and the posterior probability of $H_0$ near 1/2. (Berger & Delampady 1987, 326)

Consequently, the Bayesians have, after all, raised several serious objections to significance testing and certainly offer a number of important alternative methods to it. Furthermore, it seems to be too easy to confirm ESP-hypotheses by tests of significance. However, there is a certain incommensurability between the classical and the Bayesian approach. The Bayesian methodology works with epistemic probabilities, and normally specifies no threshold value for accepting a scientific hypothesis, whereas the classical statistician, on the other hand, rejects epistemic probabilities altogether, works primarily with frequencies or likelihoods of the form $P(E|H)$ and has no place for degrees of belief in a hypothesis. We have, therefore, to look for a convincing and more direct epistemological evaluation of the Fisherian method that classical statisticians can accept.

This new assessment method will be based on the relative frequencies with which the Fisherian method will give us the right results, but these frequencies have subsequently to be interpreted as certain expectations similar to epistemic probabilities. This is the only way to derive rational expectations from relative frequencies. Though this form of statistical syllogism is officially neither appreciated by Bayesians nor by classical statisticians (since it is the basis of an inductive logic) both approaches have to rely on it if they want to provide epistemologically relevant results (cf. Franklin 2001, Campbell & Franklin 2004, and Bartelborth 2012, ch. 5.2).

## 4 Epistemological Evaluation: The Filter Analogy

The question remains how we can directly assess significantly confirmed hypotheses $H$ in the classical frequentist framework. The likelihoods $P(E|H)$ of the data $E$ offer no direct evidence for the probability of hypothesis $H$. But, we want to know whether the hypothesis is probably true if cer-

tain data occur, and to answer that question seems to require (prior) epistemic probabilities. Of course, the error probability (or significance level) of 5% does not mean that significantly confirmed hypothesis are only false in 5% of the cases, although we sometimes come across misleading remarks in statistics textbooks hinting in this direction. Then, how can we evaluate epistemologically significance tests? We want to know whether a significantly confirmed target hypothesis $H_1$ is true or at least supported so well that we should accept it, at least preliminarily.

The Bayesian determines the epistemic probability of $H$ for this purpose, but the classical statistician accepts only objective likelihoods or relative frequencies. Thus, I will base my assessment of the plausibility of significantly confirmed hypotheses solely on relative frequencies. For this purpose we have to use the following analogy (cf. Bartelborth 2012: ch. 6.4): If we have a true null hypothesis $H_0$ (and thus a wrong target hypothesis $H$) and we accomplish a significance test on $H_0$ with significance level $\alpha$, then our test statistic will still fall into the rejection region in $\alpha{\cdot}100\%$ of the cases, and we would, therefore, falsely reject the null hypothesis. If we set $\alpha = 5\%$, that means that of 100 false target hypotheses which we send into the test on average 5 of them will be confirmed by it. The test works as a *filter* that filters out 95 of 100 wrong target hypotheses which we have sent into the filter.

Of course, we still do not know how many of the target hypotheses which pass the filter are true. In order to determine that magnitude we additionally have to know the $\beta$-error, which gives the proportion of true target hypotheses that are also stopped by the filter. Even if the target hypothesis is true (and the null hypothesis is therefore wrong), the null hypothesis will not always be rejected. In $\beta{\cdot}100\%$ of these cases we still obtain an insignificant result, and the true target hypothesis will not be confirmed as significant. These target hypotheses will remain stuck in the filter and therefore remain in the neutral area, since only those research hypotheses will be accepted that have passed the filter. Altogether, from the false target hypotheses $\alpha{\cdot}100\%$ will pass the filter and from the true ones $(1{-}\beta){\cdot}100\%$ will pass it, and will thus be accepted. Since the $\beta$-error is increasing with decreasing significance level $\alpha$ (our filter lets pass on the whole fewer hypotheses) both error types depend on each other, but in general we do not know the $\beta$-error and, at best, can estimate it. But the basic idea of significance testing is, of course, that only few false and many true target hypotheses

will pass the filter, whereby a passing of the filter can be regarded as an indicator for the truth of these hypotheses.

However, to exactly determine the overall effect of the filter, we also need the ratio of true hypotheses to false hypotheses that are sent into the filter. Suppose that, of the hypotheses which we send into the significance filter, 100·$q$% are true and 100·(1–$q$)% are false, then I will call $q$ the *prior truth-proportion*. Now we can figure out how many of the significantly confirmed hypotheses (that means the hypotheses that pass the filter) are true on average. We have to determine (or estimate) the three parameters $\alpha$, $\beta$, and $q$ to arrive at the *posterior truth-proportion* $w$:

$$w = \frac{q(1-\beta)}{\alpha(1-q)+q(1-\beta)}$$

Thus, a significance test is a method that lets a certain proportion of target hypotheses be confirmed as significant through the filter. In particular, the majority of the false hypotheses are filtered out while the true hypotheses will pass mostly through the filter. In the ideal case, we obtain with this filtering a very high proportion of true target hypotheses as output, as is depicted in the following graph:

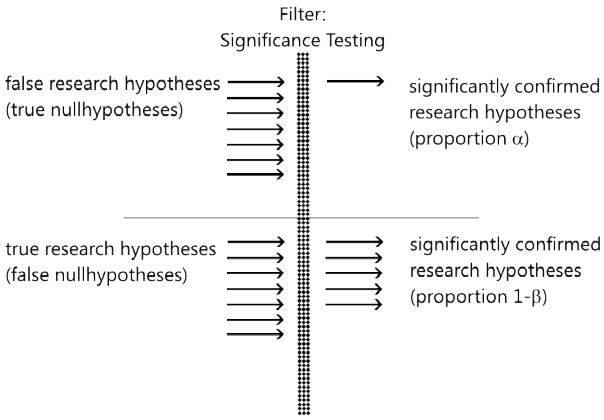


Figure 1: Significance testing considered as a filter for hypotheses

The posterior truth-proportion offers the best method for a direct epistemic evaluation of significance testing because it tells us how many true research hypotheses we find among the (significantly confirmed) hypotheses pass-

ing the filter and therefore characterizes the performance of the filter. If we want to find out the plausibility we should assign to a specific hypothesis $H$ that is confirmed by a test, we have additionally to rely on the above mentioned statistical syllogism. It says that if $x$·100% of some kind of objects have the characteristic $F$ and we have an object a of this kind and have no further knowledge whether it has $F$, then we should give $Fa$ the epistemic probability $x$ or the degree of plausibility $x$ ($P(Fa) = x$). There are good reasons to use this form of empirical calibration for our degrees of plausibility, but I will not discuss them here and rather just accept this intuitive principle at this point. It seems to be the only plausible method of how we can rationally learn from purely frequentist data and can use them as a guide to our life.

If we use this type of reasoning in our example (and how can we otherwise use statistical information), then we should take the posterior truth-proportion $w$ as our best estimate for evaluating the epistemic support which our data provide for the hypothesis $H$. If we have further knowledge about the truth of our research hypotheses, that can of course be used to estimate the prior truth-proportion. In any case, a first question is how the prior truth-proportion influences the posterior truth-proportion. We can acknowledge here similarities to the Bayesian updating procedure but the values are different as we will see later.

First of all, suppose that we use a significance test with fixed values $\alpha = 0.05$ and estimated $\beta = 0.2$. In this case the dependence of the ratio of the posterior truth-proportion on the prior truth-proportion is obvious. However, we can improve the test procedure if we choose a smaller significance level of $\alpha = 0.01$ or even $\alpha = 0.001$ and thereby increase the posterior truth-proportion (if the $\beta$-error is not too large). I assume that $\beta$ is adjusting in these cases only to 0.5 and 0.8 respectively, but we will see later that $\beta$ can even rise to critical values in concrete cases. That depends mainly on where the actual parameter that we want to determine is located.

| **prior truth-proportion $q$** | **0** | **0.001** | **0.01** | **0.1** | **0.3** | **0.5** | **0.6** |
|---|---|---|---|---|---|---|---|
| $\alpha = 0.05$ ($\beta = 0.2$) | 0 | 1.6 | 14 | 64 | 87 | 94 | 96 |
| $\alpha = 0.01$ ($\beta = 0.5$) | 0 | 5 | 34 | 85 | 95 | 97 | 98 |
| $\alpha = 0.001$ ($\beta = 0.8$) | 0 | 17 | 67 | 96 | 99 | 99.5 | 99.7 |

Table 1: Posterior truth-proportions $w$ in percent

In the second column we can see that we only obtain false hypotheses passing through the filter if all hypotheses we think of and send into it are false. This is a first explicit hint that significantly confirmed hypothesis need neither be true nor even plausible. Furthermore, if a scientist is very productive, testing many hypotheses, of which only every tenth is true, and he works with a significance level of 5%, then its significant results will be true on average only in 64% of the cases. For most research questions that does not seem to be a sufficient certainty to rely on—for example in the case of therapies. Nevertheless, if the applied p-value is much smaller than 5% that means that the posterior truth-rate is clearly rising if the $\beta$-error will not become too large. The p-value is thus a first helpful indicator of the strength of the confirmation of a target hypothesis by significant data. However, there are also other parameters to consider and we will later see that the $\beta$-errors can become very large (even approaching 1) and then can certainly also affect the posterior truth-proportion substantially. Thus, in particular, we cannot identify the posterior truth-rate only by means of the p-value. We also have to take the $\beta$-error and the prior truth-rate into consideration.

Moreover, at the least, it seems to be clear that the strong dependence of the posterior truth-proportion $w$ on our prior truth-proportion $q$ has the consequence that *only plausible hypotheses should be sent into the filter* if we want to have a good reason to believe that significantly confirmed hypotheses are true. The classical statistician must therefore—quite similar to the Bayesian—at first evaluate the initial plausibility of his research hypotheses before testing them with a significance test, since, otherwise, many false target hypotheses will pass the filter and the posterior truth-proportion becomes too low. This is obviously a serious problem for the testing of ESP-hypotheses such as Bem's precognition hypothesis. We have to at least estimate the relevant parameters in these cases in order to assess the epistemic status after the successful tests with the help of the posterior truth-proportion in such cases.

How can we accomplish such an informal plausibility assessment? At a minimum, I would like to mention some aspects which we have to take into consideration: We have, first of all, to examine whether the target hypotheses can indeed *explain* our data and whether they can do it better than any *alternative hypotheses*. To this purpose we have to search explicitly for alternative explanations of the data and have to determine how many plausible alternative explanations for the data are to be found. Furthermore, we have

to assess how well our research hypothesis fits with our other accepted scientific theories. The Bayesian attempts to provide an exact number for this initial plausibility of our hypotheses as their prior probability (which is certainly a difficult enterprise), but at least we have to give an informal assessment of the initial plausibility, which tells us that they are plausible candidates for the best explanations of our data.

From now on I will always assume that the target hypotheses and the null hypotheses have (at the start) the same initial plausibility and thus that our prior truth-proportion for the tests is approximately $q = 0.5$. But even in this favorable situation we will encounter further epistemic problems for the practice of significance testing.

## 5 Problems of Significance Testing in Concrete Examples

We will now look at some problems of classical significance testing in a simple example. It involves hypotheses about a proportion. Our null hypothesis always says that a particular coin is fair (i.e. P(head) = 0.5). The data we consider at first are the number of heads in 60 tosses of the coin. The test statistic $T$ (as the number of heads in the sample) has a binomial distribution and can therefore be calculated easily. This situation corresponds, e.g., to a hypothesis that postulates a proportion of 50% women in the total population of all students in England who enroll in a particular year of philosophy. In this case we can choose a random sample of 60 students for whom we determine, as a test statistic, how many women are to be found in the sample. If we assume that the population of all students is far greater than 60 we can again assume a binomial distribution for the test statistic. We will consider for simplicity only the one sided target hypothesis $H_1$ that the coin is biased in favor of heads (or that we find more girls in the population than boys). In the example of Bem, the null hypothesis corresponds to a random hit rate of just 0.5 and the target hypotheses are designed as one sided and postulate that the true hit rate $\tau$ is greater than 0.5.

The first problem is that the target hypothesis in our case consists of a large "disjunction" of the different values $\tau > 0.5$. Therefore, it provides no objective likelihoods and accordingly no specific $\beta$-error. To obtain determinate values we could adopt the Bayesian strategy and assume a prior density for $\tau > 0.5$ (indicating with which probability these values $q$ will occur) and then remove the unwanted parameter by integrating it out (see above),

but that cannot be accepted by the classical statistician because it necessarily depends on using epistemic probabilities. Moreover, we would thereby compare a point hypothesis with a weaker interval hypothesis, which can lead to miscalculations due to the additional free parameter. Actually, we know that some concrete value $\tau = r$ is the true value specifying the probability with which our coin comes up heads. The true hypothesis is hence a point hypothesis with some value $r$ which we do not know. We can now compare our null hypothesis with different hypotheses $H_r$ (saying $q = r$) for several values of $r$ and examine in each case how good different data (which will significantly confirm $H_r$) will really support the research hypothesis $H_r$.

I will, therefore, consider different hypothetical values of $\tau$, and then examine in each case whether the significance test interprets the data coming up intuitively correct (as confirming or not) and whether significant results will actually support our target hypothesis so strongly that we should regard it as absolutely confirmed and accept it. For this case to obtain I will require a posterior truth-proportion of at least 90%. The general (interval) research hypothesis $\tau > 0.5$, on the other hand, provides us with a certain kind of averaging over all such hypothesis, which we will consider here in isolation. Furthermore, we can compare the results with the corresponding Bayesian results for specific (point) hypotheses, and to facilitate this comparison I will always assume that the prior truth-proportion (in the classical case) and the prior probability (in the Bayesian case) is in each example 1/2.

In order to uncover some of the problems of significance testing we will consider a different specific target hypothesis according to which $H_1$ states that it is (approximately) P(head) = 0.7 or has another concrete value down to 0.51. In each of these scenarios we will assume that the null or the target hypothesis is true and can therefore also give a Bayesian analysis as an additional benchmark for our truth-proportion assessment. We can see, e.g., that if the null hypothesis and target hypothesis make very similar predictions (are close-by hypotheses) then the posterior truth-proportion of the target hypothesis shows that our hypothesis choice becomes a form of guessing. This is a problem well known to the Bayesian, but it does not seem to be similarly obvious for classical significance testing.

In our example, we consider the $\alpha$-errors 0.05 and 0.01 and 0.001 respectively, calculate the corresponding $\beta$-errors for different target hypotheses, and examine how they affect the posterior truth-proportions. For 60 tosses

of a coin the acceptance set for the null hypotheses and $\alpha = 0{,}05$ goes up to 36 times heads ($k = 36$), and for $\alpha = 0{,}01$ up to 39 times heads and at least for $\alpha = 0.001$ up to 42 times heads. In addition, I assume an optimistic truth-proportion of 0.5, and will besides determine the Bayesian posterior probability updated with results that fall just in the significant range, whereby I assume a prior probability of 0.5. Note that only significant results will be considered in the table.

Furthermore, I have marked hypotheses with posterior truth-proportions and probabilities of more than 90% as absolutely confirmed to alleviate the comparison between the two methodologies. The values for $w$ in the table give us the posterior truth-proportion that we can (on average) expect for certain significance levels $\alpha$ and a certain definite target hypothesis $H_1$ for 60 trials. A target hypothesis $H_1$ that has passed a significance test successfully has to be regarded as plausible to the degree $w$, since this is the proportion of true hypotheses in its new reference class. Bayesian calculation provides us with a different evaluation, but we have to keep in mind that I always update with the least value that is significant.

| $H_1$: $100 \cdot \tau =$ | **70** | **65** | **60** | **59** | **58** | **57** | **55** | **52** | **51** |
|---|---|---|---|---|---|---|---|---|---|
| $\beta$ (for $\alpha$=0.05) | 6 | 25 | 55 | 61 | 67 | 72 | 82 | 92 | 94 |
| $w$ (for $\alpha$=0.05) | 95 | 94 | 90 | 89 | 87 | 85 | 78 | 63 | 56 |
| P($H_1$ \| $k$=37) | 67 | 82 | 83 | 83 | 81 | 80 | 75 | 63 | 57 |
| $\beta$ ($\alpha$=0.01) | 24 | 55 | 82 | 86 | 89 | 92 | 96 | 98 | 99 |
| $w$ ($\alpha$=0.01) | 98.7 | 98 | 95 | 93 | 92 | 89 | 82 | 60 | 51 |
| P($H_1$ \| $k$=40) | 96 | 97 | 95 | 93 | 92 | 90 | 85 | 68 | 60 |
| $\beta$ ($\alpha$=0.001) | 55 | 83 | 96 | 97 | 98 | 98.6 | 99.4 | 99.8 | 99.9 |
| $w$ ($\alpha$=0.001) | 99.8 | 99.4 | 98 | 97 | 95 | 93 | 86 | 60 | 48 |
| P($H_1$ \| $k$=43) | 99.7 | 99.5 | 98 | 97.7 | 97 | 95.6 | 91 | 73 | 62 |

Table 2: Posterior truth-proportions $w$ (and $\beta$-errors) for $\alpha = 0{,}05$ and $\alpha = 0{,}01$ and $\alpha = 0{,}001$ with $n = 60$ and different target hypotheses $H_1$ (values in percent)

We can see in table 2, at first, that for $\alpha = 0.05$ and a greater distance between the null and the target hypotheses significance testing generates greater truth-proportions than the probabilities provided by Bayesian updating. This seems to be an advantage of significance testing over the Baye-

sian approach for this special situation of large or middle sized effects and only few trials. But that impression is misleading. In fact, the significance filter works well on average, because in the given situation (in which we assume that $H_0$ or $H_1$ is true) $k = 37$ only seldom occurs. As a result, in 95% of the cases the hypotheses that have passed the filter are true. The filter is definitely reliable on average, but it does not tell us the degree of confirmation given by the special datum $k = 37$. That is provided by the Bayesian analysis, and we see (which seems to be obvious in this case) that the confirmation of the target hypothesis by this datum is almost negligible. In that case, why do we not opt for the Bayesian analysis as the correct way to determine the confirmation? The classical statistician will answer that we normally are not in the comfortable position to know which point target hypothesis is the only candidate next to the null hypothesis. We just have the null hypothesis and can determine how probable certain data are if we assume it. Therefore, we have to be content with the averaging test procedure of significance testing. In table 2 we only examine how this test procedure works by looking at different possible point hypotheses, but we are rarely in the position to specify one of them as privileged.

Moreover, we can see in table 2 that the p-postulate has some truth in it in such situations, which can be established further in the following comparison with Bayesian posterior probabilities.

| **k** | **31** | **33** | **35** | **36** | **37** | **38** | **39** | **40** | **41** | **42** | **43** | **45** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p-value | 45 | 26 | 12 | 8 | 5 | 2.6 | 1.4 | 0.7 | 0.3 | 0.13 | 0.05 | 0.007 |
| $P(H_1 \mid k)$ | 1 | 6 | 27 | 46 | 67 | 82 | 92 | 96 | 98 | 99 | 99.7 | 99.94 |

Table 3: p-values (in percent) for $n = 60$ and $k$ times heads
(and $P(H_1 \mid k)$ for the target hypothesis $\tau = 0.7$)

Nevertheless, from the Bayesian point of view only for smaller $\alpha$-errors do we get a considerable conformity with the proper judgment. For $\alpha = 0.01$ and for $\alpha = 0.001$ we find in table 2 great similarities in the assessments between the two approaches.

Another important effect is that with decreasing distance between the two hypotheses the $\beta$-error increases dramatically (up to $1-\alpha$) and, thus, the posterior truth-proportion drops considerably. We already know this phenomenon from Bayesian calculations, but in the process of classical significance testing the phenomenon does not appear, because the method requires only a falsification of the null hypothesis and does not actually proceed comparatively. In cases of small effects the Fisherian methodology in-

dicates a significant confirmation although the posterior truth-proportion shows the confirmation to be rather weak. In the more extreme cases we can hardly speak of a notable confirmation at all, e.g. when the truth-proportion falls clearly below 80%. Significance testing obviously falls short of its goals for these cases, which means that these data have no real impact for the question which of the two hypotheses is probably true. The filter lets pass as many false as true hypotheses and thereby looses its original function, which was based on the assumption that the filter predominantly withholds the false hypotheses but lets the true ones mostly pass. So especially in the problematic cases of only small effects we must use other methods or, at least, work with a much smaller $\alpha$-error.

Intuitively, we can describe this phenomenon in the following way: If we find only a small distance between the null and the target hypothesis our data are often not sufficient to decide between the hypotheses. Both hypotheses are likewise compatible with the data since the hypotheses tell us similar things about the world, i.e., the likelihoods for the data diverge only slightly. In these cases, we obviously need more data to decide between the hypotheses than those data which we need in order to get a significant confirmation in the Fisherian methodology. Hence, the data are often clearly *over-interpreted* by the significance test. This problem is overlooked by the classical statistician because he does not use comparative methods and therefore the distance between the hypotheses plays no role in his procedure. The only aspect that is taken into consideration is if the data are probabilistically compatible with the null hypothesis, and it is not asked how compatible they are with the target hypothesis. In particular, no comparison takes place in form of the Bayes factor (or likelihood ratios) as in other approaches. I will call the resulting problem *the over-interpretation of data for small effects*.

This problem presumably relates to the experiments of Bem (2011), but certainly also to many results in other disciplines. This corresponds to the assessment of Wagenmakers et al. (2011), which he has, however, achieved in a different way (by the calculation of Bayes factors). Accordingly, the classical statistician comes too easily—especially for small effects—to significant results and these results provide only a small and incremental confirmation of the target hypothesis, but not the required absolute confirmation. In order to determine the validity of a significance test we have thus to

estimate the effects and then the truth-proportions before and after the experiment.

Our next question is how the number of trials or objects of investigation influences the results. Here we can find another phenomenon that should worry us:

| $H_1$: **P(Kopf) = $\tau$ =** | **0.7** | **0.65** | **0.6** | **0.59** | **0.57** | **0.55** | **0.52** | **0.51** |
|---|---|---|---|---|---|---|---|---|
| **$\alpha$=0.05**, **$n$=20**, maintaining the null hypothesis up to and including $k$=14 | | | | | | | | |
| $w$ | 94 | 92 | 89 | 88 | 86 | 83 | 78 | 75 |
| P($H_1$\|$k$=15) | 92 | 89.6 | 83 | 82 | 77 | 71 | 60 | 55 |
| **$\alpha$=0.05**, **$n$=200**, maintaining the null hypothesis up to and including 112 | | | | | | | | |
| $w$ | 95 | 95 | 94.5 | 94 | 92 | 88 | 70 | 58 |
| P($H_1$\|$k$=113) | 0.2 | 20 | 77 | 81 | 84 | 83 | 71 | 62 |
| **$\alpha$=0.01**, **$n$=200**, maintaining the null hypothesis up to and including $k$=115 | | | | | | | | |
| $w$ | 99 | 99 | 98.6 | 98 | 97 | 95 | 80 | 67 |
| P($H_1$\|$k$=116) | 5 | 75 | 94 | 95 | 94 | 92 | 77 | 65 |
| **$\alpha$=0.001**, **$n$=200**, maintaining the null hypothesis up to and including $k$=122 | | | | | | | | |
| $w$ | 99.9 | 99.9 | 99.7 | 99.6 | 99 | 97 | 81 | 64 |
| P($H_1$\|$k$=123) | 87 | 99 | 99.5 | 99.4 | 99 | 97 | 84 | 71 |
| **$\alpha$=0.05**, **$n$=1000**, maintaining the null hypothesis up to and including $k$=526 | | | | | | | | |
| $w$ | 95 | 95 | 95 | 95 | 95 | 95 | 87 | 75 |
| P($H_1$\|$k$=527) | 0 | 0 | 0 | 0.1 | 9 | 60 | 80 | 70 |
| **$\alpha$=0.01**, **$n$=1000**, maintaining the null hypothesis up to and including $k$=537 | | | | | | | | |
| $w$ | 99 | 99 | 99 | 99 | 99 | 99 | 93 | 80 |
| P($H_1$\|$k$=538) | 0 | 0 | 1 | 7 | 69 | 93 | 90 | 80 |

Table 4: Posterior truth-proportions for different values of $n$, $\tau$ and $\alpha$ (all data in percent)

First, table 4 shows that an increase in the numbers prima facie also leads to an improvement of the posterior truth-proportions similar as a reduction of the α-error. However, the problem that we know from Table 2 is not entirely avoided by larger numbers and a smaller $\alpha$-error. In addition, the in-

crease in numbers still shows a new problem for the situations of large effects. If the presumed probability according to our target hypothesis for heads is about 0.7 then the significant datum of 113 times heads at 200 tosses or 527 times heads at 1000 tosses speaks intuitively in favor of the null hypothesis, since if the target hypothesis is true we would on average expect 140 times heads in the first case or in the second case 700 times heads. Therefore, the data are closer to the results we normally expect if the null hypothesis is true than if the target hypothesis is true. Nevertheless, in significance tests the null hypothesis is regarded as falsified by these data.

This problem is also reproduced in the corresponding Bayesian probabilities and correlates to the *paradox of Lindsey*. But these data speak according to significance testing in against the null hypothesis and therefore are *misinterpreted* by the Fisherian methodology. This is once more the case because significance testing is not really comparative. It only asks if the data speak against the null hypothesis, but not whether they actually are in favor of a specific target hypothesis. We can call it the *problem of misinterpretation for large effects and with large numbers n*. More trustworthy are significance tests for smaller numbers of trials. We can see this problem in the following diagram:

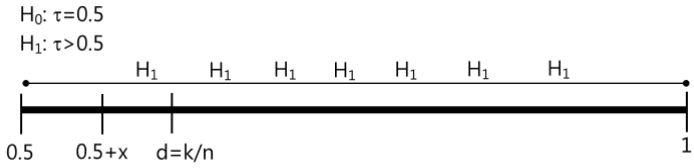


Figure 2: Distances of data and the two hypotheses

In a significance test we obtain a significant test result $d$ if $d$ is greater than the acceptance threshold $0.5+x$ (with some $x > 0$) for the null hypothesis. But by increasing $n$, $x$ will become smaller and smaller. If the true hypothesis $H_1$ is more on the right side, then we can see that (by increasing $n$) an increasing number of data $d$ are nearer to $H_0$ than to $H_1$ but are nevertheless significant. These data are misinterpreted by significance testing. They intuitively speak in favor of $H_0$, which is represented in the diagram by the greater distance to $H_1$ than to $H_0$.

In consequence, it seems that the significance testing methodology is caught in a dilemma: For small effects it is subject to the danger of an over interpretation of the data and for large effects we find even a wrong inter-

pretation of the data—and this is the case for the favorable situation that every second hypothesis tested is a true one. This is a powerful argument for, at first, estimating the initial plausibility of the hypothesis and, at second, estimating the size of the effect and on this basis estimating the posterior truth proportion. At a minimum, the examples have proven the posterior truth proportion to be a better indicator of the strength of the confirmation than the p-value. In addition, the problems strongly suggest using other evaluation methods such as Bayesian updating at least parallel to the significance tests.

The problematic is supported by an example that Beck-Bornhold and Hans-Herrmann Dubben (1996) have published in Nature, but which was not always interpreted correctly. A newer version of it is to be found in (Beck-Bornholt & Dubben 1998). We choose in this example as a null hypothesis $H_0$ that Jim has played in a lottery and as the target hypothesis that this is not the case. We now get the evidence $E$ that Jim was paid the first prize of the lottery. Intuitively, we can conclude from $E$ that Jim has (probably) played in the lottery. But, we further assume that the chance to win the first prize in the lottery is only one out of a million. Therefore the p-value for our evidence $E$ is: $P(E|H_0) = 0.000001$, and from the logic of significance testing (corresponding to a probabilistic Modus Tollens, compare Greco 2011) we should reject the null hypothesis. The problem once again results from the fact that significance testing is not comparative. Although $E$ has a small probability on assuming the null hypothesis it has an even much smaller probability if we think the target hypothesis is true. Such a capital error of the lottery company (to pay the first prize to someone who has not even participated in the lottery) has certainly taken place less often than once in a million. Therefore, the resultant likelihood ratio (or the resultant Bayes factor) is a far better indicator that speaks clearly in favor of our null hypothesis. The Fisherian methodology is obviously not designed for such cases, but it is noticeable that significance testing is the only approach from the three approaches we consider that is not appropriate in drawing the right conclusions from the evidence in this situation.

The point is that in this example we accept, at first, an information of the form $H_0 \vee H_1$ with two distinct hypotheses $H_0$ and $H_1$, both of which determine the likelihoods for many data. Then it seems obvious that we should base a comparison of the two hypotheses on the likelihoods $P(E|H_0)$ and $P(E|H_1)$. Prima facie the hypothesis *explains* the data $E$ better for which $E$ is

more probable (cf. Strevens 2000), and it therefore fits the data $E$ better than the rival hypothesis. The problem with the Fisherian methodology is that it, firstly, cannot use the special information at hand ($H_0 \vee H_1$) and that it, secondly, cannot really compare how the data fit the first *and* the second hypotheses. In many situations in science we agree on a certain finite list of hypotheses and assume one of them to be the correct answer to a scientific problem, but the significance testing methodology cannot use this background knowledge and cannot concentrate on a comparison between the hypotheses, which seems the natural strategy for a further evaluation of the hypotheses.

Significance tests seem to be more apt for cases in which one concrete point hypothesis (the null hypothesis) determines the likelihoods for many data but the rival hypothesis (our target hypothesis) is, e.g. of an disjunctive type (or an interval hypothesis) and provides no particular likelihoods for the data. Then only two options remain: First, we can try to falsify the null hypothesis or, second, we can work with (meta-) probabilities for the disjuncts of the target hypothesis (or a prior-density) in order to calculate likelihoods for the target hypothesis. These additional probabilities for the parts of the target hypothesis are normally epistemic probabilities as in the Bayesian approach, which the classical statistician wants to avoid. He thus has to bite the bullet in these cases and live with the problems of significance testing that we have acknowledged above and can additionally use the estimation of the posterior truth proportion.

For the cases of clear misinterpretation of the data he can argue that these critical data will occur only with a very low frequency. Only in these rare cases will the test procedures mislead us. Nevertheless, significance testing cannot reconstruct these situations correctly in which the data speak in favor of the null hypothesis and if in such cases the data are improbable from the viewpoint of the null hypothesis it will misinterpret them. We have to see this deficit of the Fisherian methodology clearly to avoid systematic errors in our inductions from certain data.

Greco (2011), who defends the Fisherian methodology, even argues that the cases with large effects occur only seldom. He believes that to be the case at least for hypotheses about continuous magnitudes which are rather evenly distributed over a certain region. For hypotheses about proportions, on the other hand, he suggests a maximum likelihood comparison, but does not give the details of how the comparison should proceed exactly

and in which cases we should accept certain hypotheses. Furthermore, for small effects the problem remains that the data are often over-interpreted.

Nevertheless, the problem of misinterpretation for larger effects seems to be tolerable if the posterior truth-proportion remains high for significantly confirmed hypotheses in these cases. If $H_0$ is true it will only seldom be rejected. However, a more accurate analysis of the data is provided by the Bayesian methodology especially for the cases in which different determinate (point) hypotheses concur as in the lottery example. But, to be sure, the debate will continue.

## 6 Conclusion

The filter analogy facilitated a direct frequentist assessment of the support significant data give to some target hypotheses. It reveals some weaknesses of significance testing and indicates how we can improve it. Firstly, significance testing is just as dependent on an assessment of the prior plausibility of hypotheses as the Bayesian methodology. If the prior plausibility is low as in the example of the precognition hypothesis of Bem, the significantly confirming data result only in a weak (or even negligible) support for the hypothesis. In any case, we achieve only a weak incremental confirmation and not the absolute one required in classical epistemology for accepting the hypothesis. But if we even only send sufficiently plausible hypotheses in the filter, the Fisherian methodology shows still further problems. We run into a dilemma: If we have only small effects significant data only resulting in low posterior truth-proportions, i.e. the data being often over-interpreted, and if we deal with large effects, the posterior truth-proportions may be sufficient but nevertheless we find an obvious misinterpretation of the data.

In the first situation the filter looses its proper function because the $\beta$-error has become so great that the number of false target hypotheses that will pass the filter is critical relative to the $\alpha$-error. That seems to be a problem for the precognition hypothesis even if we would attest it a sufficient prior plausibility. This result corresponds to the objections of Wagenmaker et al. (2011) who obtains only low values of the Bayes factor for Bem's hypothesis.

In the second situation, in which the effects are greater, we have to deal with the problem of misinterpretation. For $\alpha = 0.05$ and $n = 200$ a result of

113 times heads is already significant and we can reject the null hypothesis. But if the real probability for heads is 0.7 this datum obviously speaks more in favor of the null than of the target hypothesis P(head) = 0.7. The significance test cannot deal with this problem satisfactorily because it is not really comparative. Even smaller $\alpha$-errors as $\alpha = 0.01$ (or two-sided tests) cannot completely solve the problem. The need for a comparative analysis is also reflected in the lottery example.

Consequently, to assess the confirmation by a significant result, we must firstly discuss the prior plausibility of the target hypothesis and secondly must determine the posterior truth-posterior based on estimates of the effect size. Furthermore, it seems helpful to make some Bayesian calculations alongside and use, e.g., the Bayes factor as an additional indicator of the support given to the target hypothesis by the data. Altogether, we should not only consider whether the data are significant in the sense of Fisher for the hypothesis, but we have to provide a more comprehensive assessment of the confirmation the data give to our target hypothesis, which also includes an analysis of the intuitive explanatory power the hypotheses have with respect to the data, and we have to consider how good the explanations are that alternative hypotheses can provide.

## Bibliography

Alcock, J. (2011): Back from the Future: Parapsychology and the Bem Affair, *Skeptical Inquirer*: URL: http://www.csicop.org/specialarticles/show/back_from_the_future.

Bartelborth, T., *Die erkenntnistheoretischen Grundlagen induktiven Schließens*, E-Book 2012, URN=http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-84565.

Beck-Bornholdt, H. P. & Dubben, H. H. (1996): Is the pope an alien?, *Nature*, **381**, S. 730.

Beck-Bornholdt, H. P. & Dubben, H. H. (1998): *Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken*, Hamburg: Rowohlt.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anoma-lous retroactive influences on cognition and affect, *Journal of Personality and Social Psychology*, **100**,1–19.

Bem, D. J. & Utts, J., & Johnson, W. O. (2011): Reply: Must Psychologists Change the Way They Analyze Their Data? *Journal of Personality and Social Psychology*, **101**, 716 –719.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science* **2**, 317-352.

Campbell, S. & Franklin, J. (2004): Randomness And the Justification of Induction, *Synthese*, **138**, 79–99.

Feller, W. (1968): *An Introduction to Probability Theory and its Applications*, Volume I, 3rd edition, Wiley.

Fisher, R. A. (1935a): *The design of experiments*, Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935b). The logic of inductive inference, *Journal of the Royal Statistical Society*, **98**, 39-82.

Fisher, R. A. (1958). *Statistical methods for research workers*, (13th ed.)., New York: Hafner.

Fitelson, B. (2011): Favoring, Likelihoodism, and Bayesianism, *Philosophy and Phenomenological Research*, **83**, Issue 3, 666–672.

Franklin, J., (2001): Resurrecting logical probability, *Erkenntnis*, **55**, 277-305.

Galak, Jeff & LeBoeuf, Robyn A. & Nelson, Leif D. & Simmons, Joseph P. (2012): Correcting the past: Failures to replicate psi, *Journal of Personality and Social Psychology*, **103**, 933-948.

Greco, D. (2011): Significance Testing in Theory and Practice, *Brit. J. Phil. Sci.*, **62**, 607-637.

Rouder, J. N. & Speckman, P. L. & Sun, D., Morey, R. D. & Iverson, G. (2009): Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, **16** (2), 225-237.

Royall, R. M. (1997): *Statistical Evidence: A Likelihood Paradigm*, New York: Chapman & Hall/CRC.

Strevens, M. (2000): Do Large Probabilities Explain Better? *Philosophy of Science*, **67**, 366–90.

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

Wagenmakers, EJ & Wetzels, R. & Borsboom, D. & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of personality and social psychology*, **100**, 426–32.

Wagenmakers, E.-J. & Wetzels, R. & Borsboom, D. & Kievit, R. & van der Maas, H. L. J. (2011a): Yes, psychologists must change the way they ana-

lyze their data: Clarifications for Bem, Utts, & Johnson (2011) manuscript: http://dl.dropbox.com/u/1018886/Clarifications ForBemUtts Johnson.pdf

Wetzels, R. & Matzke, D. & Lee, M. D. & Rouder, J. N. & Iverson, G. J., & Wagenmakers, E.-J. (2011): Statistical evidence in experimental psychology: An empirical comparison using 855 t tests, *Perspectives on Psychological Science*, **6**, 291-298.