How to use the *p*-value in decision making

(I seem to have too little paint in my cans,
it is significant, what should I do?)

Aart F. de Vos

draft 5.0, September 12, 2012

Preface

This article is a trial to bridge the gap between statistical tests and decision making. The leading simple example is a statistical test taken from a textbook for business students. That "significance when the *p*-value is smaller than .05 " is not linked to economic decisions makes it rather incomprehensible for students. There are Bayesian alternatives but these appear to be too confusing: they are hardly taught. I suggest a justification of current practice based on sequential decision making: using the *p*-value to decide whether a phenomenon needs further (Bayesian) attention, experimenting with priors and loss functions.

Positive reactions and stimulating discussions encouraged me to investigate the consequences of my setup further. With a result that still amazes me. Apart from special cases, my conclusion is that one does not need the concept "null hypothesis" to justify the use of the p−value or its Bayesian counterpart the Bayes Factor. No more hypotheses, just decisions. Subjective decisions. Concentrating upon simple models, easily explained with a spreadsheet. "Bayes light" my wife said when I explained my solution to her. That 's it. Full Bayes is too heavy.

This article is written as an essay, with formulas and reference kept to a minimum. The introduction contains nothing new for Bayesians, it serves to focus non-Bayesians on problems with their approach. Aiming at Dutch statisticians, who in general know little about Bayes. The rest of section 1 contains the basic idea. In section 2 this is worked out in a traditional Bayesian setting, involving "hypotheses that might be true", first without loss functions and showing that the introduction of loss functions matters a lot. In section 3 only loss functions remain. And the (major) problem to specify subjective priors and loss functions.Section 4 gives an example where the interplay between *p*-values, Bayes Factors, hypotheses and loss functions becomes clear.

This version is not yet completely in equilibrium. First I described what should be done and why. And speculated about the outcomes. Next I had a program made that did what I had written should be done. That program provided concrete answers to some of my speculations. They are described in section 4.

1. The p-value

Few subjects in statistics are as controversial as the *p*-value. Bayesians despise it. "No more *p*-values" sings the chorus in a play by Antony O 'Hagan, author of the Bayesian part 2b of Kendall's Advanced Theory of Statistics. This is worrying, to say the least, as the *p*-value is a central concept in virtually any book on statistics for practitioners. It has been for at least half a century. So it is hard to imagine that it is useless.

The *p*-value is formally described as the probability that, if the null hypothesis (H_0) is true, your data are as extreme as they are or even more so. It gets meaning by comparison to a norm, the significance level α, usually chosen 0.05. If $p < \alpha$ then something is "significant".

It is a common mistake, called the "*p*-value fallacy", to think that the *p*-value is the probability that H_0 is true. This is of course wrong. This probability does not even exist in classical statistics. But it is what we need to take decisions under uncertainty, so the confusion is not amazing. What must a business student (my example in this article, but it is much more general) with a conclusion "there is significantly too little paint in my cans" other than think that

you must do something because it is very likely that something is wrong? In my view the lack of a link between testing and decision making in statistics courses is a drama that makes these courses almost useless. And even dangerous. The *p*-value fallacy is a version of the prosecutors fallacy. "If the suspect is innocent (H_0) the facts are utterly unlikely, so the suspect must be guilty". The trial of Sally Clark (two of her children died, was this murder or two instances of cot death?) is a famous example.

Even students who have had extensive training in classical statistics are shocked when I give the following example: Suppose a student takes a multiple choice examination. The exam has 50 questions, with three possible answers for each question. The student answers 27 questions correctly. Did the student guess in each question or not? If the probability of guessing right is 1/3 for each question, the probability of 27 *or more* right is 0.002. This is very significant. So he did guess? With Bayes' Theorem I show a 65% probability that he guessed may be a plausible answer. 27 answers right is less likely if he studied (and raised his probability per question to 0.75) than if he guessed. It is the likelihood ratio that matters (of 27 right, *more* is irrelevant). And if there is a 50% prior probability that he had a party and did not learn but guess the computation is easy. Posterior odds is prior odds times likelihood ratio.

Should we then ban the *p*-value? No. I think I know (at least for business students) how to justify the use of *p*-value by giving it an economic underpinning. My thesis is that the classical format for significance, $p < \alpha$, can be used for coherent (Bayesian) decision making in simple models. And that simple models are needed for an optimal sequential solution for most problems as treated in textbooks. Which bridges the divide between frequentist and Bayesian approaches. And the solution is relatively simple. The *p*-value remains what it was. The significance level α follows from a preliminary model for a possible decision, including simple priors and loss functions (I use the term loss in this article, utility would be more accurate and general, but a bit confusing)

Priors are subjective, loss functions mostly too And most decision makers feel uncomfortable when forced to specify them. Classical statisticians, following Ronald Fisher (1935), argue that it is an illusion that we are able to do so. Which is the main argument against the Bayesian approach. I think that we are able specify priors and loss functions approximately. Uncertainty about them leads to uncertainty about the relevant α. But I think that we often may make a confidence interval for α : α^- to α^+. In section 4 I will show how this can be done in an example that is applicable in many cases.

The decision procedure then consists of three steps.

The first step is to design a decision rule, a choice of α^- and α^+

1. Choose an upper bound for the significance level α^+ that makes sense in the given context.

The second step is well known:

2. Determine whether your data show significance, which is the case if $p < \alpha^+$. Followed by a decision that goes right to the heart of the concept of significance and may be paraphrased as follows:

3. If it is not significant: don't think further. If it very significant ($p < \alpha^-$): act. Else ($\alpha^- < p < \alpha^+$): hire a Bayesian.

Like in health care. The general practitioner has three diagnoses for a patient with a medical problem: innocent, suspect or wrong. If suspect he refers to a specialist. And with "innocent" he means that the probability that something is seriously wrong is too low to justify the cost and trouble of advanced diagnosis. "Hire a Bayesian" is similar. In most situations a full Bayesian analysis is difficult and thus costly. It requires formulating (partly subjective) priors and loss functions and evaluating the different possible decisions in a much more detailed way than in the first step.

The typical setting where our procedure applies is when it is a priori unlikely that the situation is much different from the normal one, traditionally described as the null hypothesis H_0, and potential losses by taking wrong decisions are within reasonable boundaries. Then the classical test is a kind of self protection. Continuously we are bothered by "significant" results

that may be due to something real but in most cases it appears to be of little importance. A full analysis of the facts requires lots of time. And the money this requires is waisted if turns out that there is not enough evidence to change our behavior. That is the case in a fraction $\pi(H_0)(\alpha^+ - \alpha^-)$ of all cases, with $\pi(H_0)$ the prior probability that nothing is wrong, when using the rule "think if the p-value is between α^- and α^+".

The classical test says "reject H_0 if $p < \alpha$". With $\alpha = 0.05$. Or sometimes 0.01 for reasons that intuitively make sense but no more than that. Students obediently learn how to specify a test statistic (S) with a known distribution so that they can compute the p-value of the "realized S". And after hard work and some failures most of them pass the exam. It would be nice to investigate whether the main reason that students find statistics difficult is the mathematics or the lack of clear interpretation of the concept significance. Teachers are more often than not mathematicians who are used to manipulate abstract concepts without reference to the real world. But for normal people this is different.

The key to this article is that most (if not all) one-sided tests of the format "significant if $S > k$" (or $S < k$, depending on the situation) also have a Bayesian interpretation. What differs in then is the motivation to choose k and the interpretation of significance. In a fully specified Bayesian setup significant means "do something" (to minimize expected loss). The decision rule can conveniently be rewritten as "do something if $p < \alpha_{im}$" With p the p-value that only depends on the data and α_{im} the "implicit α" which depends on priors and loss functions.

A very confusing aspect of the statistical literature is that there are three types of Bayesians. All Bayesians use priors, but not all Bayesians use loss functions. And Bayesians who use both are split into Bayesians that use subjective loss functions (different among decision makers) and "objective Bayesians" who search for loss functions with some form of objective validity (Bernardo(2011)).

My position goes back to Savage(1953) who showed that decision making under uncertainty requires the specification of priors and utility (loss) functions. Both subjective.

The school without loss functions, going back to Jeffreys (1939), concentrates upon truth finding: the probability that H_0 is true. With Bayes ' rule one can compute how prior belief in H_0 is changed by the data into posterior belief by the "Bayes Factor" (see section 2). The Bayes Factor is seen as an alternative for the p-value. And as an alternative for the classical "$\alpha = 0.05$ is significant, $\alpha = 0.01$ is very significant" Jeffreys suggested as measure for evidence against H_0 (the Bayes factor is measured in odds against H_0): "3:1 to 10:1 Substantial 10:1 to 30:1 Strong, 30:1 to 100:1 Very strong >100:1 Decisive". And Wikipedia just quote Jeffreys, suggesting that, just like classical statisticians stuck to Fisher 's α's, Bayesians stuck to Jeffreys' guesses. I hope to show that one can do better.

The idea of an "implied α" goes back to Andrews (1994). He proves that (asymptotically) a Bayesian posterior odds test (not involving loss functions) is equivalent to a classical test of some size (α), also for cases where there is no single statistic S. I think these proofs also apply when loss functions are involved. Also Berger (2003) and the authors who quote this article show correspondences, but not in the format of the p-value.

I will (in a setting less general than that of Andrews) show that if one does include loss functions and concentrates upon decisions, the format $p < \alpha_{im}$ remains valid, and that these loss functions are a major determinant of α_{im}. If the cost of action is high, you need more evidence before you act.

For decision rules in the format $p < \alpha$ one can conclude that the main difference between frequentists and Bayesians is that the latter have a recipe to determine α.

One of the reasons why Bayesians do not use this format is that their focus is not upon decision rules. Bayesians condition upon the data. Decision rules concern data before they are observed, which has a frequentist flavour. The distribution of the data under H_0 (needed to determine α) is of no concern.

But it makes sense if one accepts the usefulness of a decision rule, if only to make a "quick and dirty " decision in similar situations whether it is worth while to think further, the representation in terms of p and α has two important advantages. The first is that p depends only

upon H_0 and the data, and α on priors and loss functions. This implies that differences in opinion between decision makers may be summarized as differences of opinions about the relevant α. The second advantage of using the format $p < \alpha$ (there are equivalent alternatives) is that p has a nice metric (between zero and one), and clarifies the discussion with frequentists.

The coherent way to describe the decision situation, at least in the first step, is in terms of a given p-value and a prior for α. This in turn implies that in sequential decision making it is an option to obtain more information about α (hire a Bayesian). This should lead to less uncertainty about α and ideally to a completely specified problem which is equivalent to a known value of α.

The basic question in the first step is: how to formulate a prior for α without much trust in priors and loss functions? My answer is: start with a simple model that contains the most important inputs. Make a sensitivity analysis showing how α_{im} depends on the context. And hope that this, combined with practical experience with the setup will lead to reasonable choices. The result will easily lead to an improvement on the decision rule that always takes $\alpha = 5\%$ without clear motivation. The example I give in section 4 convinces me that this is quite well possible.

2. The Bayes Factor and the fixed loss case

The Bayes Factor is the Bayesian alternative for the p −value. But it has always been controversial. Asymptotically it has nice properties, but in small samples the outcome depends heavily on a prior distribution. Loss functions are not involved in the Bayes Factor, but fixed loss of wrong decisions may be incorporated. This leads to an α_{im}.

We only treat the simplest case where there is only 1 parameter θ that matters, H_0 is $\theta = \theta_0$, and for the time being the alternative is one-sided, so H_1 is $\theta > \theta_0$ (or $\theta < \theta_0$). And one statistic S gives a summary (or good approximation) of the statistical information. For many standard text examples there is even an exact Bayesian justification to use the well known (e.g. normal or student distributed regression parameters) statistic (by integrating out nuisance parameters), for other cases one may disagree about details, but these are of secondary importance.

The standard Bayesian approach is to assume prior probabilities that hypotheses are true. Formulated as $\pi(\theta = \theta_0) = \pi(H_0)$. In case of a continuous distributions that is not undisputed, but may be seen as a convenient way to express that there is a priori some degree of belief that θ is close to θ_0. A necessary ingredient to justify something equivalent to a statistical test if one does not wants to use loss functions. In section 3 I will argue that it is no longer necessary if one uses loss functions that depend on θ, but even then it may reflect the state of mind of a decision maker. In the example of section 4 it is an option to investigate the effect of specific probability mass near θ_0.

But as the Bayes Factor is the standard and provides some useful insights, I treat this setup first.

It is possible to use the Bayes Factor for the case of simple fixed loss functions when taking the wrong decision. The formula for the inequality prescribing when to take action because the expected utility is higher than doing nothing is: act if

$$BF(S) = \int \frac{f(S|\theta)}{f(S|\theta_0)} \pi(\theta|H_1)d\theta > \frac{\pi(H_0)}{\pi(H_1)} \frac{L(1,0)}{L(0,1)} = \kappa \qquad 1$$

Here $f(S|\theta)$ is the density of S; $\pi(\theta|H_1)$ the (proper) prior density of θ if H_1 is true; $\pi(H_i)$ the prior probability that H_i is true. $L(1,0)$ and $L(0,1)$ are expected losses if the wrong action is taken (see below)

On the left side we have the "Bayes Factor" ($BF(S)$). That is the ratio $P(S|H_1)/P(S|H_0)$, the information from the data (S) about the probability that H_0 is true ($P(H_0|S)$, implying $P(H_1|S)$, they sum to 1). It depends on $\pi(\theta|H_1)$, the (subjective) prior for θ if something is wrong.

On the right side we have the prior odds $\pi(\theta|H_0)/\pi(\theta|H_1)$ times the loss ratio. $L(1,0)$ is the loss of doing something while there is nothing wrong. Obviously this depends on the decision one has in mind. Often that will be taking another sample. $L(0,1)$ is the expected loss of taking no action while something is wrong.

The Bayesian posterior odds test has the format of (1): $BF(S) > \kappa$. Including fixed loss functions we still have that format. It has the drawback that both sides of this inequality depend on priors. And the right side on loss functions as well. Which implies that Jeffreys' suggestions to use the Bayes Factor as a measure whether S is worth mentioning without reference to loss functions can only be justified by assuming that $L(0,1) = L(1,0)$. This can be accommodated. If 3:1 is the ratio above which S becomes worth mentioning if the losses are equal, $6:1$ is needed when $L(1,0)$ doubles.

The implicit α in (1) is

$$\alpha_{im} = P_{S|H_0}(BF(S) > \kappa), \qquad\qquad 2a$$

an unusual mixture of frequentist ($P_{S|H_0}$) and Bayesian statistics. Notice that α_{im} does not depend on the observed S. The same choice of α can be used for any realized S in the same circumstances. In de Vos and Francke(2008) the relation between κ and α_{im} is elaborated further.

If $L(0,1)$ depends on θ, say $L(\theta)$, things are more complicated. The expected loss then depends on the posterior $p(\theta|S)$. So on the observed S. We define it as $L(0,S) = \int p(\theta|S)L(\theta)d\theta$. Using $L(0,S)$ instead of $L(0,1)$ in (1), we get a test of the form $BF(S) > \kappa(S)$. Under mild conditions this can be rewritten as $S > k$. Then

$$\alpha_{im} = P_{S|H_0}(S > k), \qquad\qquad 2b$$

again independent of S. And the break even point $S = k$ has an associated $BF(k)$. So one may formulate decision rules in the format of an "implied Bayes Factor". But I prefer the p-value because that only depends on the data. I will come back to this in the example of section 4 where the whole situation will become more clear.

How alpha depends on loss functions. An example

Before I give a full example how to choose α I give an example how to use (1) to make different decisions mutually coherent in the case of fixed losses. Suppose I am a paint manufacturer and there seems to be not enough paint in my cans. If that appears to be true on average I can get a fine. The question is: must I take another sample? Once I hired a Bayesian who computed that I should do so if my p-value is lower than $\alpha = 0.05$. But now the price of that sample doubles. So in (1) $L(1,0)$ doubles. That logically implies a lower α_{im}.

Take θ the downward deviation of the mean (the norm (say 1 liter) minus the true mean) and assume that the relevant statistic S (the mean deviation from the norm in the sample) has a normal distribution with mean θ and unit variance. H_0 is $\theta = 0$. We use the prior $\pi(\theta|H_1) = \exp(-\theta)$, an exponential distribution. From (1) and (2) we get a table. I only give some interesting values: ($\theta_0 = 0$; S the sample result; S_0 a possible result $|H_0$)

S	0.91	1.5	1.645	2	2.193	2.63
$BF(S)$	1.12	1.91	2.23	3.43	4.46	8.92
$P(S_0 > S|H_0)$	0.181	0.067	0.050	0.023	0.014	0.006

The $\alpha = 0.05$ occurs for $S = 1.645$, a familiar result. Given our assumptions the corresponding break even point for the Bayes Factor is 2.23. If $L(1,0)$ doubles according to (1) this value doubles, so becomes 4.46. The table shows that this corresponds with $\alpha_{im}= 0.014$.

If $L(0,1)$ doubles, because the fines if something is wrong double, the relevant value of $BF(S)$ becomes 1.12, so $\alpha_{im} = 0.181$. Doubling (from the default situation) the prior odds in favour of H_0 justifies $\alpha_{im} = 0.014$. And then doubling $L(1,0)$ as well, $\alpha_{im} = 0.006$ is justified. Rather large shifts compared to what one would expect after a classical training (where formally these shifts do not exist but informally a choice of α depending on the circumstances is advocated).

In section 4 I work with losses that depend on θ. There too, the loss functions appear to have a big impact on α_{im}.

Two sided tests

If the sample raises suspicion that there is (on average) not enough paint in each can, with a

one sided p-value of 0.03, should I abstain from doing something if I would also would do something when there was too much paint? Economists learn to multiply the p-value with 2 if they do a two sided test. This is rather funny. If significance testing is embedded in a decision context there is a natural solution. There are two break-even points for S that make one decide to do something. If the sample indicates that there is too much paint in the cans the break-even point (say l) depends on the loss function on the other side of θ_0, one must act if $S < l$, and this corresponds with an α_{im} for the other tail area. When the loss functions are different, the solution is asymmetric. See section 4.

The difference with the classical test (two one sided tests instead of one two-sided tests) is a result of conditioning upon S. In this case we condition upon the fact that S is below or above the value that corresponds with θ_0. By only considering the resulting two break-even points we do not condition further upon the value of S, what a full Bayesian analysis would do.

The literature on Bayes Factors in relation to two sided tests is rather confusing, I think. I will not go into this. The quite satisfactory results I obtain in section 4, that correspond with what is written above, show that no problems arise if one works with decisions and loss functions.

Generalizations

In de Vos and Francke(2008) we show that the existence of a single sufficient statistic S is not essential. If there is one statistic S that is maximal invariant with respect to nuisance parameters (of place and scale), tests in the format $p < \alpha$ are uniformly most powerful invariant (UMPI). And there is an exact correspondence with Bayesian decision rules if Jeffreys' independence (noninformative) priors are used for the nuisance parameters. For unit root tests that we treat there, there is no such S, but the more general concept of marginal likelihood appears to provide tests that almost reach the UMPI upper boundary. This is shown for the case with fixed losses. That it also holds if $L(0, 1)$ depends on θ remains to be proved. The results from Andrews(1994) suggest that similar results hold approximately for other types of nuisance parameters.

3. Decisions without hypotheses

It took a while before I, after reading Bernardo(2011), realized that working with loss functions and decisions, hypotheses are no longer needed. If we have a prior $\pi(\theta)$, loss functions $L(0, \theta)$ for doing nothing and $L(d, \theta)$ for one or more decisions d, and a statistic S, the ingredients are there to take the decision that minimizes the posterior expected loss (more expected utility)

$L(d|S) = \int L(d, \theta)p(\theta|S)d\theta$

And, under reasonable regularity conditions, we can compute the break-even point k such that we decide to do something if $S > k$. And we can transform this to $p < \alpha_{im}$. Extension to two sided "tests" is trivial. If too much paint in the cans also involves losses a decision rule for low values of S results. Which also can be written in the $p < \alpha$ format for the other tail. With a different α if the loss functions are asymmetric.

So we do not need the concept "null hypothesis " to take decisions. We may ask the question: why H_0?

On the origin of the null hypothesis

Decision based upon priors and loss functions may be written in the format $p < \alpha$. Where p depends on the data and α on the priors and loss functions.

What to do if one does not want to specify priors and loss functions because that is very difficult and/or subjective? How to specify then a justification of a sensible rule to guide you? Where sensible is the format $p < \alpha$. With p the message from the data.

That's the problem that was faced by classical statisticians. Fisher (the design of experiments, 1935) found an ingenious solution. He introduced the concepts "null hypothesis "and "significance". And $\alpha = 0.05$ came out of thin air with an appeal to common sense by requiring that one should avoid the error of the first kind: calling something significant while H_0 is true.

Once one realizes that the same recipe can be motivated with priors and loss functions, the

most amazing aspect of the wide acceptance of Fisher's solution is that abstract concepts are preferred over concrete though disputable ones.

Anyhow, in many non-experimental and decision oriented settings, I think it might be better to reject the "null hypothesis" as a useful concept. It may be an unnecessary and confusing ingredient.

Bayes Factors revisited.

The precise null hypothesis $H_0 : \theta = \theta_0$ is used by Bayesians as something that might be true. With prior probability $\pi(H_0)$. Mathematically this works. I used it in my equations. But it is a bit problematic. In a continuous setting the probability that $\theta = \theta_0$ is zero. It is possible to introduce a zero point mass, as a convenient representation of probability in a small interval near θ_0 but the separation of the prior for θ into two parts remains rather arbitrary. And is not necessary, as long as the loss function only involves θ. Only $\pi(\theta)$ matters, and if it is given, it can be split as $\pi(\theta) = \pi(H_0)\pi(\theta|H_0) + \pi(H_1)\pi(\theta|H_1)$ in an infinite number of ways. Each leading to the same answer, so not incorrect in itself. The crucial question is whether one is able to ask practical questions that allow practitioners to give a sensible split.

So the Bayes Factor format

$$\frac{P(H_1|S)}{P(H_0|S)} = \frac{\pi(H_1)}{\pi(H_0)} \frac{\int p(S|\theta)\pi(\theta|H_1)d\theta}{\int p(S|\theta)\pi(\theta|H_0)d\theta}$$

(posterior odds is prior odds times Bayes Factor) has two drawbacks. The split between H_0 and H_1 is often artificial and it does not provide the solution to the decision problem if loss functions are involved.

An exception may be the situation where the hypotheses correspond to physical states. Suppose my paint filling machine is calibrated on θ_0. Calibration cannot be perfect so the result is a $\pi(\theta|H_0)$ concentrated around θ_0 (which will correspond to slightly more than a liter to avoid underfilling). Something can change or go wrong, leading to a $\pi(\theta|H_1)$. One can construct the mixture prior $\pi(\theta) = \pi(H_0)\pi(\theta|H_0) + \pi(H_1)\pi(\theta|H_1)$ then. And given S the posterior odds for the two states. But this is relevant only if the losses do not only depend on the value of θ, but on the reason why θ deviates from θ_0 as well.

Bernardo(2011) also treats testing as a decision problem. And avoids the use of probability mass concentrated in θ_0. Acting as if $\theta = \theta_0$ is in his framework the relevant decision. That is rather similar to our doing nothing, or rather not bothering. But he derives fixed "objective" decision rules while my approach is subjective with different outcomes depending on the circumstances. Among many comments on Bernardo 's article, there is one by Dennis Lindley, one of the godfathers of Bayesian inference. He writes:

"My view is that the scientific method, and statistics as part of that method. is fundamentally subjective, objectivity only appearing when scientists reach agreement. I therefore argue in favour of statistical procedures that are based on subjective probabilities: probabilities that reflect your beliefs."

I could not agree more. But the problem remains how to find representations of real life problems such that prior beliefs can be incorporated.

My experience is that the best way to reveal priors is to ask people in what situations they are confident that they can take the right decision. A model with priors and loss functions should reproduce those decisions. Asking for the probability that a hypothesis is true is asking for trouble.

Nevertheless I use in the example of section 4 probability mass around θ_0. If only to investigate the effect. Which appears to be substantial. The question remains: what questions should one ask practitioners to reveal specific prior information related to the area close to θ_0?

Does it matter whether hypotheses are true?

"Truth is nothing but an opinion held by all", I remember from an article by Ed Leamer.

Maybe it is not even an opinion, but just a way to motivate our behavior. It is tempting to think further on hypotheses and decisions. We do not need hypotheses to make decisions. We just don't bother about a lot of things because they are not important and/or unlikely. If my general practitioner says that my complaint is not a sign of some severe illness, I stop bothering. The reverse holds too: though only one scientist says that taking an aspirin a day prevents cancer it changed my behavior: aspirin is cheap, there are hardly side effects and potential losses are huge.

I was involved in two Bayesian analyses that may be interesting in this respect. I reread them after writing the above. In de Vos (1987) I analyzed (in Dutch) the link between prone position of babies and cot death. A dutch physician had found a significantly higher incidence of cot death when babies slept in prone position, which was advocated at that time. It was an accidental discovery, the data had not been gathered to investigate this link, and it was certainly not a controlled experiment. But it was published, and a debate resulted whether this was justified. In my analysis I motivated the choice of priors, combined the outcome with the fact that prone position and cot death had both increased over the years, incorporated doubts on the causality and the non-representative sample. A series of posteriors, changing with each step, for the ratio of cot death risks in prone and other positions was the result. Ending with the advise to avoid prone position for the time being. It was a long story, with many debatable assumptions, without sharp null hypotheses. Just a mathematical representation of all relevant thoughts. No more, no less. It is 34 years later now. Many studies have appeared since, most of them confirming the statistical link but none finding a really satisfactory explanation. But prone position of babies is discouraged world wide. A good example of "start thinking when you find something significant". And in this case of "take a preliminary decision even if there is a probability of say 50% that it is not true " as well.

In Tol and de Vos(1998) we gave a Bayesian analysis of the link between the concentration of carbon dioxide and the temperature on earth. Focusing upon the significant coefficient for CO_2 concentration in regressions explaining the rise of the temperature during a century. Again a long story with many debatable assumptions. Too much to summarize here, and too little to mention as so much has been written on this subject. But the correspondence with the cot-death case is striking. World wide efforts are made to reduce CO_2 emission, but doubts remain and research continues. And what about the question "what is the probability that CO_2 does cause global warming?"

I once explained a brilliant mathematician the example I started with. The student with the multiple choice questions (by the way: an example of two physical states with corresponding hypotheses). What is the probability that he gambled?, I asked.

The next day I met him again. He had discussed for hours with a friend about this problem. And his answer was:

You may not ask that.

4. A spreadsheet that explains much.

After writing a previous version of this article a had a spreadsheet program made for illustration. All elements of the discussion are there. The p-value, the Bayes Factor, the prior and the loss functions. And one can change many parameters to see how they affect the decision rule. α_{im} is computed for each combination of relevant inputs.

The crucial question is whether playing with scenarios one gets enough confidence to choose boundaries for α : α^- and α^+. According to me that seems quite well possible, and less far apart than I had expected.

I hope it can be the basis of a system that can be used by practitioners. That may require an interface that transforms inputs to categories like: large deviations are "unlikely ", "very unlikely " or "extremely unlikely ". Pressing the right buttons and giving some rough estimates of cost should provide the practitioner with sensible choices of α as required for step 1. The formulation of the right questions to reveal priors is the most difficult task. The whole setup depends heavily on the ability to do so.

One must keep in mind that the goal is only to provide reasonable values. Exact optimization is not possible and not needed. The expected loss as a function of α is likely to be rather flat in the neighborhood of the optimal value. So the consequence of a choice that is approximately good will hardly differ from the optimal one. Avoiding stupid choices is the goal. This is also the reason why a simple model is sufficient in most cases.

The Inputs

The spreadsheet program represents the most basic example of statistical inference: A sample mean that has a normal (or student) distribution. And priors and loss functions one can play with.

The decision rule we want to study might refer to a sample of the content of cans of paint. And we take the decision whether or not to interrupt the filling process and recalibrate the filling machines, which involves cost, independent of the mean deviation θ. Doing nothing involves expected losses that do depend on θ. For $\theta > 0$ (too little paint) we take a quadratic loss function. For $\theta < 0$ a linear one (which makes some sense in this case).

Many problems have the same structure. Doing something involves cost, independent of θ. Doing nothing may involve losses, in most cases increasing with the distance of θ from θ_0.

The scaled likelihood function and the p-value

Figure 1 provides a print of a specific parameter constellation. The upper picture gives prior, scaled likelihood and posterior for a given value of S (the main input).

The horizontal axes represents θ. The scaled likelihood is a normal distribution of θ with an expectation S and standard deviation $\sigma = 0.1$. Which is also the standard deviation of S. Understanding this requires basic knowledge of Bayesian thinking. Which might begin with the explanation that $(S - \theta)$ has a Normal distribution with expectation 0 and

To be explained perhaps to students in the sense that $(S - \theta)$ has a normal distribution that has two interpretations: the classical S given θ and the Bayesian θ given S (and a noninformative prior). And that the p-value also has two interpretations: the classical one defined with the help of a confidence interval for S and the Bayesian one: the area in the graph left of $\theta = 0$. The fact that the number you get is always the same may help. A real good explanation however may require a three dimensional picture with θ and S on different axes.

Once one accepts this, it is clear that by shifting S, the scaled likelihood shifts and the p-value changes. Which completes the classical analysis. In fig 1 S is chosen such that the p-value is .0228. A classical "significant " situation.

The prior and posterior

The next element is the prior. This consists of a middle part, an area around zero with width w, so $-w/2 < \theta < w/2$. In the picture $w = 0.04$. The distribution is taken uniform. The height is 12.5 so the surface is 0.5. To make the correspondence clear with the Bayes Factor approach without loss functions values of θ in the interval are reckoned to belong to H_0. So $\pi(H_0) = 0.5$. The values of w and $\pi(H_0)$ are inputs.

The left and right tails are exponential, possibly with different parameters and probability mass. Here the right tail has paramer $\lambda_r = 4$, on the left side we have $\lambda_l = 3$ and the probability masses left and right are equal (so 0.25).

The prior is now complete and provides by multiplication with the likelihood function and scaling the posterior of θ. The right moment to show to students how

$$p(\theta|S) \propto p(\theta)p(S|\theta)$$

works in practice.

As $P(H_0|S) = P((-w/2 < \theta < w/2)|S)$ this is the surface of the posterior in the H_0 interval. In the picture this is $0,33$. The Bayes Factor is best explained from "posterior odds is prior odds times Bayes Factor". The prior odds are 1:1, the posterior odds 2:1, so the Bayes Factor is 2:1.

It is important to note that neither the p-value nor the Bayes Factor or $P(H_0|S)$ involve cost.

These are used in the second picture.

Loss functions

In this picture the ax represents S as well as θ. As these have the same metric this is possible and it provides useful insights.

Shown are the loss functions $L(1)$ (independent of θ) and $L(0,\theta)$. The most important curve shows the "expected cost when doing nothing " (in red).

$E[L(0,S)] \;\; = \int L(0,\theta)p(\theta|S)d\theta$

Note that the loss functions refer to θ and the expected loss functions to S.

As $L(1,\theta)$ is constant it coincides with $E[L(1,S)]$. So we define

$L(1,\theta) = E[L(1,S)] \;\; = L(1)$

(That $L(1,\theta)$ does not depend on θ is no loss of generality as only the difference between $L(0,\theta)$ and $L(1,\theta)$ matters)

The posterior $p(\theta|S)$ of the picture above combined with the curve for $L(0,\theta)$ provides one point on the $E[L(0,S)]$ curve. A kind of weighted average.

Changing S only changes the picture above. Increasing the sample size while S remains the same makes that $E[L(0,S)]$ approaches $L(0,\theta)$.

Break even points

For any cost level of taking action the intersection of $L(1)$ with $E[L(0,S)]$ provides the break even points for S. We notate them as S_{BP}. Any point S has an associated p-value, that is given by the graph below, the two S_{BP} points give the α values. Note that at $S = 0$ the relevant p-value (left or right) switches. In the picture the left intersection is at $S = -0.2$, giving a left side α of 2.28% as well. That is a coincidence. Increasing $L(1)$ (shifting the green line upward) shifts the S_{BP} points further away from θ_0 and lowers the α values asymmetrically.

The question how to treat two-sided hypotheses has a clear answer now. The effect of the prior and the loss function in "the other half " on the position of S_{BP} is negligible. This is because the likelihood function is negligible apart from an area on the left close to θ_0, but there the loss is small. So the two sided pictures can be viewed as two one sided pictures. So in the discussion we confine ourselves to the right side.

Note that the break even point for the Bayes Factor, $BF(S_{BP})$, which is on the right side 2:1 now, may differ from that on the left side. The position of the break points may differ (they do not in the picture, but that is a coincidence) and the prior may differ (like in the picture). Bayes Factors and loss functions are not a happy marriage.

If one is primarily interested in the right half side (too little paint) one only needs to consider the right hand α_{im}. But what if then a value of S occurs that seems to indicate that there is too significantly too much paint? The answer depends on the question whether one is able to specify priors independent from the fact that S has occurred (what one would have thought before S occurred?). If so, one just can construct the left hand side afterwards to see what should be done and what the decision rule for these cases should be in the future.

Explanation of some features

The dominant feature of all pictures is the connection between $L(0,\theta)$ and $E[L(0,S)]$. Cost functions matter a lot. The position of θ_{BP}, the break even point one would use if θ would be known, is a major determinant for the position of S_{BP}. For a linear $L(0,\theta)$ S_{BP} is further from θ_0 than θ_{BP}. The distance grows with the variance of S. Quadratic loss shifts S_{BP} closer to θ_0.

These features can easily be understood. For the chosen specification they can even be derived analytically, but that is no longer possible if we work e.g. with a Student distribution instead of a normal one.

The easiest feature is the influence of the loss function $L(0,\theta)$. If this is linear, only the expectation of the posterior counts. If $L(0,\theta) = a + b\theta + c\theta^2$, with $c > 0$, $E[L(0,S)]$ is $E(\theta|S) + cVar(\theta|S)$. So the $E[L(0,S)]$ curve shifts upward and $BF(S)$ shifts towards θ_0 . One has

to be more careful if one is afraid that something might be seriously wrong.

The basic result of the prior is that expected cost $L(0.S)$ are lower than $L(0,\theta)$ through the interaction between uncertainty and prior considerations. In a testing situation values of θ are a priori more likely the closer they are to θ_0. This shifts the posterior to towards θ_0.

In S_{BP}, which is the only point that is relevant for the decision rule, and in interesting situations moderately far from θ_0, two features of the prior are relevant: the exponential decay λ of the prior in the relevant half ($p(\theta|H_1, \theta > 0) = \lambda e^{-\lambda\theta}$) and the probability mass $\pi(H_0)$.

An exponential prior results in case of a normal likelihood in a shift towards θ_0 of the posterior depending on the degree of exponential decay λ. If the likelihood is negligible at the other side of θ_0, the posterior is the scaled likelihood shifted $\lambda\sigma^2/2$ towards θ_0. Otherwise the same happens, but the posterior is truncated.

A double exponential prior results in a posterior where both sides are truncated normal and the total is rescaled to unit surface.

Adding a mass point at θ_0, the traditional Bayesian approach, so $P(\theta = \theta_0) = M$

$$\frac{P(H_1|S)}{P(H_0|S)} = \frac{\pi(H_1)}{\pi(H_0)} \frac{\int p(S|\theta)\pi(\theta|H_1)d\theta}{\int p(S|\theta)\pi(\theta|H_0)d\theta}$$

The expectation of
$E[\theta|S] = P[H_0|S]E[\theta|H_0,S] + P[H_1|S]E[\theta|H_1,S]$
As $E[\theta|H_0,S] \approx 0$ only the second term counts. So
$E[\theta|S] \approx (1 - P[H_0|S])E[\theta|H_1,S]$,
and the shift in expectation by increasing $\pi(H_0)$ only depends on $P[H_0|S]$.

As $p(S|H_1) = \int p(S|\theta)\pi(\theta|H_1)d\theta$ does not or hardly change by introducing $\pi(H_0)$, the shift depends on the size of $\pi(H_0)$ and
$p(S|H_0) = \int p(S|\theta)\pi(\theta|H_0)d\theta$.

With $\pi(\theta|H_0)$ uniform on a small interval around θ_0 this is the mean level of $p(S|\theta)$ in this interval. Which gets larger as the width w increases because the likelihood function $p(S|\theta)$, in the normal case proportional to $\exp(-0.5(S - \theta)^2/\sigma^2))$, is such that values at the right side of the interval are relatively so big that they dominate the average. This makes the Bayes Factor, and so $P[H_0|S]$ as well, rather sensitive for w. An aspect to worry about. Moreover the sensitivity for the form of the likelihood function in the tail area is high.

As it is $P[H_0|S]$ that matters for the shift in $E[\theta|S]$, the prior odds for the Hypotheses are as important as the Bayes Factor. So their is simultaneous uncertainty about the guesses for $\pi(H_0)$ and w. And as increasing w changes the meaning of H_0 such that a higher value for w logically implies a higher value for $\pi(H_0)$, the uncertainties reinforce each other.

A sensitivity analysis

In table 2 we give a sensitivity analysis. The first line (default) corresponds to the picture, the next lines give the effects of changes of the input parameters.

Table 2: sensitivity of criteria for inputs							
Inputs (if not: default)				Outputs for break even point			
$L(1)$	λ	$\pi(H_0)$	w	S_{BP}	$\alpha(\%)$	$P(H_0|S_{BP})$	$BF(S_{BP})$
3	4	0.5	0.04	0.20	2.28	0.33	1:2.0
			0.02	0.21	1.79	0.28	1:2.6
	9			0.21	1.79	0.28	1:2.5
		0.2		0.15	6.68	0.22	1:1.2
		0.9		0.27	0.35	0.48	1:9.5
2				0.16	5.48	0.45	1:1.2

Noteworthy is the sensitivity of $BF(S_{BP})$ for $\pi(H_0)$. This is due to two effects. For $S = 0.20$ and $\pi(H_0) = 0.9$ the Bayes Factor is 1:4.1. The shift of S to $S_{BP} = 0.27$ causes a further rise to $BF(S_{BP}) = 1 : 9.5$.

For the position of S_{BP} (and thus α) two factors are really important: $\pi(H_0)$ and the cost level of action $L(1)$.

And now the crucial question. Can we confidently specify values α^- and α^+?

In case of cans of paints the loss functions are probably reasonably well known. The main uncertain factor is $\pi(H_0)$. Meaning the prior probability that θ is very close to 0 (and no more than that). If 0.9 is the upper bound and 0.2 the lower bound one has to use $\alpha^+ = 0.067$ and $\alpha^- = 0.0035$. And in case the p-value is between these boundaries one must try to find a Bayesian who can help to think about $\pi(H_0)$. Or (if this Bayesian cannot be found or is too expensive, draw another sample. Or just use the default value of $\alpha = 0.0228$ to decide. It would be interesting to evaluate the expected loss of the latter procedure.

A riddle in case of the student distribution

Finally table 3: the same outcomes for a student(4) distribution as would result from a sample of 4 normally distributed items with an estimated standard error of 0.2 (so $s/\sqrt{n} = 0.1$). Similar but with fatter tails.

Table 3: sensitivity in case Student(4) distribution							
Inputs (if not: default)				Outputs for break even point			
$L(1)$	λ	$\pi(H_0)$	w	S_{BP}	$\alpha(\%)$	$P(H_0|S_{BP})$	$BF(S_{BP})$
3	4	0.5	0.04	0.21	6.33	0.36	1:1.8
			0.02	0.22	5.76	0.31	1:2.2
	9			0.24	4.79	0.31	1:2.3
		0.2		0.15	12.1	0.22	1.2:1
		0.9		0.35	1.97	0.55	1:7.3
2				0.16	10.4	0.46	1.2:1.

A rather shocking result is that the position of S_{BP} does not change much while the values for α increase dramatically. This could disqualify the p-value as a reasonable robust measure. On closer inspection this result appears to be the effect of the quadratic loss function $L(0,\theta)$. The fat tail near θ_0 increases the impact of $\pi(H_0)$ and pulls S_{BP} to the left. But the fat tail on the right hand side pulls S_{BP} to the right. For linear $L(0,\theta)$ or one with decreasing slope (an asymptote

might be reasonable) the picture changes. It is possible to make configurations such that the α values remain comparable to Table 2. So there is no clear conclusion other than one should be extra careful in case of small samples.

5. Conclusion

It is possible to cast decision problems in the format $p < \alpha_{im}$. But the implied α strongly depends upon priors and loss functions. This can be made clear in a spreadsheet that business students should be able to understand. And the outcomes make sense in so many ways that it might convince them that statistics is fun and useful. It should enable them to make a reasonable first step in a decision process: when to treat an outcome as "significant". Even a guess of the uncertainty about break-even points in decisions, because priors and maybe loss functions are uncertain, may well be made. Hiring a Bayesian may be restricted to situations that remain unclear because the p-value falls in the "confidence interval for α".

It is tempting to speculate further on the implication of these insights for the meaning of "significance " in other situations. I think it is quite general. Decisions matter more than abstract concepts. But I leave that to others.

A rather shocking discovery was the inadequacy of the Bayes Factor in decision situations. As moreover the Bayes Factor depends on priors, I prefer the p-value. Together with the size of S that gives a reasonable summary of the data provided a sample is not too small (and it is not doubled because it was a two sided test). Given S and the p-value anyone can, with the help of a spreadsheet like mine, decide whether he or she should care about some result. Which will differ among people. Because subjectivity matters and their loss functions may differ.

Epilogue

"There's no theorem like Bayes Theorem" (George Box 1960) is a song that I often sang in front of my classes when I taught. Unfortunately, I often found that I was singing in the desert. At the end of my career, I begin to understand why. Maybe Bayesians want too much. In many cases a simple recipe in the frequentist format will do. On the occasion of my pensioning a symposium will be organized about the theme of his article. In Amsterdam, on September 30. Bernardo, Berger and O 'Hagan will attend and join in "the p-value debate.

My hope is that many more will comment. Come to the debate. And send a mail to a.f.de.vos@vu.nl if you have any comments.

References (except classics before 1970)

Andrews, D. W. K. (1994). The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica* 62, 1207–1232.

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

Bernardo, J.M. (2011). Integrated Objective Bayesian Estimation and Hypothesis Testing. *Bayesian Statistics* 9, 1-66.

Tol, R.S.J. & de Vos, A.F. (1998) A Bayesian analysis of the Enhanced Greenhouse Effect. *Climatic Change* 38, 87-112

de Vos, A.F. & Francke, M.K. (2008) Bayesian Unit Root Tests and Marginal Likelihood (my website or www1.fee.uva.nl/pp/bin/1015fulltext.pdf)

de Vos, A.F. (1987) Statistiek en Wiegedood (Statistics and Cot Death). *Kwantitatieve Methoden* (and my website)

Distributions

P value:
2,28%
P(H0|S):
0,3300
Bayes Factor:
2,03:1

—— Prior

—— Likelihood (for one S)

—— Posterior

S, Theta

Costfunctions and intersects S_left and S_right

S; 0,2

—— L0(theta)

—— L1(theta)

—— E[L0|S]

—— S_left

—— S_right

—■— S

Costs

S, Theta

P-value and alphas

—— p-value

—— alpha

—— S_left

—— S_right

Probabilty

S