The ethics of statistical testing

The Ethics of Statistical Testing

Jan Sprenger* and David Teira Serrano†

July 21, 2011

Statistics is a mathematical discipline that provides advice in the making of uncertain choices: for instance, if we want to invest in a company, we would like to see a projection of future prospects before making a decision. There are various statistical tools that we may use to cope with such uncertain choices. In this paper we will focus on *significance testing*, the most widely used statistical tool for quantitative analysis in science and business. We want to explore in what sense significance testing can help in making ethical decisions, and in what sense it may obstruct them.

Statistics is most useful for consequentialist approaches to ethics, where actions are assessed in terms of their consequences. However, not every statistical tool allows us to justify our choices in a consequentialist manner. Statistical tests can either be interpreted behaviorally, as guiding actual decisions that we make, or evidentially, as providing evidence about the truth or falsehood of a particular claim. In daily statistical practice, significance tests are often used for both ends, for inference *and* decision-making. It is this tension between the behavioral and the evidential interpretation that stands at the heart of our paper. After all, we need to be consistent in our interpretation of statistical methods if we want a proper assessment of the uncertain prospects we face, and a sound consequentialist appraisal of our choices.

In the first section of this paper, we analyze a simplified model of ethical decision-making, showing how consistency in the assignment of probabilities is a prerequisite for any consequentialist justification of our choices. In the second section we provide a short introduction to significance testing and its two main interpretations. In section three we point to inconsistencies in the actual practice of significance testing. Finally, in section four, we discuss several proposals for a consistent use of statistical tests in practical decision-making.

1 Ethics and statistics

Long before the establishment of mathematical statistics as a discipline, the ethical dimension of uncertain decisions was appraised in an Aristotelian tradition,

*Contact information: Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: http://www.laeuferpaar.de.

†Contact information: Dpto. de Lógica, Historia y Filosofia de la ciencia, Universidad Nacional de Educación a Distancia, Madrid, Paseo de Senda del rey 7, 28040 Madrid, Spain. Email: dteira@fsof.uned.es. Webpage: http://www.uned.es/personal/dteira/.

namely in terms of their *prudence* (Aubenque 1986): a good choice depended on finding the correct means for the correct goal. According to Aristotle, there was no general rule for dealing with ethical choices under uncertainty. Rather, like ancient medicine, good decision-making as thought of as a craft where one had to apply one's practical wisdom (*phronêsis*). In the same way that we defer to the advice of a medical doctor when we are sick, we should defer to the practically wise in questions of ethical decision-making. His or her voice settles disagreement about what we should do. One of the most salient examples in Greek history was Pericles, the wise manager of the city of Athens in the 5th century BC.

With Kant, almost two thousand years later, prudential choices are left out of the proper realm of ethics. The highest ethical good does not consist any more in achieving a certain goal, but in the *good will*. Moreover, in sharp contrast to the Aristotelian deference to experts, Kant developed the categorical imperative as a universal rule of action: to act in a way that could be generalized to a general law. This single maxim is to be followed independently of the uncertainty of the alternatives and the practical consequences they yield. For example, we are not allowed to lie, even if as a consequence of our truth-telling, a malevolent dictator will be able to track down innocent refugees. In such an act of lying, we would use the person that we are lying to as an means to an end, something that is incompatible with Kant's vision of human autonomy and dignity.

Evidently, statistical advice is most relevant for those approaches in ethics that appraise the rightness of our choices in terms of their consequences. Remember that for Kant, we are morally compelled to abide by the categorical imperative: you are not allowed to protect a refugee by lying about her whereabouts. This emphasis on universal maxims and duties is a *deontological* approach. By contrast, a *consequentialist* in the prudential tradition (cf. Sinnott-Armstrong 2008) would consider such action wrong, given the likely consequences (the refugee being tortured and/or killed), and would have justified lying. Here, a statistical analysis can step in, by weighting the likely consequences of our actions against each other: maybe the refugee will be able to escape despite our collaboration with the regime, so telling the truth might not be such a bad thing.

Notably, a statistical analysis does not impinge on our goals: these are taken as given. But if our decision depends on the likelihood of attaining these goals, a statistical analysis may evaluate the ethical correction of our decision. For instance, if a hedge fund manager invests her customers money on the basis of careless calculations, we will consider her morally blameworthy: we need accurate estimates of the consequences of our investment decisions in order to justify them. Here arises a source of epistemic and ethical concerns: since the correction of our choice depends on the correspondence between our models and the actual risks we are dealing with, how do we know that our model adequately captures such risks? The gist of Nassim Taleb's (2007) best seller *The Black Swan* is that risks in financial markets (as in other domains) are often not adequately described: we mistakenly assume that the real risks can be structured by a simple probabilistic model, such as the Normal distribution.

Due to the idealizing nature of such assumptions (e.g, the extremely thin tails of the Normal distribution), we are ill prepared to estimate the real likelihood of high-impact events.[1]

The recent financial crisis illustrates that we can rarely apply statistics blindly, as if we had a mechanical algorithm: statistical analysis depends on a number of assumptions about the data and the proper way to handle them. Intuitively, the decision-maker seems to have the responsibility to check those assumptions. For instance, according to Michael Lewis (2010), there were a number of traders who anticipated the 2007 crash of the subprime mortgage market and actually earned significant amounts of money by selling insurance against it. The standard procedure to redistribute the risk of a mortgage defaulting was through a collateralized debt obligation (CDO), a bond in which thousands of loans were gathered in tranches with different levels of risks, under the assumption that they would not all default together. According to Lewis, it took just a simulation of the effects of home price appreciations on these loans to convince an insightful trader (such as Gregg Lippmann) that default rates would violate the CDO assumptions: they could very easily collapse simultaneously.

Should we blame the sellers of CDOs for not conducting such simulations? Before we answe "yes", we need to be aware of what we may legitimately expect and require from statistics in order to attribute the responsibility for a proper or improper use. In particular, such a responsibility cannot be easily attributed, unless we have a regulative ideal against which to evaluate a particular choice.

The classical regulative ideal in consequentialist decision-making is *Subjective Expected Utility Theory (SEUT)*, developed by, inter alia, Ramsey (1926) and Savage (1954). By now, it has become the standard model of decision-making under uncertainty in social science and in moral and political philosophy. The classical justification proceeds by outlining an intuitive axiom system for individual preferences, demanding that they be complete, transitive, respect the sure-thing principle, apply to mixed bundles of goods, and so on. Then, it is shown that such a system of preferences admits a (unique up to affine transformation) representation in terms of a real-valued utility function over the outcomes and a probability function representing the subjective uncertainty of the agent. That is, if $a_1, \ldots, a_n$ denote the available actions, $p(\cdot)$ denotes our subjective probability function over states $s_1, \ldots, s_n$, and u_{kl} the utility of

[1] Frank Knight (1921) famously argued that statistical theory could not be applied to business decisions. When a businessman is making a choice between uncertain alternatives, this uncertainty arises from so many particular circumstances that there is no way of telling if such a decision will ever take place again. For Knight, each choice is entirely unique and cannot be made part of a class of similar choices arising from a general decision rule. On the other hand, statistical decision theory is a theory of probabilistic decisions, and at Knight's time, probabilities in statistical inference were usually explicated as relative frequencies. We could estimate how *risky* a decision rule is analysing how frequently it yields successful choices, but if each decision is entirely singular, as Knight argued, we cannot quantify the risk: we are dealing with real *uncertainty*. Modern financial economics assumes precisely the opposite: there are precise mathematical models of the risks involved in most of our economic decisions, and these models allow us to determine which option is best.

action k in state l, then action a_i is better than action a_j if and only if

$$\sum_{k=1}^{n} p(s_k)u_{ik} > \sum_{k=1}^{n} p(s_k)u_{jk}. \tag{1}$$

In other words, the averaged or expected utility of a_i exceeds the expected utility of a_j with respect to one's subjective probability funtion, hence the name Subjective Expecte Utility Theory.

As a descriptive model of the average economic agent, SEUT is often contested (Allais 1953; Ellsberg 1961). However, it is often defended on normative grounds (Jallais et al. 2008). If you are a consequentialist, not taking into account the principles of probability will put you in a difficult position. Take, for example, the representation of uncertainty by a probability function – an essential cornerstone of SEUT. If our degrees of belief violate the axioms of probability, a malicious bookie can set up a gamble (according to our degrees of belief) whose set of odds and bets guarantees a profit for him, whatever the actual outcome (Vineberg 2011). Since degrees of belief are standardly operationalized via betting behavoir or judgments on the fairness of bets, non-probabilistic degrees of belief are arguably self-defeating.

Still, even if we are convinced by this "Dutch Book Argument" in favor of coherent probabilities, it has not been demonstrated that we should maximize the *average* expected utility. The standard argument to that end goes that in the long run, acting in accordance with SEUT delivers practical success. In his 1951 essay "Why *should* statisticians and businessmen maximize moral expectation?", Jacob Marschak tried to derive from the rule of maximizing expected utility the satisfaction of "the rule of long run success": under certain assumptions, it will be almost certain that a sequence of strategies maximizing expected utility will outperform any other consequentialist decision rule (Marschak 1951, 504-505). Unlike the Dutch book argument, Marschak's case was about winning, rather than not losing, appealing to the practical rationality of businessmen. Still, this argument has, apart from doubts about the plausibility of its assumptions, often been challenged – particularly by the empirical findings of Kahneman and Tversky (1979).

Objections put aside for the moment, we see two different consequentialist justifications for SEUT as a standard of rational and ethical choice. If the moral correctness of our decisions depends on an accuracy of their consequences, SEUT contributes to it in two ways. There is, on the one hand, (probabilistic) coherence: make your choices in a way that it is not self-defeating for your aims. On the other hand, there is success: make your choices in a way that actually maximizes your chances of attaining your goals. We will, in the remainder, use SEUT as a regulative ideal against which we evaluate different approaches to statistical testing. If we do not apply our statistical techniques consistently, we cannot expect statistics to increase our chances of success. Hence, from a consequentialist perspective, we will lack a proper statistical justification of our decisions. We will be just deceiving ourselves or misleading our audience into the incorrect belief that we have such a justification.

2 Two varieties of frequentist statistics

Statistics tries to anticipate random events by drawing on the data that have accumulated in our experience. We try to discern a pattern in the random distribution of these data (past, present and future): we form hypotheses about such distributions and we use statistical tests to check whether our hypotheses are correct. In its simplest form, a hypothesis test compares two hypotheses H_0 and H_1 about an unknown quantity of interest, represented by the parameter $\theta \in \Theta$. Sometimes we deal with a precise hypothesis about θ, e.g., $H_0 : \theta = \theta_0$ – the null hypothesis – and oppose it to an unspecified alternative $H_1 : \theta \neq \theta_0$. It is then tested whether the data are compatible with the null, or whether a significant deviation is present. Such *hypothesis testing* is the prime activity of frequentist inference – inference that shuns subjective assessments of uncertainty and only builds on the probability of events under the tested hypotheses, that is, the sampling distribution.

There are two main approaches to hypothesis testing within frequentists statistics. The first one, devised by Jerzy Neyman, argues that statistical testing is about *making decisions* about the acceptability of a hypothesis. The second one, due to Ronald A. Fisher, claims that statistical tests should only provide an *assesment of the evidence* for or against a particular scientific claim. Both interpretations are often confused in practice. Below, we spell out the difference: conflation of both approaches goes at the expense of conceptual consistency that we seek in order to make properly informed decisions.

Together with Egon Pearson, Neyman designed a hypothesis test as a proper decision rule, that is, as a function $T : \mathcal{X} \to \{\text{accept}\, H_0, \text{reject}\, H_0\}$, $\mathcal{X}$ being the sample space. Think, for example, of industrial quality control. Should we accept a delivery of bulbs which we have sampled for defective elements? The answer will, inevitably, depend on how many elements in our sample have been found to be defective. We might make the wrong decision if, by chance, we pick a nonrepresentative sample, but if the test is properly designed, only a small number of our decisions will be mistaken.

From an epistemological point of view, the Neyman-Pearson approach receives its justification by the associated *error probabilities*. Let the null hypothesis be that in our delivery of bulbs, there are not more than 10% defect elements, and let the alternative posit that there are more than 10% defect bulbs. (Assume that 10% is the highest proportion of defect bulbs at which it is still economically advantageous for us to accept the delivery.) The test statistic T is then so designed that the null hypothesis is rejected in at most 5% of all cases where it is true, that is, where the delivery is acceptable.[2] This *type I error level* – the probability of an erroneous rejection of the null – can also be chosen to be 10%, or 1%, etc. – the cutoff is purely conventional and reflects how important we find it that the null is not erroneously rejected.

Evidently, there are various tests that satisfy this property. Trivially, even a test that always accepts the null (and the delivery of bulbs) will have a type

[2] Mathematically, this is done by assigning the acceptance region a weight of 0.95, that is $\int_{T=0} P(x)dx = 0.95$.

I error level as low as 0%. While such a test appears desirable in theory, it is practically unsound: the decision does not depend at all on how many defect bulbs are found. In other words, the test is not responsive to the strength of the evidence. Therefore, the acceptance region should be chosen such that, for a type I error level deemed acceptable, say, 5%, the *type II error level* – the *probability of an erroneous acceptance* of the null – is minimized. We say in that case that the *power* of the test, its ability to recognize the alternative when it is true, is maximized conditional on the level of the test being 5%. In this way, both possible types of error are controlled, and the optimal Neyman-Pearson test will rarely lead to a wrong decision:

> we shall reject H_0 when it true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H_0 sufficiently often when it is false. (Neyman and Pearson 1933, 291, notational details changed)

Such a behavioral rationale is well-suited to inform real decisions with concrete, immediate impact. Neyman's approach emerged from the world of industrial quality control, where every decision has costs and benefits. Statistical tests à la Neyman were aimed at hedging costs, conforming to the consequentialist spirit presented in the previous section. But not every statistician shared such an applied perspective: many "decisions" in science are just preliminary and subject to further evidence. A behavioral interpretation of statistical testing was considered inferior to an evidential, inferential interpretation, where we assess the truth of a hypothesis, independently of the consequences of a wrong assessment. As R.A. Fisher put it:

> In the field of pure research no assessment of the cost of wrong conclusions [...] can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence. (Fisher 1935, 25–26)

Two arguments are implied here. First, we cannot quantify the utility that correctly accepting or rejecting a hypothesis will eventually have for the advancement of science. The far-reaching consequences of such a decision are beyond our horizon. Second, statistical hypothesis tests should state the *evidence* for or against the tested hypothesis: a scientist is interested in whether she has reason to believe that a hypothesis is true or false, and her judgment should not be obscured by the long-term consequences of working with this rather than that hypothesis. For Fisher, testing an hypothesis requires an assesment of the significance of the evidence against it. By his emphasis on evidence rather than decisions, Fisher departs from Neyman and Pearson's consequentialist reasoning – a change that severely affects the interpretation of those statistical testing procedures.

Significance tests aim at determining whether a perceived effect in the data is real or possibly due to chance. If the discrepancy between data and null hypothesis is large enough, we are entitled to infer to the presence of a significant effect. Suppose we have a precise null hypothesis $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

For measuring the discrepancy in the data $x := (x_1, \ldots, x_N)$ with respect to postulated mean value θ_0 of a population with known variance σ^2, one canonically uses the standardized statistic

$$z(x) := \sqrt{N}\,\frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{\sigma^2}} \tag{2}$$

Thus, we may re-interpret equation (2) as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}}. \tag{3}$$

Determining whether a result is significant or not depends then, on the distribution of the value of z. Practitioners usually use the so-called *p-value* or *significance level*, the "tail area" of the null under the observed data (see figure 1), which can be computed as

$$p := P(|z(X)| \geq |z(x)|) \tag{4}$$

that is, as the probability of observing a more extreme discrepancy under the null than the one which is actually observed. On that reading, a low significance level indicates evidence against the null since the chance that z would take a value at least as high as $z(x)$ is very small, if the null were indeed true. Conventionally, one says that $p < 0.05$ means significant evidence against the null, $p < 0.01$ very significant evidence, etc. To repeat, p-values serve, in the first place, the purpose of statistical inference, not the purpose of statistically informed decision-making.

Fisher has interpreted significance levels as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher 1956, 43). What is more, Fisher is explicit that some cutoff value for p should be regarded as necessary for speaking about the presence of a scientifically significant effect:

> Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance. (Fisher 1935, 504)

The possibility of integrating these two approaches to statistical inference into a consequentialist framework are remarkably different. Neyman incorporates an explicit consequentialist dimension: we can justify the acceptance of a hypothesis in terms of the balance between the number of successes and failures we will obtain if we consistently apply our decision rule. If we are willing to bear a mistaken decision about hypothesis in 5 out of every 100 tests, an appropriate hypothesis test provides the statistical tools to ensure this error rate in the long run. In this way, frequentist statistics can be naturally integrated into reponsible decision-making.

However, the majority of inferences and decisions in science and business are derived from observed significance levels, in line with Fisher's evidential approach. Indeed, Fisher's above quote demonstrates that the borderline between

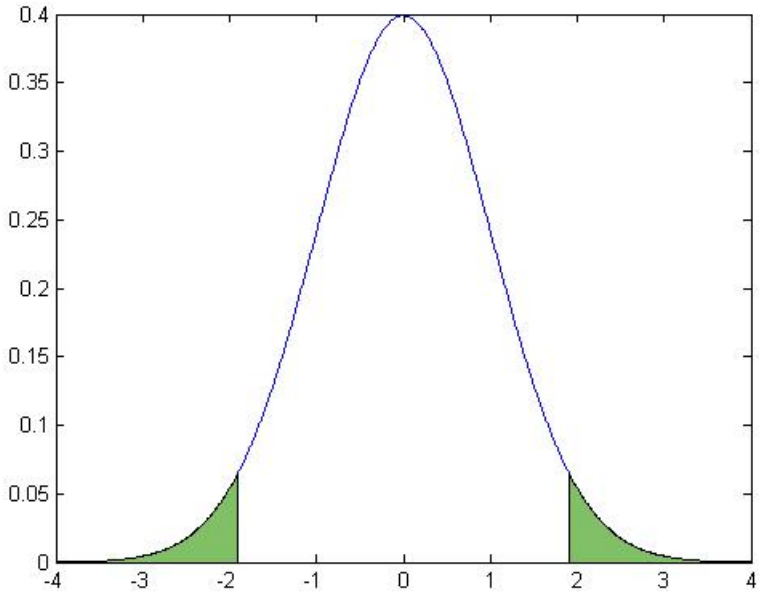

Figure 1: The rejection region for testing the mean of a N(0,1)-distributed random variable at the 5% level.

evidence and practical decisions is thin ("ignore entirely all results which fail to reach this [significance] level"). Significance tests do not quantify how often will we suceed or fail if we apply such rules, and leave ample room for interpretation when we try to apply them in practice. As a consequence, they are often misused, without no clear way to attribute responsibility for the failures. The next section illuminates those criticisms in detail.

3 Misuses of significance testing

As mentioned above, significance tests are, although devised as procedures for stating the evidence against the null, frequently used for substantiating practical decisions, e.g., the null is either accepted or rejected depending on the strength of the evidence. Is this practice compatible with the regulative ideal of SEUT? Do significance tests give a valid assessment of our uncertainty about the tested hypothesis?

Concretely, we have to ask whether p-values can be meaningfully related to subjective posterior probabilities (that is, probabilities conditional on the observed evidence) that enter the expected utility analysis. While a subjective analysis is often charged with being arbitrary, it cannot be doubted that in some cases, e.g., when reasoning in games of chance, subjective probability assignments can be objectively grounded. In these canonical cases, p-values should give a valid cue about subjective posterior probabilities.

However, the analyses of Berger and Delampady (1987) and Berger and Sellke (1987) have shown that p-values tend to grossly overstate evidence against the null, to the extent that the posterior probability of the null – and even the *minimum* of $p(H_0|x)$ under a large class of prior uncertainty assessments – is typically much higher than the observed p-value. In other words, even a subjectivist analysis that is maximally biased against the null is still less biased than a p-value analysis. This has led statisticians to state that "almost anything will give a better indication of the evidence provided by the data against H_0" (Berger and Delampady 1987, 330). The main source of the problem is that p-values do not make use of the full information contained in the data – namely that the observed discrepancy is *equal* to z – but only of the information that the discrepancy is greater or equal to z, cf. equation (4).

The situation is further complicated if we focus on the justification for using (4) as a statistic that measures the strength of the evidence against the null. Fisher famously argued that a low p-value, that is, a highly significant finding, means that "either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true" (Fisher 1956, 39). That is, in the face of surprising results, we make an inference to the best explanation, namely to the falsity of the null. On a superficial glance, this inference rule provides a natural implementation of Popper's critical rationalist attitude into statistical reasoning. According to that school of philosophy, scientific method consists in the successive testing and refutation of conjectures one comes up with. However, the analogy is superficial. Fisher's Disjunction can be rephrased as the inference

$p($`Data`$|$`Null Hypothesis`$)$ is low.
`Data` is observed.

`Null Hypothesis` must be false.

Many arguments and counterexamples have been raised in order to show that this probabilistic *modus tollens* is invalid (Hacking 1965; Cohen 1994). First of all, what is valid in deductive logic need not be valid in probabilistic logic. Second, only with respect to a well-specified set of alternatives we can meaningfully say that a certain set of data constitute evidence against a hypothesis. The idea of testing a hypothesis "in isolation", without consideration of alternatives, has proved to be incoherent (Edwards, Lindman and Savage 1963; Spielman 1974; Royall 1997). In other words, even for purely evidential purposes, we should not use significance testing without a very careful consideration of the way we frame a hypothesis and the alternatives.

In actual practice such caution in the use and interpretation of signficance testing is often missing. In economics, for instance, the economists Deirdre McCloskey and Stephen Ziliak have launched strong attacks against significance tests in a series of papers and books (McCloskey and Ziliak 1996, 2008; Ziliak and McCloskey 2004, 2008). Let us give their favorite example.

Assume that we have to choose between two diet cures, based on pill A and pill B. Pill A has an average effect of making you lose 10 pounds, with an an average variation of 5 pounds.[3] Pill B will make you lose 3 pounds on average,

[3]The concept of "average variation" is intuitively explicated as the statistical concept of

with an average variation of 1 pound. Which one leads to more significant loss? Naturally, we opt for pill A, in spite of the higher variation, because the effect size is so much larger.

However, if we translate the example back into significance testing and identify the null hypothesis with the default claim that there is no effect at all, the order is reversed. Observing a three pounds weight loss after taking pill B, with a known standard error of one pound, is stronger evidence for the efficacy of B than observing a ten pounds weight loss after taking pill A, with a known standard error of five pounds, is evidence for the efficacy of B:

$$z_A(10) = \frac{10-0}{5} = 2 \qquad\qquad z_B(3) = \frac{3-0}{1} = 3$$

Thus, there is a notable discrepancy between our intuitive judgment about which pill is effective in making a patient lose weight. This occurs because statistical significance is supposed to be "a measure of the strength of the signal relative to background noise" (Hoover and Siegler 2008b, 58). On this score, pill B indeed performs better than pill A, and reasonably so because there is quite some noise in the effects of pill A. But what really matters, what economists, businesswomen and policy-makers are interested in, is the effect size, not the signal strength/noise ratio captured by significance tests, argue McCloskey and Ziliak. We are not interested in whether we can ascertain the presence of *some* effect, but whether we can demonstrate a *substantial* effect. In other words, we have to state in which currency we measure effects, and what a deviance of one, two, or three standard errors actually means for the intended application.

According to McCloskey and Ziliak, economists and other social scientists frequently commit the fallacy of neglecting this fundamental difference. By scrutinizing the statistical practice in the top journal *American Economic Review*, as well as by surveying the opinion of economists on the meaning of statistical significance, they arrive at the conclusion that most economists are unaware of the proper meaning of statistical concepts.[4] In practice, "asterisking" prevails: e.g., in correlation tables, the most significant results are marked with an asterisk, and these results are the ones that are supposed to be real, big, of economic importance. Whereas the other correlations are neglected. This neglects two salient pitfalls: first, an effect need not be statistically significant to be big and remarkable (like pill A), second, a statistically significant effect can be quite small and uninteresting (like pill B).

Even more disconcerting is that according to empirical surveys, many practitioners believe that if a result speaks highly significantly against the null, then it must be wrong (Oakes 1986). In other words, the null is believed to be highly improbable if a highly significant result is observed. But posterior probabilities of the null hypothesis don't have a place in the frequentist inference framework that we have assumed so far. Even if that particular fallacy is avoided, conflation often reigns: p-values are often perceived as the probability of replicating

standard deviance: for a random variable X, we calculate $\sqrt{E[(X - E(X))^2]}$.

[4] Their results are disputed by Hoover and Siegler (2008a) and Spanos (2008), but reaffirmed in McCloskey and Ziliak (2008). It is fair to say that the discussion of this point is still open. See also Zellner (2004).

an effect of at least the same size, as the level of type I error, etc. None of these standpoints are statistically valid.[5]

A main danger of conducting significance tests is thus that misinterpretation is so prolific, distorting the results. Methodologists such as Fidler (2005) and Cumming and Finch (2005) have therefore suggested to drop significance tests altogether and to report confidence intervals for the parameter of interest instead. Taken together with the more theoretical criticisms of significance tests, it seems that the practice of basing business and science decisions on statistical data is often far from the ideal of ethically responsible and prudential decision-making.

4 Discussion

Our analysis has revealed that despite their apparent simplicity, significance tests are difficult to interpret and to practically use in a consistent manner. From a consequentialist perspective, we should not value much decisions that are grounded on misinterpreted significance tests. If our hedge-fund manager had made an investment on the basis of any such test, we may consider him morally blameworthy, but if the confusion is so widespread (as the CDO case seems to point out) no court will declare him guilty.

Therefore, it is not only an epistemological, but also an ethical requirement to publicly agree on standards for consistent statistical practices. As for significance testing, there is much room for improvement. We briefly sketch three possible options within the frequentist paradigm, none of them being entirely unproblematic.

1. Keep significance tests, but interpret them properly, e.g., by including effect size and power measures. This proposal by Hoover and Siegler (2008a,b) accepts that failure to distinguish between significance level and effect size is a fallacy, but argues that significance testing *does* have an important function in science and particularly economics: namely to decide whether the error in a statistical model can be regarded as truly random, or whether a systematic bias is present. To our mind, however, pure significance testing without considering explicit alternatives has trouble to be embedded into a coherent logic of inference.

2. Move to confidence intervals. A solution that has become increasingly popular in psychology and that has, in the meantime, reached out to editorial policies (Wilkinson et al. 1999). Confidence intervals replace significance level by providing 95%/99%/etc. coverage areas for the data, given a particular value of the parameter of interest. It has been argued (Cumming 2008) that they are a much better indicator of *effect replication* than

[5]Even sophisticated defenders of significance testing in economics, such as Hoover and Siegler (2008b, 58), sometimes go wrong, e.g., when they call a significance level a "type I error probability". To recall, error probabilities are pre-experimental characteristic of a decision procedures, significance levels are measures of discrepancy between data and null.

significance levels, and that they are more stable for the purpose of meta-analysis. However, these intervals must *not* be interpreted as credible intervals in the sense that with 95% probability, the parameter is contained in the confidence interval. If scientists already have trouble to distinguish between p-values and posterior probabilities, they may be equally likely to commit the natural fallacy of interpreting confidence intervals along the lines of degrees of belief about the location of the parameter value.

3. Retract to Neyman's behavioral interpretation of statistical tests. That is, statistical tests are not used for finding out whether a hypothesis is right or wrong, but only for supporting a particular decision. This proposal has never found many supporters in practice. One of the most salient reasons is that any statistical analysis would then be highly idiosyncratic, dependent on the interests and the particular loss function of the individual that conducts the analysis. Intersubjective communication of results and conclusions would, as a consequence, suffer. Moreover, the technical problems – how to compute the level and the power of more complicated testing problems, how to avoid the slippery slope to a subjectivist framework – are far from trivial. – An intermediate position between Neyman and Fisher is advocated by Mayo and Spanos (2006) under the label of *error statistics*.

A radically different solution goes back to the grounding of decisions and tests in SEUT. The idea is to conceive of statistical inference as providing the relevant probabilities for what might or might not be the case, and to feed these probabilities into an expected utility analysis. That is, we inform good decisions by means of well-calibrating our assessments of uncertainty. The standard way of doing so is via *Bayesian Conditionalization*. If we are revising our degree of belief in hypothesis H in the light of evidence E, our new degree of belief in H can be computed as

$$p_{\text{new}}(H) := p(H|E) = p(H)\frac{p(E|H)}{p(E)}. \tag{5}$$

Statistical analysis can thus inform right decisions in the following way: we start with a *prior* probability function $p(\cdot)$ that represents our initial uncertainty, revise it in the light of statistical evidence E by equation (5), and apply the principle of maximizing expected utility with our new *posterior* probability function $p(\cdot|E)$.

This subjective understanding of statistical inference is called Bayesian inference. It dominates in moral and political theory, decision theory and game theory. One of its big advantages is its coherence, simplicity and universality: it is by construction in sync with SEUT. Moreover, while complex applications demand mathematical sophistication, the basic conceptual framework of equation 5 remains unscathed. Finally, the epistemic and the ethical/utility-related aspects of the analysis are separated from each other (an advantage vis-à-vis Neyman's approach), and the probabilistic assessment precedes and serves as an input for the actual decision-making.

However, many scientists – in particular those coming from the natural sciences – have problems with the subjective interpretation of probability, or consider it unsuitable for scientific analysis. According to that view, science should deal with objective facts, objective evidence, not with revising subjective (and ultimately arbitrary) degrees of belief. A fortiori, scientific inference must not proceed by Bayesian Conditionalization. The main attraction of the frequentist paradigm is, to repeat, that it eschews degrees of beliefs, builds on the view of probability as relative frequency and devises statistical methods that fit these parsimonious modeling assumptions. But in the light of the problems that the most popular frequentist testing procedures experience, the Bayesian paradigm may deserve more attention. Indeed, recent developments in statistical methodology support a trend towards increased use of Bayesian methods.

Thus, a responsible decision-maker needs to think carefully about statistical methodology: too much can depend on choosing an adequate or inadequate interpretation of a statistical test. We believe that striving for ethically sound decisions does not commit oneself to Bayesianism or frequentism; however, the frequentist stance may be loaded with more challenges and pitfalls with respect to applying it consistently.

References

Allais, Maurice (1953): "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine", *Econometrica* 21, 503–546.

Aubenque, Pierre (1986): *La prudence chez Aristote*. Paris: Presses universitaires de France.

Berger, James O., and Mohan Delampady (1987): "Testing Precise Hypotheses (with discussion)", *Statistical Science* 2, 317–352.

Berger, James O., and Thomas Sellke (1987): "Testing a point null hypothesis: The irreconciliability of P-values and evidence (with discussion)", *Journal of the American Statistical Association* 82, 112–139.

Cohen, Jacob (1994): "The Earth is Round ($p < .05$)", *American Psychologist* 49, 997–1001.

Cumming, Geoff, and S. Finch (2005): "Inference by eye: Confidence intervals, and how to read pictures of data", *American Psychologist* 60, 170–180.

Cumming, Geoff (2008): "Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better", *Perspectives on Psychological Science* 3, 286–300.

Edwards, Ward, Harold Lindman and Leonard J. Savage (1963): "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70, 450–499.

Ellsberg, Daniel (1961): "Risk, Ambiguity, and the Savage Axioms", *Quarterly Journal of Economics* 75 , 643–669.

Fidler, Fiona (2005): *From Statistical Significance to Effect Estimation*. Ph.D. Thesis: University of Melbourne.

Fisher, Ronald A. (1926): "Arrangement of Field Experiments", *Journal of Ministry of Agriculture* XXXIII, 503–513.

Fisher, Ronald A. (1935): *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, Ronald A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Hacking, Ian (1965): *Logic of Statistical Inference*. Cambridge University Press, Cambridge.

Hoover, Kevin D., and Mark V. Siegler (2008a): "Sound and Fury: McCloskey and Significance Testing in Economics", *Journal of Economic Methodology* 15, 1–37.

Hoover, Kevin D., and Mark V. Siegler (2008b): "The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey", *Journal of Economic Methodology* 15, 57–68.

Jallais, S., P. Ch. Pradier, and D. Teira (2008): "Facts, Norms and Expected Utility Functions", *History of the Human Sciences* 21, 45–62.

Kahneman, Daniel, and Amos Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47, 263–291.

Knight, Frank (1921): *Risk, uncertainty and profit*. Boston/New York: Houghton Mifflin Company.

Lewis, Michael (2010): *The big short: inside the doomsday machine*. New York: W.W. Norton.

Marschak, J. (1951): "Why Should Statisticians and Businessmen Maximize Moral Expectation?", in: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 493–506. Berkeley: University of California Press.

Mayo, Deborah G., and Aris Spanos (2006): "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *The British Journal for the Philosophy of Science* 57, 323–357.

McCloskey, Deirdre N., and Stephan T. Ziliak (1996): "The Standard Error of Regressions", *Journal of Economic Literature* 34, 97–114.

McCloskey, Deirdre N., and Stephan T. Ziliak (2008): "Signifying Nothing: Reply to Hoover and Siegler", *Journal of Economic Methodology* 15, 39–55.

Neyman, Jerzy, and Egon Pearson (1933): "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society A* 231, 289–337.

Oakes, M. (1986): *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Ramsey, Frank P. (1926/1990): "Truth and Probability", in: Philosophical Papers (eds.), *D. H. Mellor*. Cambridge: Cambridge University Press.

Royall, Richard (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Savage, Leonard J. (1954): *Foundations of Statistics*. Dover: New York.

Sinnott-Armstrong, Walter (2008): "Consequentialism", in: The Stanford Encyclopedia of Philosophy (eds.), *Edward N. Zalta*. http://plato.stanford.edu/archives/fall2008/entries/consequentialism/, retrieved on July 18, 2011..

Spielman, Stephen (1974): "On the Infirmities of Gillies's Rule", *British Journal for the Philosophy of Science* 25, 261–265.

Taleb, Nassim (2007): *The black swan: the impact of the highly improbable*. New York: Random House.

Vineberg, Susan (2011): "Dutch Book Arguments", in: The Stanford Encyclopedia of Philosophy (eds.), *Edward N. Zalta*. http://plato.stanford.edu/archives/fall2008/entries/dutch-book/, retrieved on July 18, 2011..

Wilkinson, L., and Task Force on Statistical Inference (1999): "Statistical Methods in Psychology Journals", *American Psychologist* 54, 594–604.

Zellner, Arnold (2004): "To Test or Not to Test and If So, How?", *Journal of Socio-Economics* 33, 581–586.

Ziliak, Stephen T., and Deirdre N. McCloskey (2004): "Size Matters: The Standard Error of Regressions in the *American Economic Review*", *Journal of Socio-Economics* 33, 527–546.

Ziliak, Stephen T., and Deirdre N. McCloskey (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.