Abstract

The *p*-value is the probability under the null hypothesis of obtaining an experimental result that is at least as extreme as the one that we have actually obtained. That probability plays a crucial role in frequentist statistical inferences. But if we take the word 'extreme' to mean 'improbable', then we can show that this type of inference can be very problematic. In this paper, I argue that it is a mistake to make such an interpretation. Under minimal assumptions about the alternative hypothesis, I explain why 'extreme' means 'outside the most precise predicted range of experimental outcomes for a given upper bound probability of error'. Doing so, I rebut recent formulations of recurrent criticisms against the frequentist approach in statistics and underscore the importance of random variables.

On the Correct Interpretation of P-values and the Importance of Random Variables

1. Introduction. A frequentist approach[1] to theory testing (FA) dictates the following decision rule with regards to the null hypothesis (H0)[2]. It can be stated as follows:

If the probability of observing an experimental outcome that is at least as extreme as the one we have actually observed is too low given that H0 is true, then we should reject H0.

That probability is called 'the *p*-value' and this general definition is one on which every statistician (scientist) can agree. For instance, here is how Eric-Jan Wagenmakers defines the *p*-value:

> The probability of encountering a value of a test statistic that is at least as *extreme* as the one that is actually observed, given that the null hypothesis is true (Wagenmakers 2007, 799) [emphasis added].

Complications arise when we wish to cash-out the meaning of the term 'extreme'. According to one particular interpretation that we often encounter in the philosophical and in the scientific literature, 'extreme' means 'improbable'. It is such that we can demonstrate that (FA) is inadequate.

In this paper, I explain why we must discard that faulty interpretation. Under minimal assumptions about the alternative hypothesis, I argue that the word 'extreme' means 'outside the most precise predicted range of experimental outcomes for a given upper bound probability of error'. By the same token, I show why random variables are important. They

[1] There are two main schools of thought in frequentist testing: the Fisherian and the Neyman-Pearson. The decision rule presented here is more adequate for a Neyman-Pearson framework. According to the latter, the rejection of H0 implies the acceptance of an alternative hypothesis (H1). The Neyman-Pearson approach accordingly aims to minimise the probability of rejecting H0 when H0 is true (the type-I error) and to minimise the probability of rejecting H1 when H1 is true (the type-II error). Fisher, on the other hand, was against a formal treatment of the type-II error. He also criticised the 'accept/reject' procedure and preferred to interpret the *p*-value as providing degrees of evidence against H0. I will alert the reader when the differences can matter.

[2] The null hypothesis is the default hypothesis. It is the one that we accept unless the evidence suggests that we should reject it.

allow us to give precise measures of dispersion and of central tendencies under H0, which in turns allow us to determine a precise range for our predictions. The main motivation for this work is to root-out many criticisms of (FA) that are now entrenched[3] in the philosophical and scientific literature.

In the first part of this paper, I make it clear that (FA) is inadequate if we take 'extreme' to mean 'improbable'. Firstly, I demonstrate how a frequentist decision procedure could suggest two incompatible courses of action to be taken at the same time (reject H0 and do not reject H0). To do this, I combine two very similar arguments against (FA). One of them has been put forward by Daniel Greco (2011) and the other, by Elliott Sober (2008). Secondly, I show that many inferences would not make any sense. To bring this point home, I present an argument that has been brought back to light by Wagenmakers (2007). I also discuss an interesting variation on that argument (Greco 2011).

In the second and third part, I explain why it is a mistake to interpret 'extreme' as 'improbable' and argue that it means 'outside the most precise predicted range of experimental outcomes for a given probability of error' when we make minimal assumptions about the alternative hypothesis. I also underscore the fact that random variables are particularly valuable to frequentist theory-testing. As a result, I solve the problems mentioned in the first part of this paper.

Unless specified otherwise, I will always assume that we are in a Fisherian context, *i.e.*, one where there is no formal treatment of the alternative hypothesis. I will use expressions such as 'reject H0' and 'decision rule'. But a true Fisherian might want to ready 'we have evidence against H0' and 'rule of inference'.

[3] What I mean by 'entrenched' is that they are recurrent and appear in high-profile publications.

2. A Faulty Interpretation of P-values

2.1 One Experiment, Two Incompatible Decisions to Take

There are many critical arguments against (FA) that rest on an interpretation of the word 'extreme' as 'improbable'. In this section, I present two of them and reach a contradiction. The first one has recently been put forward by Daniel Greco (2011). Greco maintains that Fisher's decision procedure to reject H0 (see footnote 1) validates the following argument called 'the probabilistic *modus tollens*' (PMT)[4]:

"(P1) If the null hypothesis is true, then the value for the test statistic will probably not be at least as *extreme* as x.

(P2) The value for the test statistic is at least as extreme as x. Therefore:

(C) Probably, the null hypothesis is false" (Greco 2011, 611) [emphasis added].

He then goes on to show that PMT is invalid with the help of the following example:

> What's wrong with PMT? I roll a die 10 times. The sequence of number showing on the face of the die is as follows: 4, 4, 1, 3, 1, 3, 6, 3, 4, 3. Call this sequence S. Now, consider the hypothesis that the die is fair –each face is equally likely to come up, and each roll is independent from the rest. The probability that I should obtain sequence S upon rolling the die 10 times, on the hypothesis that the die is fair, is quite low (in particular, it is the same as the probability for any other particular sequence: 1/6 to the tenth power). But I did obtain sequence S. PMT would tell us to conclude that the die is probably not fair. But this would be silly (Greco 2011, 311-312)

Here, Greco correctly points out that every possible experimental outcome will be very improbable (1/6 to the tenth power). This means, according to him, that for any x that we might choose in (P1), our observations will be at least as improbable (*i.e.*, as extreme) as x. S certainly is, thus we should reject H0. In fact, if we follow this line of reasoning, we will

[4] Elliott Sober coined the expression 'probabilistic modus tollens'. I shall also explain why he claims that it is invalid.

always reject H0, which is very silly indeed[5]. Therefore, (FA) cannot be wholly adequate.

But things get even worse when we interpret that very same experiment with the help of the *p*-value, like Sober does (Sober 2008, 55)[6]. Here is the formal interpretation of the *p*-value when we take 'extreme' to mean 'improbable':

(Def1) A *p*-value is the probability of the disjunction of all the possible experimental outcomes that are at least as improbable as the event that we observed given that H0 is true.

In order to determine the *p*-value in this case, that definition implies that we should compute the sum of the probability of obtaining S and of all the other possible outcomes that are at least as improbable as S, *i.e.*, all of them. This obviously implies that our *p*-value will be equal to 1 and that we will never reject H0. Now that is troubling to say the least. We can actually infer a contradiction if we follow both lines of reasoning: A (We should reject H0) and not-A (It is not the case that we should reject H0).

Of course, neither Greco nor Sober claim that frequentists cannot find a better way to test whether or not a die is fair. But they have to endorse the idea that such a contradiction can be inferred if we follow the frequentist rules of decision: the probabilistic modus tollens and the inference rule based on the *p*-value (see introduction). I will ultimately show that they are mistaken.

2.2 Inexplicable Decisions

The previous problem appears to be relatively local. It depends heavily on the fact that experimental outcomes are equiprobable. But there are even more serious problems

[5] Ian Hacking traces back the origin of that fallacy to John Arbuthnot (1710) (Hacking 1965, 75).
[6] Sober actually discusses an experiment involving a coin. But the point is essentially the same.

afflicting (FA) when we interpret 'extreme' as 'improbable'. In what follows I expound an argument that we can find in (Wagenmakers 2007, 782)[7].

Consider two distributions under H0: f(x) and h(x). Both are defined in Table 1. Suppose that both of our test statistics is equal to 5.

Table 1: different p-values for x=5

Distribution	**x=1**	**x=2**	**x=3**	**x=4**	**x=5**	**x=6**
f(x)|H0	0.5	0.3	0.1	0.06	***0.03***	0.01
h(x)|H0	0.5	0.3	0.1	0.045	***0.03***	0.025

As we can see, they are both as extreme (improbable) given their corresponding distribution under H0. Their probability is equal to 0.03. However, the *p*-values are quite different (see Def1). The *p*-value associated with f(x)|H0 is equal to 0.04 and the *p*-value associated with h(x)|H0 is equal to 0.055. This means that we will reject H0 with the test involving f(x) but not with the test involving h(x) if the significance level[8] of our tests is equal to 0.05. We will do so even if the test statistic is as extreme (improbable) in both cases. This is incomprehensible and this problem is quite serious because it does not depend on any particular kind of distribution.

The incomprehension stems from two assumptions. Firstly, if an observation is too improbable (extreme) to keep f(x)|H0, then an equally improbable (extreme) observation under h(x)|H0 should lead us to reject h(x)|H0 as well. Secondly, we should not take into

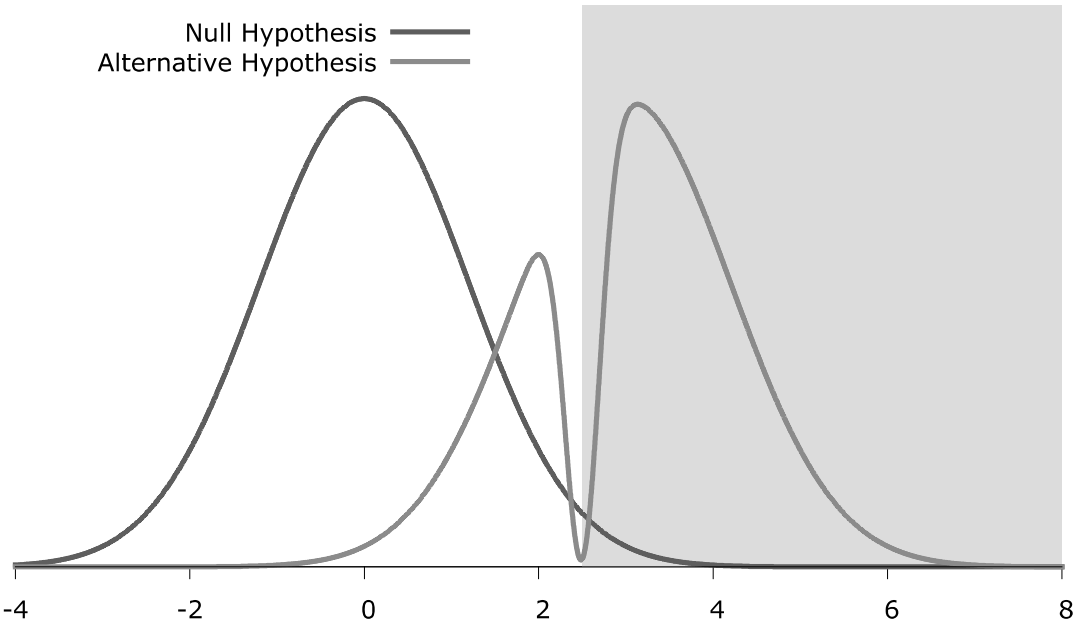
[7] Wagenmakers' article also provides references to other scientific work in which we can find the same argument.
[8] The significance level of a test (α for short) is the threshold that determines if a *p*-value is low enough to reject H0.

account the probability of an unobserved state (x=6) in order to implement our decision to keep or reject the hypotheses. Here is how Harrold Jeffreys ironically describes the situation: "What the use of *P* implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observation results that have not occurred. This seems a remarkable procedure" (Jeffrey 1961, 385).

The second assumption is obvious from a Bayesian or a likelihoodist perspective. From a Bayesian point of view we are interested in computing the posterior distribution over our hypotheses, given the actual observations only. From a likelihoodist perspective we are interested in comparing the probability of the actual observations given under competing hypotheses. But in the cases presented above we are not comparing H0 with any other hypotheses and we are not even interested in computing the probability of H1. In fact, from a frequentist point of view, hypotheses do not have probabilities. Clearly, we are facing different inferential paradigms and the frequentist one seems thick with paradoxes.

Now, we can further exploit the difference between the *p*-value and the probability of a test statistic in order to reach another kind of unacceptable result in a context where we would specify an alternative hypothesis –as one would do if she were following the Neyman-Pearson methodology. Here is an interesting example by Greco. Let us look at Figure 1 (the original figure from Greco's article) and imagine that our test statistic is 2.5 and that our *p*-value is less than 0.01.

Figure 1



According to Greco, while it is true that the probability of observing a statistic at least as extreme as 2.5 is very low according to H0 and very high according to H1, the actual probability of observing 2.5 is lower according to H1 than it is according to H0: "In fact, while it is quite unlikely on the null hypothesis that the test statistic should take the value of 2.5, it is even more unlikely on the alternative hypothesis" (Greco 2011, 622-23). Therefore, says Greco, the *p*-value leads to the wrong conclusion. The test statistics is clearly more extreme under H1…or is it?

Greco actually makes a mistake here. The fact is that the probability of observing a test statistic of 2.5 is the same under H0 and H1. It is equal to 0 because we are dealing with probability density functions. When a random variable is continuous, a probability is defined as an integral of the associated density function and the area under a curve at any point is

always 0. We can only determine to probability that our test statistic falls within a given interval.

More specifically, it is a mistake to interpret the value of a density function f(x) for a given x as a probability. A density function f(x) can take values that are much greater than 1. For example, a uniform distribution f(x) over the interval [0, 0.2] will always be equal to 5.

This does not mean that Greco's point is lost. We could easily imagine a 'discretised' version of H0 and H1 (where the mass distributed on each outcomes is a probability) in order to make the same point. However, Greco's mistake raises an interesting question. If every possible value that a continuous variable X can take is attributed a probability of 0, then how can we ever determine that one value is more extreme (improbable) than any other? How could (Def1) seriously be applicable at all? It looks as if it is impossible to define a reasonable critical region[9] for a test that involves density distributions. This 'puzzle' should be a serious hint that we might have been working with the wrong definition of 'extreme' all along.

In the next section, I solve this 'puzzle', and explain why 'extreme' was never meant to mean 'improbable'. I show that all of the problems that have been discussed so far dissolve if we take the time to understand the kind of inference we wish to make within (FA). I also revisit the 'fair die' experiment, to make my case more vivid.

My aim in this paper is not to show that (FA) embodies the best inferential procedure. I aim to show that (FA) stands on its own, *i.e.*, that it does not generate paradoxes of the kind I have presented in this section. Hence, I will not make a thorough comparison between (FA) and other approaches, like the Bayesian approach. Instead, I will reach three main objectives.

[9] A critical region is a set of extreme outcomes such that we would reject H0 if our test statistic belonged to it. If every possible outcome is as extreme as any other, then the critical region includes (or excludes) all of them, which is unreasonable.

Firstly, I will show why it is false to claim that if an observation is too extreme to keep f(x)|H0, then an equally improbable observation under h(x)|H0 necessarily lead us to reject h(x)|H0. Secondly, I am going to justify why frequentists need to take into account the unobserved values over which we define a distribution. Finally, I shall argue that a frequentist test will not suggest that we should both keep and reject H0, as it was implied by the combination of Greco's and Sober's critical comments. Doing so, I will propose a sound definition of the *p*-value.

3. Extreme values are not determined independently from a given distribution

The criticisms presented in the previous section all rest on the same fundamental mistake. To highlight it as clearly as possible, I will make an analogy by expounding a simple (non-probabilistic) inference about a distribution. Suppose that we wish to make an inference about the way in which a cooperative shares its profits. We assume that it shares them equally among its members. There are 100000 members and an amount of 10 dollars to share. Therefore, our assumption implies that everyone will receive 0.0001 dollar.

Naturally, we cannot falsify the assumption about the distribution of the profits simply by pointing out that a member has received the very small amount of 0.0001 dollar. We do not first determine what constitutes an amount of money that is too big or too small (extreme); make observations; and then make an inference about the way in which the cooperative distributes its profits. This is absurd. We cannot define the set of extreme amounts (anything above or below 0.0001 dollar) independently from the distribution that we wish to test.

The same goes for a frequentist test. We cannot define what is extreme regardless of the distribution that we are testing. For instance, consider f(x)|H0 as defined in Table 1. It would be mistake to stipulate that 0.03 is too improbable (extreme); observe x=5; and then infer that f(x)|H0 is an inadequate distribution because it implies that the probability of observing x=5 is 0.03. With such an inferential method, we would always reject every continuous distribution under H0 and every discrete distribution under H0 that has a mode (*i.e.*, the most probable outcome) with a probability that is smaller or equal to 0.03. In other words, we would be able to make inferences about mass or density distributions without even knowing any of their properties. That would be a remarkable procedure.

The mistake lies in the fact that when we take 'extreme' to mean 'improbable', we stipulate that there is a degree of improbability that qualifies an outcome as being extreme and that degree is determined independently from the distribution under H0 that we are putting under test. That mistake is being made when it is claimed that x=5 is equally extreme under both distributions in Table 1 because they are equally improbable. It is also committed when it is claimed that our observation is more extreme under H0 than under H1 in Figure 1 because it is less probable under H0. Furthermore, it is at the root of Greco's 'die experiment'.

In that example, he stipulates that S is too extreme (improbable) by standards that are independent from the distribution that he is testing; observes S and then rejects H0. Now, I am not disputing the fact that this is an incorrect inference. I am saying that this is not an inference that would be sanctioned by the frequentist approach.

In a frequentist test, H0 specifies the parameter(s) of the distribution of a test statistic.

The goal of the test is to figure out if the distribution of the mass or of the density[10] over the possible outcomes of our experiment is reasonable under H0. To achieve that goal, we need to make a prediction with the distribution under the assumption that H0 is true. But in many cases we cannot predict which value a random variable will take. We can define distributions for which the probability of each possible experimental outcomes is as close to 0 as we want.

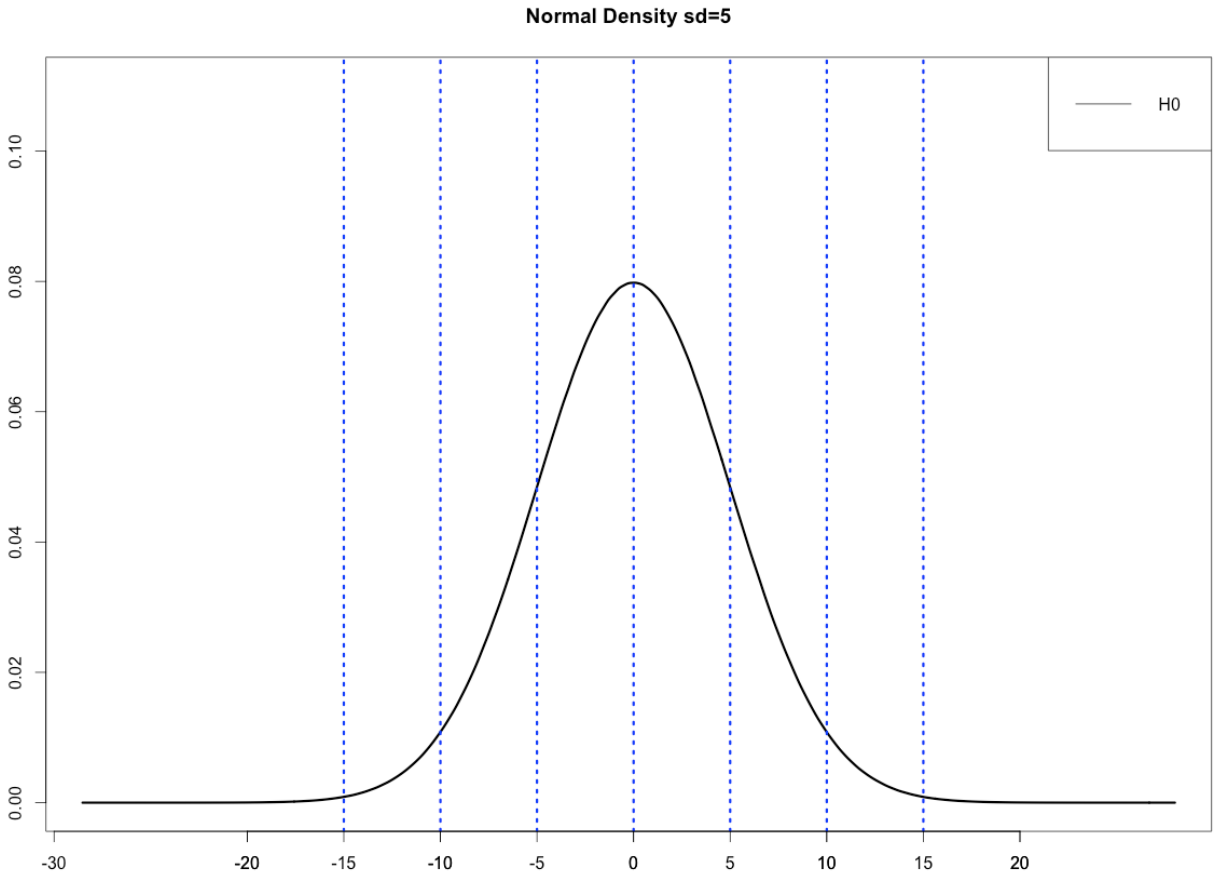
However, a distribution can allow us to predict a probable range of experimental outcomes. For example, we can predict with probability 1 that our experiment will yield one of the possible outcomes allowed by a distribution. This would be 100% accurate, but very imprecise. Of course, the idea is to strike the perfect balance between accuracy and precision.

In fact, when we have no idea about what would be the true density if H0 were false, a frequentist test consists in predicting the most precise (smallest) range of experimental outcomes, given a certain upper bound probability of error, under the assumption that H0 is true. If the result of an experiment falls outside the predicted range, then H0 is rejected. Extreme outcomes are thus defined as those who fall outside the most precise predicted range of outcomes for a given upper bound probability of error. This implies that we need to know the distribution that we are testing in order to define what is extreme. This was not the case when we considered extreme outcomes to be improbable outcomes.

Consider the following example. Figure 2 represents a normal distribution under H0 with a mean of 0 and a standard deviation of 5. The vertical blue lines determine three standard deviations from the mean.

[10] When we are dealing with discrete variables, we talk about the distribution of mass and when we are working with continuous variables, we talk about the distribution of density.

Figure 2



The random variable involved here is continuous and it is such that if we have a very precise instrument to measure the magnitude of this variable, then it is *certain* that we will observe something that is very improbable. Therefore, it would be a mistake to make an inference about this distribution based on the probability of our observation. It is simply impossible to make an interesting prediction about one outcome.

But if we examine the mean of this distribution and the dispersion of the random variable with respect to that central tendency, we see that the most precise predicted range of outcomes for a given upper bound probability of error will be centred on the mean. The

length of that range will vary according to the upper bound probability of error that we allow. Naturally, the extreme outcomes will lie in the tails of that distribution. In other words, an extreme observation in that case would be one that is too far from the mean of that distribution (*i.e.*, too deviant).

This is actually how R. A. Fisher analyses tests involving normal distributions (0,1). He makes it clear that the 'extreme' values are not those that are improbable (they are all improbable), but those that are too deviant:

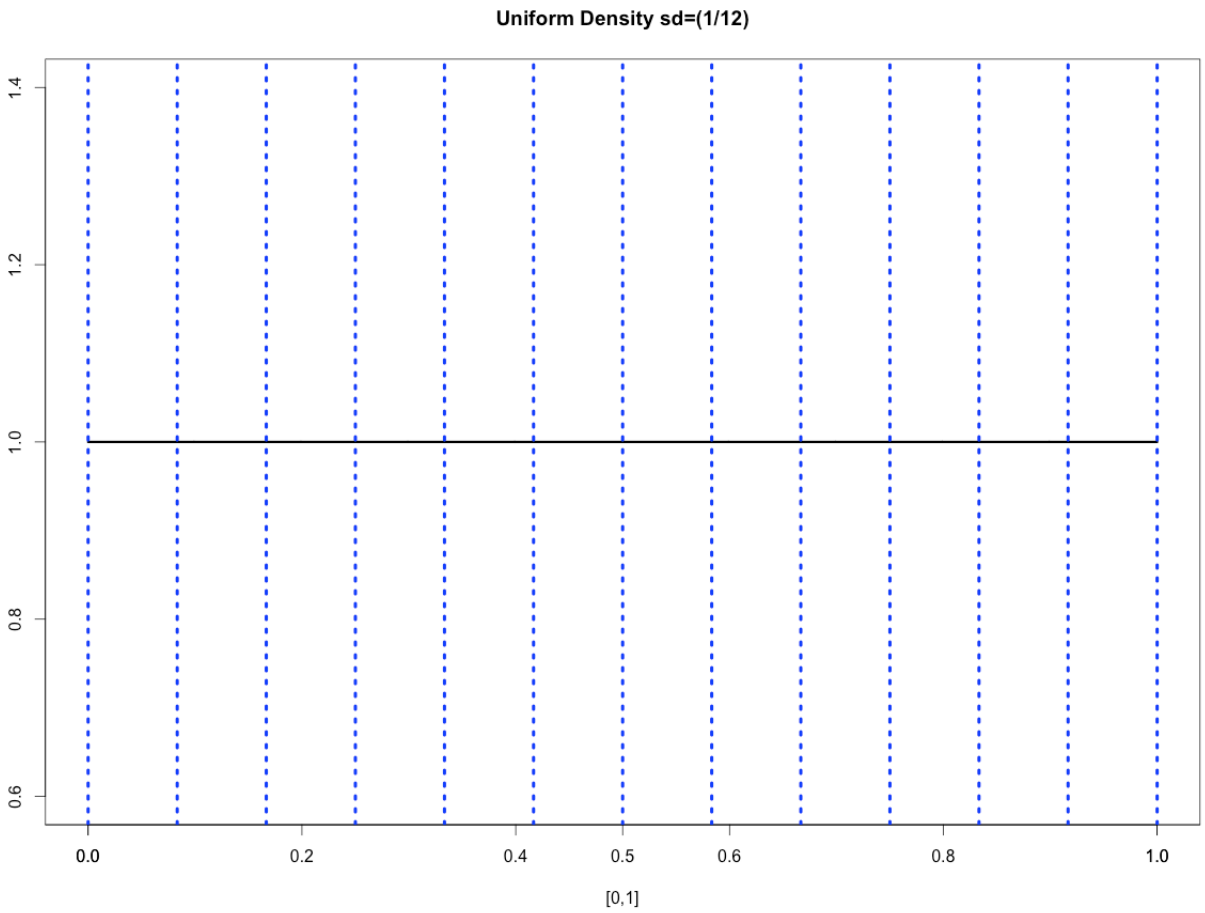
> Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which P =.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant (Fisher 1925, 47-48).

On the other hand, if we examine the uniform density distribution over the interval [0,1] that is depicted in Figure 3, we quickly realise that the random variable is equally dispersed around the mean (0.5) such that our experimental outcomes will have no tendency to be close or far from the mean within that interval. Each blue line determines 1 standard deviation from the mean and they all contain the same density.

This means that there is no such thing as the most precise range of prediction given an upper bound probability of mistake that is greater than 0. For example, if we set the upper bound probability of error to 0.05, we will find an infinity of ranges of prediction that are equally precise (of the same length). Only an observation that would fall beyond the interval [0,1] would lead us to believe that the distribution is inadequate. In such a theoretical scenario, the probability to obtain an observation that lies in the critical region under H0 is

zero. Such a result would be decisive against H0.

Figure 3



In sum, whether or not we are dealing with a density or a mass function, the inferential procedure is relatively simple. Firstly we identify the most precise predicted range of outcomes under H0 for a given upper bound probability of error. Secondly, we determine if our observation falls outside that range. If it does, then we reject H0. This inferential procedure is very different from the one depicted by Greco in section 2.

Now, it is possible to find the most precise predicted range for a given upper bound probability of error under H0 for any kind of distribution. For example, if the distribution is

asymmetric (see Figure 4), the required predicted range will be longer on one side of the mode and shorter on the other side. Moreover, if the distribution is multimodal, then we will find a union of disjoint predicted ranges.

Once we have made our predictions and performed our experiment, it is possible to determine what would have been the smallest upper bound probability of error that would exclude our observation from the predicted range. That probability will give us the *p*-value. If it is smaller equal to the initial upper bound probability of error, then we know that our observation fell outside our predicted range.

Under minimal assumptions about the alternative hypothesis, the *p*-value can thus be defined as follows:

(Def2) A *p*-value is the smallest upper bound probability of error under H0 that would exclude our observation from the most precise predicted range of experimental outcomes.

This definition is more precise than the more general definition given in the introduction. When we do not know anything specific about the alternative hypothesis, (Def2) gives the probability to obtain a result that is at least as extreme as the one that we have observed.

It is also important to notice that (Def2) gives a central function to random variables. They allow us to give precise measures of distance that allow us to determine the precision of our predicted range under H0. They also us to identify the most precise range of prediction because they make it possible to determine a variety of central tendencies and measures of dispersion. For example, without random variables there is no such thing as a mean and there is no measure of dispersion like a variance. It is therefore not surprising to see that some

textbooks straightforwardly define a statistical hypothesis as a statement concerning the distribution of random variables:

"A statistical hypothesis is a statement about the probability distribution of a *random variable*" (Hines et al. 2003, 266) [emphasis on 'random variable' added]

"**Definition 3**. A statistical hypothesis is an assertion about the distribution of one or more random variables" (Hogg & Craig 1995, 284) [emphasis on 'random variable' added][11]

4. Setting the Record Straight

4.1 The Strange Case of Table 1

Equipped with this corrected definition of the *p*-value and of the word 'extreme', we can now solve the problems expounded in section 2 more explicitly. If we look at Table 1, we can now easily explain why we would reject f(x) under H0 and not h(x) under H0. When we consider f(x)|H0, the most precise predicted range for an upper bound probability of error of 0.05 includes the elements in the following set: {1, 2, 3, 4}. Therefore, the element in the set {5, 6} are extreme values. Since we observe x=5, then we reject H0.

On the other hand, there is no such thing as the most precise predicted range for h(x)|H0, given an upper bound probability of error of 0.05 (we cannot find a unique one). The best we can do is to give the most precise range of prediction for an upper bound probability of error of 0.025. Accordingly, the predicted range would include the elements in the following set: {1, 2, 3, 4, 5} and x=6 would be the only extreme value. Since we observe x=5, we do not reject H0.

[11] That definition is more precise since there might be more than one variable involved in a statistical test.

Given the two tests, we would also reach the exact same conclusion with the *p*-value. Given that we have observed x=5, the smallest upper bound probability of mistake that would exclude x=5 from the most precise predicted range is 0.04 when we consider f(x)|H0 and 0.055 when we consider h(x)|H0. Thus for an upper bound probability of error equal to 0.05, we know that x=5 falls outside the predicted range when we consider f(x)|H0, but not when we consider h(x)|H0.

Obviously, under this interpretation, x=5 is not as extreme under f(x)|H0 and under h(x)|H0. Hence there is no paradox. It is not true that if we observe an extreme outcome under f(x)|H0, then we should also reject h(x)|H0 because what we observe is equally improbable given their respective distribution. The evidence that we have against a distribution under H0 is that we have observed something that falls outside the most precise predicted range for an upper bound probability of error of 0.05. That range is simply not the same for different distributions.

Furthermore, we can see why it is justified to take into consideration the probability of outcomes that are not observed in order to make an adequate inference. We need to consider the mass or the density over every possible outcomes in order to determine the predicted range and the range of extreme outcomes. Pace Jeffreys, there is nothing particularly counterintuitive about this procedure.

Here is an analogy to illustrate this. To evaluate an archer, we can measure just how much further from the centre the arrow could have hit an officially regulated target. If that distance is very small, then we can have an indication that the archer was not very good on this occasion. In other words, it makes a lot of sense to make an inference based on possible

(unobserved) hits.

4.2 The Puzzle of Figure 1

Moving on to another puzzle, we saw that Figure 1 sets the conditions for a Neyman-Pearson test where an alternative hypothesis H1 is defined. In that specific scenario, we know that the alternative hypothesis defines a distribution with a mass that is essentially located to the right of the distribution under H0. Under these conditions, we will impose a slightly different restriction on our predicted range under H0 and the definition of 'extreme' and 'p-value' will have to be slightly modified as well. This is because the assumptions about the alternative hypothesis are not minimal.

We can think of a Neyman-Pearson test as providing two mutually exclusive predicted ranges of possible outcomes. One prediction will be made under H0 and the other one under H1. Both ranges are constructed such that if an outcome does not belong to one of them, then it belongs to the other.

The conditions of the test are such that for a given upper bound probability of error under H0 we will maximize both the accuracy (not the precision) of our predicted range under H0 and the accuracy of our predicted range under H1. This equivalent to maximizing the probability to make a true prediction under H1 given an upper bound probability of error under H0. The point is to maximise the probability to obtain a correct prediction under H1 when H1 is true for a small probability to obtain an incorrect prediction under H0 when H0 is true.

Just like before, the extreme outcomes will be the ones that fall outside the predicted

range under H0 and the *p*-value will be the smallest upper bound probability of error that excludes the outcome that we have obtained from the predicted range under H0 that satisfies the previous restrictions. The rule of decision will be the same as before.

Now, Figure 1 shows that an extreme outcome is not necessarily less probable under H0 than under H1 (I am pretending here that the distribution is discrete and not continuous). But once we carefully explain the inferential procedure behind a Neyman-Pearson test, there is nothing particularly puzzling about that fact. Perhaps one might wonder if we could not minimize the upper bound probability of error under H0 by including our observation (2.5) in the predicted range under H0. The gain in accuracy under H0 would be greater than the loss of accuracy under H1. But this is not a problem for (FA). It simply suggests that the test can be improved. One cannot criticise an inferential approach with a poorly designed test. Hence, we can solve the puzzle presented[12] by Figure 1.

4.3 Is the Die Fair or Not?

Finally, it is now possible to understand why the 'fair die' test does not lead to a contradiction. In a nutshell, it was a mistake to rely on the improbability of the experimental outcome in order to infer that we should reject H0. That much was said in section 3. But the mistake runs deeper.

In fact, there is very little that we can do if the distribution of our test statistic under H0 is a uniform distribution. The only extreme outcomes are the ones that fall outside the

[12] I would like to point out that the puzzle is not very convincing. The only difference between the two distributions in Figure 1 should be a difference of parameters. It is not obvious to see what kind of parameter would create both distributions when we change its value.

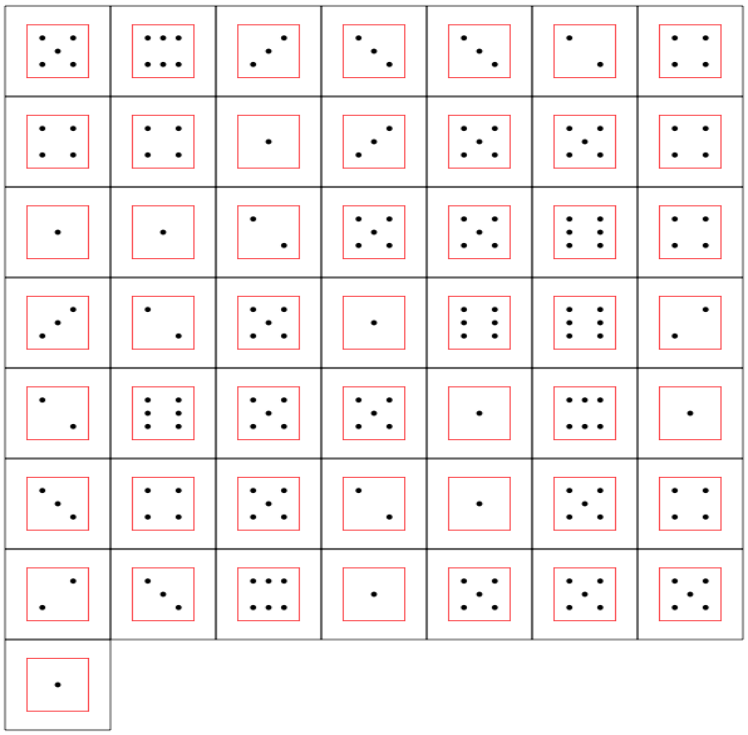
interval defined by its parameters (*i.e.*, results that should not happen). But an extreme result only mean that those parameters are wrong. In other words, an extreme outcomes implies that the die has more than 6 faces and this is not what we intend to test when we want to know if our die is fair or not. Not only did Greco misidentified the critical region of the test, but the test was not even the right one.

If we define our test correctly, our decision procedure will not suggest two incompatible courses of action to be taken at the same time. Here is how the experiment could be made. For starters, we roll a fair die 50 times[13] and obtain the following sequence S (the sequence should be read from left to right and from the top to the bottom)[14].

[13] Here I make 50 rolls instead of ten because it validates the following chi-square test and it makes every possible vectors very improbable.

[14] A computer simulation of a fair die generated the latter (see Annex).

Rolling a Fair Die X50 (simulation on R with the TeachingDemos package)



Secondly, we study the 'behaviour' of the random vector that corresponds to the possible frequencies of each possible values of the die. That vector follows a multinomial law that can be stated as such:

$$P(X_1 = x_1, \dots, X_6 = x_6) = \frac{50!}{x_1! \dots x_6!} {\pi_1}^{x_1} \dots {\pi_6}^{x_6}$$

Thus, what we have here is a multinomial experiment and our aim is to tell if our six parameters π_i are equal to $1/6$. The statistical hypothesis that we wish to test in this case can be written as follows:

$H0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = 1/6,$

and our observed statistic is (9, 7, 7, 7, 13, 7).

At this point, what is important to realise is that all the vectors that we might observe will be very improbable when H0 is true. When we maximise the mass function, we obtain a probability of 0.0001081195 (see Annex). Thus, we are one step away from reaching one of the disastrous conclusions that we inferred in section 1. Indeed, if we consider the probabilistic modus tollens; take 'extreme' to mean 'improbable'; and consider 0.0001081195 to be 'too improbable'; then we will always reject H0 because every possible vector will be at least as improbable as 0.0001081195.

But if we interpret 'extreme' as 'outside the most precise range of prediction for a given upper bound probability of error', then we safely avoid that disastrous conclusion. What H0 implies is *that our observation should not be 'too far' away from the mode of the multinomial law* because our range of prediction will be centred on that mode. Therefore, we have to focus our attention on the improbability of our observations to be at least as distant from the mean under H0 as the values of the possible experimental outcomes associated with the significance level α, as opposed to the probability of our observations *per se*. In this case, our observation does not fall outside the predicted range.

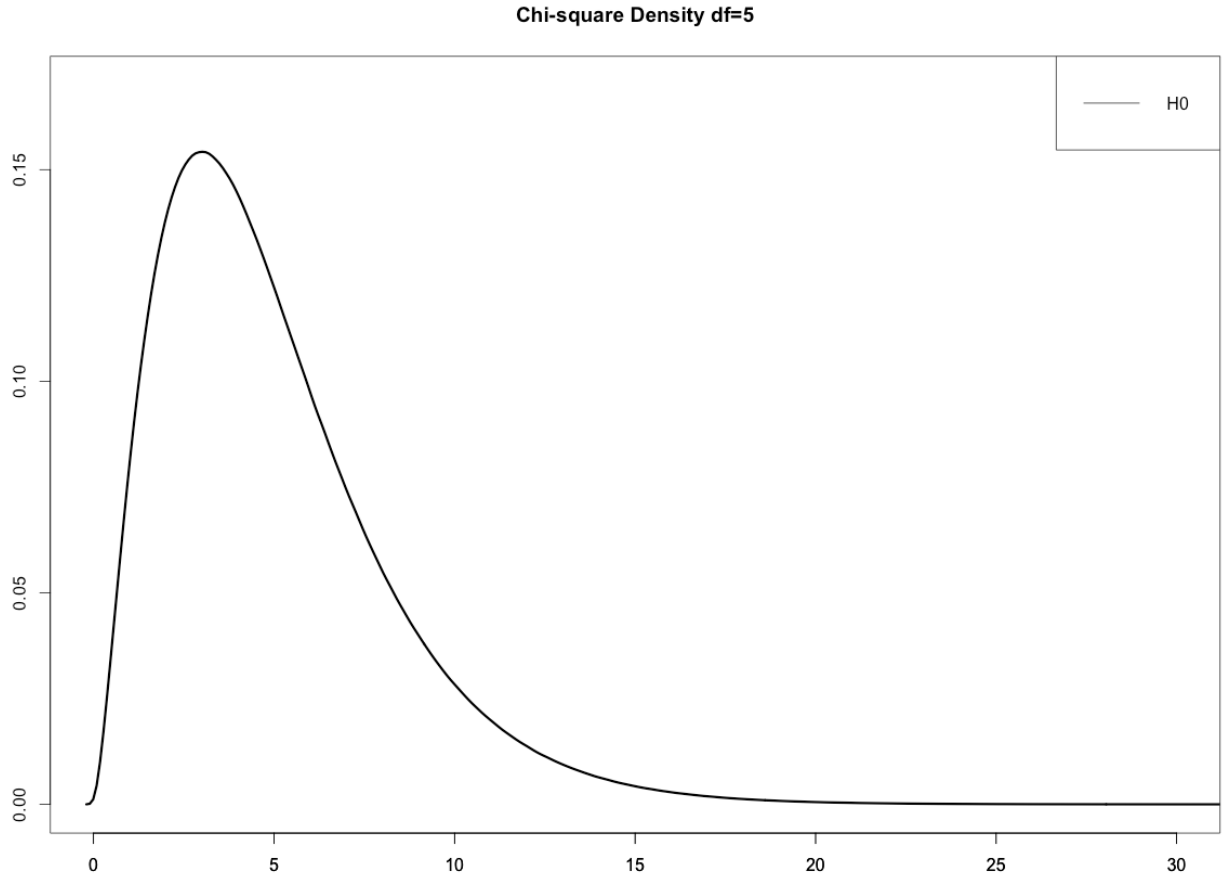
In order to show this (things are easier to visualise in 2 dimensions), we can actually take a shortcut and study a different test statistic by using a "Goodness of Fit" test. Such a test examines the difference between the frequencies that we are supposed to observe under H0 and the frequencies that we actually observe. If we do that test (see Annex), we obtain a chi-

square statistic of 3.52 with 5 degrees of freedom (df=5)[15], and a *p*-value of 0.6204.

According to (Def2), we do not reject H0 with a significance level of 0.01 because the *p*-value indicates that our observation does not fall outside the predicted range. To convince ourselves of this, we can look at Figure 4, which shows a chi-square density function with 5 degrees of freedom. We can immediately tell that the most precise predicted range will be on the left side of the positive real number axis and that it will be of a longer length on the right side of the mode than on the left side of the mode. We also tell that 3.52 will sit comfortably within the predicted range for a reasonable chance of error (not too big).

[15] We define the distribution with 5 degrees of freedom because once we have counted the observed frequencies for 5 dimensions of our random vector, we can simply deduce the frequency associated with the remaining dimension.

FIGURE 4



In short, what I have shown is that we cannot avoid the hard work of studying the central tendencies and the dispersion of the distribution under H0 in order to make a 'fair die' experiment. It is the only way to find out if our test statistic is too deviant for the distribution (whether it is discrete or continuous) to be reasonable. Of course, as my examples show, the probability of our test statistic under H0 can be very low. But this is to be expected under H0. Therefore, this information does not provide any evidence against the adequacy of the distribution under H0.

5. Conclusion. In this paper I have argued that it is a serious mistake to interpret 'extreme' to mean 'improbable' as we try to analyse the nature of frequentist statistical inferences. That mistake is the source of many critical arguments against the frequentist approach. In this paper, I have argued that we can dismiss such arguments by explaining why 'extreme' cannot possibly mean 'improbable' and why it should mean 'outside the most precise predicted range for a given upper bound probability of error', when we make minimal assumptions about the alternative hypothesis.

One of the key ideas that I have put forward is that we cannot define what is extreme independently from the distribution that we are considering. I have also given an appropriate definition for the *p*-value. Doing so, I have also stressed the importance of random variables in order to give precise measures of dispersion and of central tendencies. They are essential to establish the precision of our predictions.

For the most part of this paper, I have worked under a Fisherian framework (where there is no formal treatment of the type-II error). I have also assumed that we do not know anything specific about the alternative hypothesis. I did so because those are the conditions under which most the criticisms presented in section 2 are defined. But I have also explained how to extend my conclusion to a context where we make stronger assumptions about the alternative hypothesis. As discussed in section 4.2, the resulting definition of 'extreme' and '*p*-value' will essentially stay the same but the restrictions on the predicted ranges will be slightly different.

Overall, my aim was to show that the frequentist approach to theory testing was not undermined by the kind of paradoxes that were presented in section 2. In that respect, I argued that it was not internally incoherent. But whether or not this approach provides a

good epistemic framework in comparison with other inferential paradigms remains open for debate. This is a difficult question because the frequentists do not assign probabilities to hypotheses. Nevertheless, it would be interesting to study the extent to which an adequate epistemic interpretation of frequentist tests and of the *p*-value would yield compatible (incompatible) epistemic judgments in comparison with other approaches, such as the Bayesian approach. This is a topic for a future work.

Annex

Here is the program that I used to obtain S with R:

```
  library(TeachingDemos)

dice(rolls=50, ndice=1, sides=6, plot.it=TRUE)
```

Here is the program that I used to perform a 'Goodness of Fit' test with R:

```
 vect=c(9,7,7,7,13,7)

vectprob=c(1/6,1/6,1/6,1/6,1/6,1/6)

chisq.test(vect, p=vectprob)
```

Here is the program that I used to maximise the multinomial mass funtion with R:

```
a=factorial(50)

  b=factorial(8)

  c=factorial(9)

denom=(c^2)*(b^4)

d=(a/denom)

frac=1/(6^50)

d*frac
```

References

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Greco, D. (2011), "Significance Testing in Theory and Practice," *British Journal for the Philosophy of Science*, 62: 607–37.

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

Hines, W. W. et al. (2003), *Probability and Statistics in Engineering*, New York: Wiley, fourth edition.

Hogg, R. V. and Craig, A. T. (1995), *Introduction to Mathematical Statistics*, Englewood Cliffs, N.J.: Prentice Hall, fifth edition.

Jeffreys, H. (1961), *Theory of probability*, Oxford: Oxford University Press.

Sober, E. (2008), *Evidence and Evolution*, Cambridge: Cambridge University Press.

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problems of *P* Values", *Psychonomic Bulletin & Review*, 14: 779-804.