A Crisis of Confidence:

Adjusting p-values for Multiple Hypothesis Testing

Paul van Rensburg

Frank Robb Professor of Finance

University of Cape Town

paul.vanrensburg@uct.ac.za

Second draft. Comments welcome.

Abstract

This paper addresses, perhaps, the most important issue in empirical science today: the 'replicability crisis' that arises when using a p value <0.05 in multiple hypothesis testing. To define a 'crisis', a benchmark of expected non replication is required and the p value is flawed in this regard. The Non Replication Rate (α^*) is the ratio of false positives to all positives rather than false positives to the total number of tests (α). Using Bayes Theorem it is demonstrated that $\alpha^* = P(H_0|D)$ rather than $\alpha = P(D|H_0)$. The former unlike the latter is actually a test of the null hypothesis. As illustrated through an example in the field of asset pricing, the conventional p value associated with a pre-specified Non Replication Rate (α^*) is calculated and used as an adjusted 'hurdle of significance' for multiple hypothesis testing.

Introduction

Contemporary science is characterised by Big Data and the low cost of computer power. This allows multiple hypothesis testing to be rapid, cheap and indiscriminate. The empirical documentation of correlation has dominated over deliberation over causation and, across disciplines, a substantial number of the relationships identified as statistically significant have subsequently not been replicated. There is a currently a "*crisis of confidence*" in the statistical methods we employ. A concise survey of selected highlights of contemporary science, from the macro to the micro level, reflects the large amounts of data that are currently being processed and the multiple hypotheses being tested:

For 10 days over the Christmas period in 1995 the Hubble space telescope took a picture of a dark area of the sky roughly the size of a pinhead being held at arm's length. Within it approximately 3 000 galaxies were observed. With the naked eye a person can see about 2000 distinct stars in the clear night sky - all of which are in our own galaxy. Consider that, in a pinhead of the sky there are more galaxies than this. Such was the richness of data collected that it was shared as a public service and by 2014 there were over 900 citations that referred to the paper associated with this data set (Williams et al, 1996).

The SETI@home project produces 100 to 200 terrabytes of data daily for enthusiasts to analyse and search for signals of extra-terrestrial intelligence. In 2006, it performed the largest calculation in history comprising 10^{21} floating point operations (Guinness Book of Records, 2008, https://en.wikipedia.org/wiki/SETI@home). The most well known radio signal from space is the 1977 "*Wow!*" observation made by a volunteer at Ohio State University. Despite repeated searches, this signal has not been replicated. Indeed, this field of enquiry appears to have had a 0% replication rate, so far.

Since its launch in 2009, the Kepler spacecraft has been dedicated to search for planets outside our solar system. By October 2017, Kepler it had identified 5011 exoplanets, of which 2512 have since been confirmed (NASA Exoplanet Archive, 2019). A false positive rate of about 50%. It can be noted that this area of research beneficially adopts the approach of "*triangulation*' to confirm its hypotheses. Not only are three observations of a shielding of light from the orbited sun required for a "*transit dip*" to be recorded but also evidence of a gravitational wobble in the sun's location can be used for a "*radial velocity*" confirmation. In this example, the planet's existence is indicated through two independent mechanisms (gravity as well as light).

Large data sets also exist on the micro level. The human genome has been found to consist of about 3.3 million base-pairs within which there are 22300 protein producing genes. The field of genetic epidemiology, which attempts to relate these genes and their combinations to the occurrence of diseases, has become notorious for non-replication. Ioniddas's (2005) widely cited paper "*Why most research findings are false*" was prompted by the widespread non replication problem experienced in this field.

In July 2012, after some 6 quadrillion collisions in the Hadron Large Collider (LHC), the discovery of the Higgs Boson was announced. Its decay pattern exists for a zepto (10^{-21}) second - about the time it takes for light to travel across the length of a hydrogen atom. To do the large amount of hypothesis testing involved in processing this data, the Worldwide LHC Computing Grid was developed. In 2017 it incorporated over 170 computer centres in 42 countries. By 2012, 25 petabytes of data per year was being produced by the LHC (http://wlcg.web.cern.ch). Notably the Higgs boson discovery was only formally announced when three independent teams had observed the same decay pattern, each as a 5 sigma event (https://en.wikipedia.org/wiki/Higgs_boson).

In our home discipline of finance, a "*zoo*" of factors explaining the cross-section of equity returns has been documented (Cochrane 2010). Mclean and Pontiff (2016) look at 97 previously documented anomalies. They find "*Portfolio returns are 26% lower out-of-sample and 58% post publication*". Harvey, Liu and Zhu (2016) re-evaluate 296 previously documented anomalies: *"We argue that most claimed research findings in financial economics are likely false".* Hou, Xue and Zhang (2017) examine 447 previously documented anomalies. Using updated (but overlapping) data they find a 64% non-replication rate at the 95% level of confidence.

Across almost all disciplines, non-replication has been well in excess of the 5% "*expected level*" when applying a p value of 0.05. Among the consequences of this "*replication crisis*" are, for example, that the journal "*Basic and Applied Social Psychology*" stopped publishing papers using p values in 2015. The American Statistical Association (2016, 4) was sufficiently moved to put out a statement on p values "…*intended to steer research into a post p<0.05 era*". However, aside from stressing caution in their interpretation, this statement contained very little substance as to how this problem could be corrected. The response by researchers has essentially been to raise the hurdle rate. The field of physics has used five "sigma" (effectively a t statistic of 5) since the mid 1990s. In a recent paper in Nature co-authored by over 70 academics a argument is made to "*Redefine Statistical Significance*" (Benjamin et al, 2017). The proposition they all signed up to is to increase the hurdle p value from 0.05 to 0.005. They acknowledge that this is not a complete solution to the problem but it is a practical step that increases the likelihood of replicability.

The aim of this paper is ascertain the correct adjustment for conventional p values in multiple hypothesis testing to address this 'crisis of non-replication'. In order to do so, it attempts to calculate the p value associated with a pre-specified Non-Replication Rate.

The paper is organised as follows. Section 2 reviews prior approaches and Section 3 considers their shared weaknesses. Moving towards a solution, Section 4 proceeds by addressing the most fundamental of these: standard p values give you the probability of observing the data given the null hypothesis is true, $P(D|H_0)$, and not more relevant $P(H_0|D)$. Bayes Rule is used to flip these conditional probabilities. It is argued that the Non Replicability Rate is equivalent False Report Probability Ratio of (Wacholder et al, 2001) and it is analytically demonstrated that this measure is also $P(H_0|D)$. The switching of conditional probabilities requires the input of a prior i.e. $P(H_0)$. In Section 5 the case for a 'positivist' Bayesian approach is made whereby priors are empirically estimated or analytically derived. It is argued that there is no need for subjective probabilities to enter the prior and that its omission is not the avoidance of prejudice but an error of logic. Indeed, the implicit application of the uniform prior in conventional hypothesis testing when its true value is likely be much lower due selection bias and overfitting is a key source of the 'replication crisis'. It is demonstrated how, in multiple hypothesis testing, the prior can be empirically estimated knowing the total number of tests conducted (M), the power of the test (1-β), the level of significance used (α) and the portion observed to be statistically significant at the threshold α, P(D)). Essentially, the approach developed in this paper provides the *conventional p value, P(D|Ho), that is consistent with a prespecified Non-Replication Rate= P(Ho|D) =α**. As a practical example, section 6 proceeds to do so on a case of multiple tests in the field of asset pricing i.e. Hou, Xue and Zhang's (2017) re-testing of 447 previously documented equity market anomalies. Section 7 summarises and concludes with a consideration of some of the implications for an effective scientific methodology going forward.

Additionally, based on the work of Raftery (1995) and Wagenmakers (2007), an appendix provides a concise overview of a practical Bayes Factor hypothesis testing approach that can be used alongside t statistics in order to address its primary remaining problem problem (that of neglecting the magnitude of an effect size in favour of the relative accuracy of its estimation). As this approach it only requires standard regression output, it avoids many of the computational difficulties usually associated with calculating Bayes Factors.

2. Prior Approaches to Adjusting the p-value for Multiple Hypothesis Testing

The approaches most commonly applied are based on the Bonferroni (1936) adjustment:

$$\alpha^* = \frac{\alpha}{\mathrm{M}}$$

Where: α = the level of significance applied to a single test (usually 0.05)

M = the number of hypotheses being tested

α^* = the adjusted level of significance used for hypothesis testing

The Bonferroni approach aims to keep the "*family wise error rate*" i.e. the probability that all M null hypotheses in the group being tested are true simultaneously at the α level. The adjusted significant level in order to do so is $1-(1-\alpha)^{\frac{1}{M}}$. As an approximation, the Bonferroni adjusted hurdle of significance is simply the level of significance applied to a single test divided by the total number of tests conducted. This simple adjustment can only be justified by fairly restrictive assumptions and is likely to be extremely punitive in situations where a large number of tests have been conducted. Note that is one was to apply the equivalent approach of multiplying up estimated p values by M, it would be quite possible to get p >1.

The Holm (1979) adjustment follows a similar theme and is relatively more lenient than that of Bonferonni:

$$\alpha_m^* = \frac{\alpha}{\mathrm{M}+1-\mathrm{m}}$$

Ranking the tests from lowest to highest p values m=1,…, M, each test m has its own hurdle rate. The first test (with the lowest p value) would have the same hurdle as Bonferonni. However, for the second, the first test is effectively 'removed from the pool' and the divisor becomes M-1. This procedure is repeated down the ranked list of tests. The first test that does not meet the hurdle is rejected together with all the remaining tests that have higher p values.

With a critique that also applies to the Holm (1979) adjustment Pergneger (1998, 1236) points out that *"…Bonferroni adjustments are concerned with the wrong hypothesis. The study wide error rate applies only to the hypothesis that the two groups are identical on all … variables (the universal null hypothesis) … Such information is usually of no interest to the researcher who wishes to assess each variable in its own right*". Catering for this critique, unlike the Bonferroni and Holm adjustments which aim to avoid a single false discovery, the Benjamini and Hochberg (1995) adjustment allows an expected error rate of α:

$$\alpha_m^* = \frac{\mathrm{m}.\,\alpha}{\mathrm{M}.\,\mathrm{c(M)}}$$

Where: $c(M) = \sum_{m=1}^{M} \frac{1}{m}$

Again, each test has its own hurdle rate. Especially when M is large, it is generally more lenient than the two former approaches. Harvey, Liu and Zhu (2015) employ an equally weighted composite of the above three measures to calculate an adjusted t value of 3 for current asset pricing tests.

3. Weaknesses in Prior Approaches

There are certain key problems that all of these adjustments share.

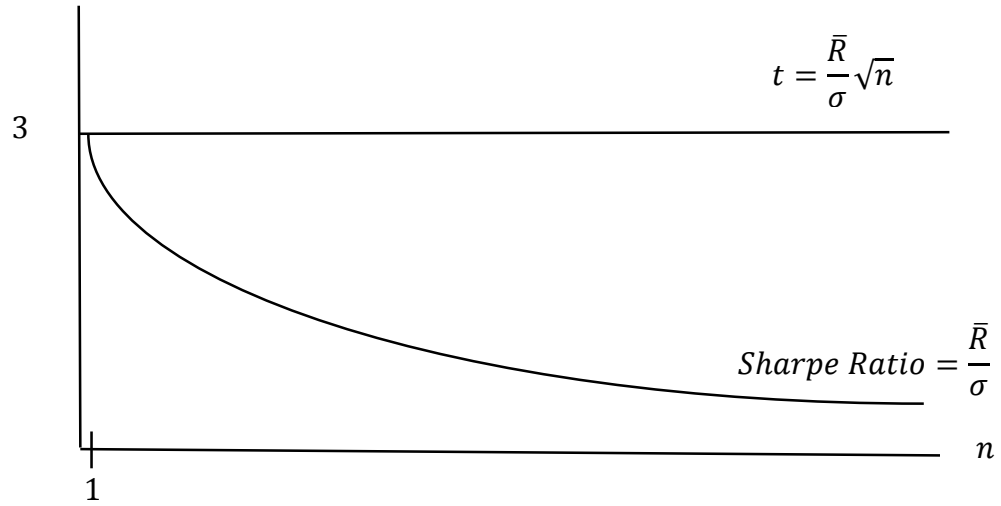
First, unlike the approach advocated in this paper, these adjustments takes no account of Type II errors and the resulting decrease in the power of the tests after the adjustment. Adjustments intended to be conservative for the Type I errors could well be reckless for Type II errors. In many cases, the Type II errors generated are likely to be more serious than the Type I errors that they are intended to prevent.

Second, there is no account of the correlations between tests. For example, various value measures (e.g. earnings yield, dividend yield, book to price) all involved dividing by price and, thus, by construction are likely to be correlated. In such a realistic situation, the number of independent tests is likely to be correspondingly overstated. Recognising this problem, Harvey and Liu (2014) conduct simulation studies to investigate the effect of correlated explanatory variables.

Third, the total number of tests conducted (M) is rarely directly observable. Due to a selection bias toward significant results, many tests that have been conducted but have turned out insignificant have not been published. Basing a tally on prior published research is likely to vastly understate M. However, when conducting an out of sample test of multiple hypotheses, M is directly observable and this fact is used later in this paper. This direct observability should not be incorrectly denied and conflated with the fact that the prior in the out of sample is test will be lower the more that the initial selection of M is based on noise (see Section 5). As will be seen, ignoring the prior leads to a number of logical errors such as this.

Fourth, p-values based on t statistics have the problems of misrepresenting effect size. To Illustrate, Harvey et al's (2015) suggestion of raising the hurdle for significance to a t-statistic of 3 is used. Reading a table of p-values associated with this t statistic would imply that one would have more confidence in the existence of an effect as the number of observations increases. The relationship between the t statistic and Sharpe Ratio of an investment strategy as the number of observations increases is illustrated in Figure 1 below.

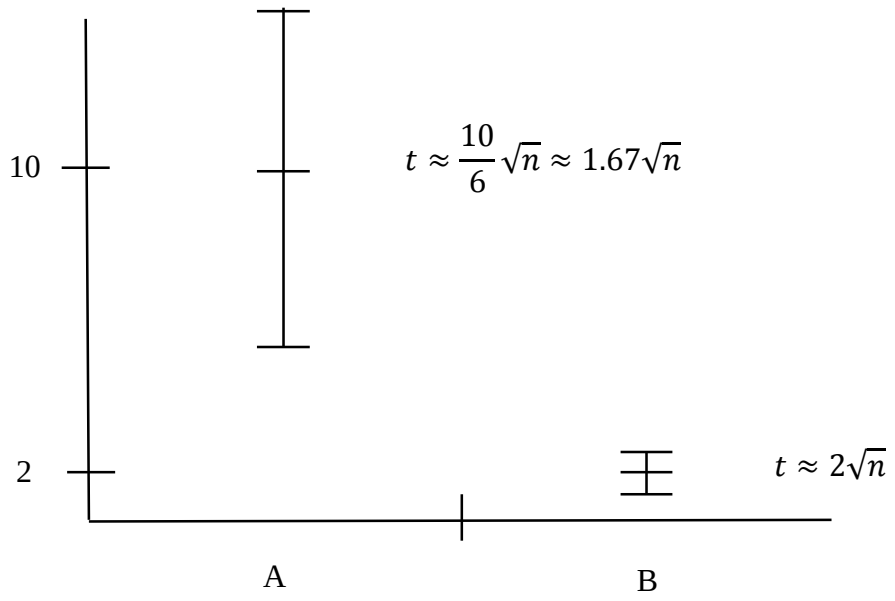
Figure 1. The Sharpe Ratio under a constant t statistic as the number of observations (n) increases



It is clear that, for a constant t-statistic (3 in this example), as the number of observations increases the Sharpe Ratio i.e. the strength of the strategy actually gets weaker! For a given t statistic, the Sharpe Ratio needs to decrease in order to compensate for the effect of additional observations. Absurdly, the p value of the strategy will get more significant as it does so.

The companion, Figure 2 considers the alternative situation where the number of observations is kept constant across examples A and B. It is clear that case B has the higher t statistic due to the lower proportional estimation error around its mean. However, even in a worst case scenario, A outperforms B. This illustrates that even when the number of observations are constant, the t-statistic takes inadequate account of effect magnitude. Given a very small effect size, the more certain you are that it is, indeed, very small – the more statistically significant it gets!

Figure 2. Effect Magnitude and t statistics with a constant number of observations. In both case A and B there are the same number of observations (n). In Panel A the expected effect size is 10 with a best case scenario of 16 and a worst case of 4 (implying a mean absolute deviation of 6). In Panel B the expected effect size is 2 with a best case scenario of 3 and a worst case of 1 (implying a mean absolute deviation of 1). For the purposes of illustration, the mean absolute deviation is taken as a proxy for the standard deviation in order to calculate a t statistic in each scenario.



Fifth and most fundamentally, all of the above measures is that the commit the "*standard error of science*". An explanation of this error requires returning to the first principles of hypothesis testing. Reconsider a streamlined version of Fisher's (1935) famous example: Can Dr Muriel Bristol (her real name) tell the difference if her tea is made by adding the milk first or not? If five cups are used in the taste test and she gets all 5 right there is a $1/(2^5) = 1/32 = 3.125\%$ chance that we would get this result (the observed data, D) under the H_0 of her having no skill i.e. $P(D|H_0)$. However, upon careful consideration, $P(D|H_0)$ is not particularly useful

to know in isolation and, indeed, is quite a clumsy concept. It seems to be almost universally interpreted as the much more useful $P(H_0|D)$ which, in contrast, is actually measure of whether the null hypothesis in true or not. Indeed, $P(H_0|D)$ rather than the reported $P(D|H_0)$ is what we really want to know and any adjustment to p-values (be it analytic, bootstrapped or ad hoc) that does not account for this is fundamentally flawed at its core.

4. The Bayesian Flip

Bayes Law allows us to switch conditional probabilities. When applied directly to the problem of hypothesis testing:

$$\mathrm{P}(H_0|\mathrm{D}) = \mathrm{P}(\mathrm{D}|H_0)\frac{\mathrm{P}(H_0)}{\mathrm{P}(\mathrm{D})}$$

Where: $\mathrm{P}(H_0)$ = the "prior" (before getting data D)

$\mathrm{P}(\mathrm{D}|H_0)$ = the "likelihood" (conventional p-value in this application)

$\mathrm{P}(\mathrm{D})$ = $\mathrm{P}(\mathrm{D}|H_0)\ \mathrm{P}(H_0)$ + $\mathrm{P}(\mathrm{D}|\mathrm{not}\ H_0)\ \mathrm{P}(\mathrm{not}\ H_0)$

$\mathrm{P}(H_0|\mathrm{D})$ = the "posterior" (after getting data D)

However, in order to conduct the 'flip' additional information in the form of an estimate of the probability of H_0 being true, $\mathrm{P}(H_0)$, prior to receiving the new data (D) is required . Note that P(D) is a normalising term that makes the posterior probabilities add up to one and does not provide any new information to the analysis. Also note that only in the case of an uniformed or uniform prior, where: $\frac{\mathrm{P}(H_0)}{\mathrm{P}(\mathrm{D})} = 1$, is conventional hypothesis testing valid as $\mathrm{P}(H_0|\mathrm{D}) = \mathrm{P}(\mathrm{D}|H_0)$. This is plausible for an initial investigation proceeding from a position of ignorance but, in the highly likely scenario where this implicit prior does not hold, this procedure will simply give incorrect results for $\mathrm{P}(H_0|\mathrm{D})$. In the dichotomous case used here for simplicity, an implicit prior of a 50% chance of the null hypothesis being true, is consistent with classical hypothesis testing. Given that selection bias and overfitting are endemic in choosing the hypotheses that get tested, it is likely that in many cases the true prior is much lower than this. Ignoring the prior -which is effectively implicitly using the uniform prior - is the underlying source of the non-replication problem in multiple hypothesis testing.

Incorporation of the prior has been the most controversial feature of applying Bayes Law as it has been criticised as a means for subjective biases to enter the analysis. This emphatically need not be the case. The "*Positivist Bayesian*" approach advocated in this paper can use (i) empirical analysis or (ii) definitions and mathematical relationships to establish a prior and does not require the use of subjective probabilities.

An example of (i) using empirical analysis to estimate the prior by doing tests on a broad based random sample of people to estimate that the base rate for a certain disease in the overall population, $\mathrm{P}(H_1)$, is say 1%. Combined with $\mathrm{P}(\mathrm{D}|H_1)$ i.e. the probability of getting the data of a positive test result for the disease given that you have it, it allows calculation of the probability that you have the disease given the test result $\mathrm{P}(H_1|\mathrm{D})$. Omitting the prior in this calculation is, as mentioned, simply an error of logic. To illustrate, if the p value of the test is 0.05 and the base rate is 0.01, even if you have a significantly positive test result, the odds are effectively 5:1 (83.19% to be precise) that you do not have the disease (for simplicity, a power of 1 for the test is assumed in this example. Figure 3, to be discussed in Section 5 below, visually depicts these probabilities and discusses the formula for the False Positive Report Probability). Given these odds, it is no wonder that there will a high chance of non-replication! Researchers seeing high rates of no-replication of this test may be moved to make ad hoc increases the hurdle p value or do elaborate simulation studies to inform them as to the magnitude of the appropriate hurdle. However, the underlying source of the problem is ignoring the prior base rate.

Importantly, estimating the base rate is just as valid an empirical analysis as the conventional estimation of $\mathrm{P}(\mathrm{D}|H_0)$. Note the prior need not imply that it is the probability before getting *any* data. It is the prior before getting the additional data provided by the test concerned. Furthermore, as more data accumulates and follow up tests is conducted, the initial prior becomes less relevant as Bayes Theorem provides a valuable mechanism to rationally update the prior based on new evidence. For example, given the initial positive result, a second independent test for the disease would commence from a 17% rather than 1% initial prior. In this case, if a second positive result was found, the chance of a false positive diagnosis drops to 19.62%.

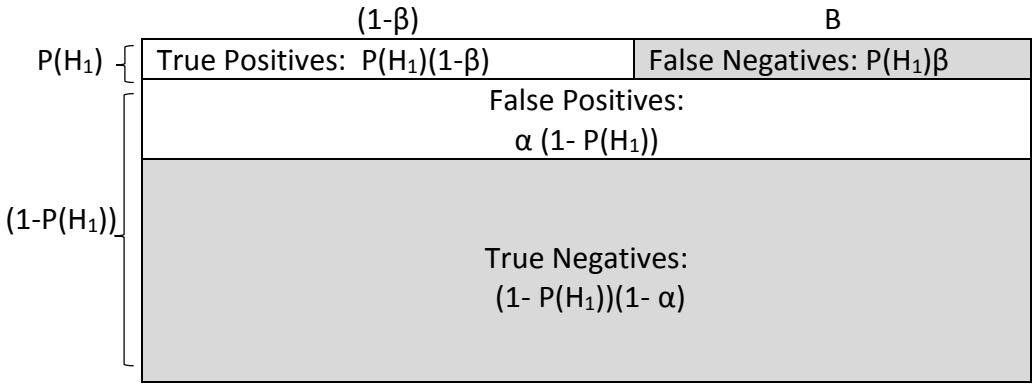
An example (ii) of using an apriori prior is the assumption that you are using a standard pack of playing cards. Thus, for example, you can state the base rate probability of randomly drawing a heart is 25%, based on the definition of what constitutes a standard pack. If you are provided with the data that the card is red you can revise your probability that the card is a heart to 50%. Note that using the uniform prior for obtaining a heart i.e. that $\mathrm{P}(H_1) = \mathrm{P}(\mathrm{D})$, the inference of 100% would be obtained . This is actually $\mathrm{P}(\mathrm{D}|H_1)$ i.e. the probability that the card is red given that it is a heart which is simply incorrect.

5. An Empirical Prior

It is possible to provide an empirically based estimate of the prior in multiple hypothesis testing as follows. **Table 1** considers the breakdown of test results against true relationships. For simplicity, it is assumed that all of the tests in the batch have the same power and α level. The total area represents all the tests conducted in the set (M). Here, $P(H_1)$ represents the unobserved proportion of cases where there is a true relationship i.e H_1 =True. In certain batches, where a field is well established, prior research may result in an informed selection of candidate variables to test and $P(H_1)$ may be of a high value. In contrast, exploratory studies may have a lower base rate. The correlations between the null hypotheses tested will also affect $P(H_1)$.

Figure 3. Test Results and True Relationships

The total area of the rectangle below represents the total number of tests conducted in this batch (M). Where $P(H_1)$ represents the unobserved rate of rate of cases where there is a true relationship, α refers to the level of significance used and (1-β) to the power of the test. The unshaded area represents the portion observed to be statistically significant at the level α, P(D)).



The top row of Figure 3 represents the proportion of true relationships, $P(H_1)$. Some will provide positive test results (True Positives) and others negative test results (False Negatives). The proportion of each is determined by the power of the test (1-β) where β represents the probability of a Type II Error (i.e. rejecting the null hypothesis when it is true). The area below the top column $(1-P(H_1))$ represents all cases where there is no true relationship. This is broken down into False Positives and True Negatives. The proportion of each is determined by the Type I error rate (α) of the tests.

Based on an analogous 2x2 grid of True and False, Positives and Negatives, Wacholder *et al* (2004) derive a ***False Positive Report Probability (FPRP)***. This represents the probability that there is no True relationship when a Positive test result is observed. Thus, it represents the False Positives divided by the total number of Positive test results (which is False Positives plus True Positives as represented by the unshaded area in Figure 1):

$$FPRP = \frac{\alpha(1 - (P(H_1))}{\alpha\big(1 - P(H_1)\big) + P(H_1)(1 - \beta)}$$

This is the formula used in calculating the probability of having a disease example of the previous section. Unlike the conventional α value –this measure is divided by the probability of getting a positive result. It thereby has the informative interpretation as the "

Non-Replication Rate" i.e. what proportion of **positive** (not total) results will be false.

The influence of the power of the tests (1-β) has been neglected by the prior p value adjustments reviewed in Section 2. Here it can be seen that as the power of the tests weaken less true positives are observed and the denominator of this expression gets smaller. This increases the proportion of false positives despite a reduction in the total number of positives reported. As an aside, it is also noted that when positives are reported in tests with weak power that they are likely to overestimate effect size as larger random observations are required to pass tests of statistical significance.

The FPRP can be brought into a Bayesian framework by defining $P(D)$, where the data (D) is a positive test result:

$$P(D) = \alpha\big(1 - P(H_1)\big) + P(H_1)(1 - \beta)$$

And thus,

$$FPRP = \frac{\alpha(1 - (P(H_1))}{P(D)}$$

Substituting $\mathrm{P}(D|H_0)$ for α allows us to solve for $\mathrm{P}(H_0|\mathrm{D})$ using Bayes Law:

$$\mathrm{P}(H_0|\mathrm{D}) = \mathrm{P}(\mathrm{D}_0|H_0\,)\frac{(1-\mathrm{P}(H_1))}{P(D)} = FPRP$$

Thus, the Non Replication Ratio that we set out to prespecify as a hurdle for our hypothesis tests is the FPRP and the $\mathrm{P}(H_0|\mathrm{D})$.

Ioanndis (2005) uses essentially the same framework as Wacholder et al (2014) and derives the "complement" to the False Positive Probability which he calls the "*Positive Predictive Value (PPV)"* of a test. It represents the proportion of Positive results that are True:

$$PPV = \frac{P(H_1)(1-\beta)}{P(D)}$$

Clearly, *PPV+FPPR=1* and, thus, these measures are rearrangements of the same information. It can be similarly shown that, PPV ≡ $P(H_1|D)$.

When applied to multiple hypothesis testing the FPRP and PPV measures have the practical difficulty that $P(H_1)$ is not directly observed. However, the proportion of positive results for a batch of M out of sample tests at a given α and β, $P(D)$, where the data (D) is defined as a positive test result is observable. It is possible to work back from this observation and solve for $P(H_1)$. As before, P(D) is the sum of the False Positives and the True Positives.

$$P(D) = \alpha\big(1-P(H_1)\big) + P(H_1)(1-\beta)$$

Simple algebra follows for expanding and grouping terms:

$$P(D) = P(H_1)(1-\beta-\alpha) + \alpha$$

And solving for $P(H_1)$:

$$P(H_1) = \frac{P(D)-\alpha}{(1-\beta-\alpha)}$$

This argument is not circular as P(D) is new empirical input that can be estimated from the batch of tests being conducted. What was previously but a normalisation term is now a

source of empirical information that allows the 'backing out' of an estimate of the prior. For example, a batch of 1000 tests may be observed have 15% positive results. The numerator of this expression conveys the intuition that α (e.g. 5%) of the results are expected to be false positives by chance. So in this case 15% - 5% = 10% of the positive results are likely to be valid. This simplification would hold if the denominator of this expression is one i.e. as the power of the test and the α level approach unity and zero respectively. As the power of the test weakens, the prior increases to compensate for false negatives. Often in an out of sample replication study the number of time-series observations is lower and, hence, the power of the test will be weaker *ceteris paribus*. This will lead to a lower proportion of significant results being found. The above caters for this effect.

A key precaution in estimating the prior is to avoid using the same data twice. For this reason, it is more precise to exclude the test result of the particular test being evaluated and estimate $P(D)$ out of the remaining M-1 tests. As M becomes large these figures converge on each other. When applied to a batch of M tests an estimate of P(D), $\hat{P}$, is obtained with a standard error of: $\sqrt{\frac{\hat{P}(1-\hat{P})}{M-1}}$. A refinement may to divide by the number of orthogonalised strategies instead of M-1. This provides confidence intervals for the estimate of the prior and stress testing for the robustness of p values. Again, a larger batch of tests, M, is likely to mitigate estimation error. This approach is likely to work best on large batch tests where the Bonferroni based adjustments are weakest.

Thus, for the in sample tests where the 'base case' or prior is observable it can be used directly to calculate the Non Replication Rate of the individual test. In the case of out of sample multiple hypothesis testing, the prior may be estimated in the manner described above. The methodology proposed in this paper is valid for any prior provided to it and not only that estimated by the new method outlined above.

6. Applying the Bayesian Flip

The approach adopted in this paper is to apply a threshold, such as 0.05 to the Non-Replication Rate (α*) i.e. $\mathrm{P}(H_0|\mathrm{D})$ and then to solve for the p value (α) that is consistent with this threshold. This is the adjusted hurdle level for conventional p values in this batch of tests. Substituting into Bayes Law:

$$\alpha^* = \mathrm{P}(\mathrm{D}|H_0)\frac{\mathrm{P}(H_0)}{\mathrm{P}(\mathrm{D})}$$

And rearranging with the variable to be solved for on the left hand side:

$$\mathrm{P}(\mathrm{D}|H_0) = \alpha^* \frac{\mathrm{P}(\mathrm{D})}{\mathrm{P}(H_0)}$$

This equation cannot be directly solved algebraically due to the interdependence of $\mathrm{P}(\mathrm{D})$ and $\mathrm{P}(\mathrm{D}|H_0)$ where each of these terms is needed to calculate the other. However, simple iterative procedures converge on a solution easily.

This approach is applied below to an example in the field of asset pricing: Hou, Xue and Zhang (2017) examine a large batch comprising 447 previously documented firm specific variables that have been found to significantly explain the cross section of equity returns. Using updated (but overlapping) data they find a 64% non-replication rate at 95% level of confidence. It was enquired from these authors how many of the anomalies were still found to be significant in the non-overlapping later period. While this test had not explicitly been done an upper estimate of 20% was provided (email correspondence with Prof Zhang). There exist well established methods of determining the power of a test on a case by case basis. For the purposes of this example, the simplifying assumption of a power level of 80% is applied across all tests. It is assumed that 15% of results are found to have a p value of 0.05 in the out of sample period.

Table 1. Illustrative example of adjusting the level of the p-value in multiple testing

In this example, α refers to the conventional p value used in the batch of tests which is set to 0.05 in this example. β refers to the type II (false negative) error rate. P(D)) is the proportion of relationships found to be significant in the batch of M tests. P(H) is the probability of an hypothesis being true and P(D) is the probability of observing the data. The upper panel Table 1.a estimates the prior probability of an effect being present, $P(H_1)$. This is an input into Table 1.b. which calculates the non-replicability rate, $P(H_0|D)$ at an inputted p value of 0.05. Table 1.c applies Bayes Theorem to calculate a conventional p-value consistent with desired hurdle level, α^* , for $P(H_0|D)$. Outputs are formatted in **bold**.

Table 1.a Estimating the Prior: $P(H_1) = \frac{P(D)-\alpha}{(1-\beta-\alpha)}$

α	β	P(D))	$P(H_1)$
0.05	0.2	0.15	**0.13**

Table 1.b Calculating: $\mathrm{P}(H_0|D)$

	Likelihood P(D|H)	Prior P(H)	Raw Posterior P(D|H)*P(H)	Posterior P(H|D)
H_1 (is an effect):	0.80	0.13	0.1067	**0.71**
H_0 (no effect):	0.05	0.87	0.0433	**0.29**
P(D):			0.15	

Table 1.c Doing the 'Bayesian Flip': $\mathrm{P}(\mathrm{D}|H_0) = \alpha^* \frac{\mathrm{P}(\mathrm{D})}{\mathrm{P}(H_0)}$

	Likelihood P(D|H)	Prior P(H)	Raw Posterior P(D|H)*P(H)	Posterior P(H|D)
H_1 (is an effect):	0.80	0.13	0.1067	0.95
H_0 (no effect):	**0.0065**	0.87	0.0056	0.05
P(D):			0.1123	

Table 1.a Calculates the prior $P(H_1)$ given inputs of α, β and proportion of tests found to be significant, P(D)).

Table 1.b. Calculates the $P(H_0|D)$ given inputs of α, β and $P(H_1)$. The leftmost column of table 1.b has 0.05 inputted for $P(D|H_0)$ and 0.8 for $P(D|H_1)$. Note that $P(D|H_1)$ is not (1- α) but rather the power of the test (1- β). Calculated figures in this example are: The chance of a True Positive (when there is an effect and the test is powerful enough to pick it up) is 0.107. The chance of a False Positive is 0.043. The total chance of getting any positive result, P(D), is the sum of these two: 0.15. This value is to be expected as it is the empirically observed input

that is consistent with the estimated prior. Scaling the "Raw Posterior" numbers calculated above by P(D) defines the Posterior distribution to sum to unity. In this case, the chance of a false positive given a positive test result is 0.0433/0.15 =0.29. This is $P(H_0|D)$. It also is the False Positive Report Probability (FPRP) of Wacholder et al (2004) and the Non-Replicability Rate discussed earlier. $P(H_1|D)$ corresponds to the Positive Predictive Value (PPV) of Ioanndis (2005).

Table 1.c. Calculates the $P(D|H_0)$ i.e. conventional p value, given inputs of α^*, β and $P(H_1)$. Here the rightmost column of table 1.c has 0.05 inputted for α^*, the desired hurdle for $P(H_0|D)$. By comparison with Table 1 1.b it can be seen that the more stringent p value decreases the chance of a false positive to 0.0056 and as a result drops the total proportion of positives to 0.1133. These figures are solved for such that such that false positives as a proportion of total positives is lowered to 0.05. The conventional p value calculated to be associated with this hurdle is 0.0065.

7. Conclusion

This paper argues that, primarily due to the error of implicitly selecting the uniform prior, the non-replication of 'significant' results has been well in excess of the incorrectly "*expected level*" of 5% when using a conventional p value of 0.05. In response, suggested hurdle t stats have gone up (Harvey et al, 2016; Benjamin et al, 2017). This is a practical but incomplete solution to the problem.

The Non Replication Rate is not the probability of a Type I Error but that of a false positive given that the test result is positive. This paper provides and illustratively applies an adjustment to p values for multiple hypothesis testing that is consistent with a pre-specified Non-Replication Rate. This is a more appropriate benchmark to assess if there is a 'crisis of replication'. It is analytically demonstrated that the latter is equivalent to P(Ho|D) and also the FPRP of Wacholder et all (2001). It addresses a key weaknesses common to previous attempts conducted within the classical framework of Fisher (1935): that the conventional p value provides the probability of obtaining the data given that the null hypothesis is true, P(D|Ho), and not the reverse - which is actually a test of the null hypothesis. A "*Positivist Bayesian*" approach (that uses empirical observation and logical relationships rather than subjective beliefs) to estimate the necessary prior is adopted. However, the Bayesian framework presented in this paper can also accept a prior from another source. This adjustment to p values considers both Type I and Type II errors in its formulation rather than just restricting the former as is the case in Bonferonni (1939), Holm (1979) and Benjamini and Hochberg (1995).

A few insights for effective application of the scientific method emerge from the analysis in this paper:

1. Insignificant Results Matter.

Despite the reluctance of journals to publish p>0.05 results these are important and may convey useful insight. Their lack of reporting also tends to understate the total number of hypothesis tested. An approach adopted by the journal "*Psychological Science*" is the preregistration of studies. A detailed research proposal is essentially pre-reviewed and, if approved, the findings are published no matter what the significance levels of the relationships found (Lindsay, 2015).

2. Replication Studies Matter.

Despite the perceived lack of glamour in repeating another's methodology and the reluctance of journals to publish this 'less innovative' form of research, replication studies are an essential part of the scientific process. Studies should be designed and transparently described with the intention of them being updated by others at some point in the future.

3. Try to test the same hypothesis in different ways.

This is, in essence, the inverse of testing yet another hypothesis. As stressed by Munafò and Smith (2018) "*Robust research needs many lines of evidence*". Researchers should be alert and invocative in identifying such opportunities. An example, is that in the field of behavioural finance, not only share return behaviour but also psychological evidence (such a questionnaires applied to the holders of past winner and losers) can be simultaneously be used to test for evidence of overreaction.

4. Effect Magnitudes Matter

Importantly these are missed by t statistics where the *effect size is dominated by the relative accuracy of its estimation* (see Section 3) This critique also applies to the p-value correction that is developed in this paper. An ancillary approach that can be applied that addresses this weakness is the Bayes Factor which is currently the standard method of Bayesian hypothesis testing. However calculating the Bayes Factor can be computationally challenging as the likelihood is integrated over the prior to get the posterior distribution. Early analytic work looked at conjugates – distributions whose properties are retained over integration. However, the same increase in computational power that facilitates multiple hypothesis also allows use of numerical methods such as Markov Chain Monte Carlo (where use of the Gibbs

Sampler is popular). Alternatively, and more practically, to promote the use of the Bayes Factor, Appendix A summarises a practical way that Bayes Factor can be approximated from standard regression output (based on Raftery, 1995 and Wagenmakers, 2007).

References

American Statistical Association Releases Statement of Statistical Significance and p-values, 7 March 2016; available at www.tinyurl.com/ASA-on-pvalues

Benjamini Y and Y Hochberg (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", Journal of the Royal Statistical Society, Series B, 289-300

Benjamin D et al (2018) "Redefine Statistical Significance", Nature Human Behaviour, 2, 6-10

Cochrane J (2011), American Finance Association Presidential Address: "Discount Rates", Journal of Finance, 66, 1147-1108

Fisher RA (1935, 1971), "The Design of Experiments" (9th ed.), MacMillan

Goodman S (2008), "A Dirty Dozen, Twelve P-Value Misconceptions", Seminars in Hematology, 45, 135-40

Harvey CR and Y Liu (2014a), "Multiple Testing in Economics", Working Paper, Duke University

Harvey CR and Y Liu (2014b), "Evaluating Trading Strategies', Journal of Portfolio management, 40, 108-118

Harvey CR, Liu Y and H Zhu (2016), "…and the Cross Section of Returns", Review of Financial Studies, 29(1), 5-68

Holm S (1979), "A Simple Sequentially Rejective Multiple Test Procedure", Scandinavian Journal of Statistics", 6, 65-70

Hou K, C Xue and L Zhang (2018), "Replicating Anomalies", Review of Financial Studies (forthcoming)

Ioannidis JP (2005), “Why Most Research Findings are False”, PLoS Medicine, 2, 694-701

Lindsay, D (2015). "Replication in Psychological Science". Psychological Science. 26 (12): 1827–32

Mackenzie D (2004), “Vital Statistics”, New Scientist, 36-41

Munafò, M and G Smith (2018). "Robust research needs many lines of evidence". Nature. 553 (7689): 399–401.

Masson MEJ (2011), “A Tutorial on a Practical Bayesian Alternative to Hull-Hypothesis Significance Testing”, Behavioural Research, 43, 679-690

Mclean RD and J Pontiff (2014), “Does Academic Research Destroy Stock return Predictability”, Journal of Finance

NASA Exoplanet Archive, 2019, www.exoplanetarchive.ipac.caltech.edu

Perneger T (1998), “What’s Wrong with Bonferroni Adjustments”, British Medical Journal, 316, 1236-1238

Raftery AE (1995), “Bayesian Model Selection in Social Research” in PV Marsden (ed.), Sociological Methodology, Cambridge MA, Blackwell, 111-196

Wacholder S, S Chanock, M Garcia-Closas, L El Ghormli and N Rothman (2004), “Assessing the Probability that a Positive Report is False: An Approach for Molecular Epidemiology Studies”, Journal of the National Cancer Institute, 76, No 6, 434-442

Wagenmakers EJ (2007), “A Practical Solution to the Pervasive Problem of p-values”, Psychonomic Bulletin and Review” , 14(5), 779-804

Williams R, B Blacker, M Dickinson, Dixon W, H Ferguson, A Fruchter, Giavaliso, M, Gilliand M, R Gilliland, I Heyer, R Katsanis , Z Levay, R Lucas, D McElroy, L Petro, M Postman, Adorf H and R Hook (1996), "The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry", Astronomical Journal, 112(4), 1335-1752

Appendix: A Practical Approach to Applying Bayes Factor in Hypothesis Testing

A Bayes' Factor approach can be conducted in a supplemental manner alongside traditional hypothesis testing and the p value adjustment advocated in this paper. Based on the work of Raftery (1995) and Wagenmakers (2007) the aim of this appendix is to provide a concise overview of how this can be done in practice and, thereby, expose readers to and popularise this approach.

Bayes' Factor (BF) is derived from the ratio (odds format) of two Bayes Laws (where P(D) cancels out):

$$\underbrace{\frac{\mathrm{P}(H_1|\mathrm{D})}{\mathrm{P}(H_0|\mathrm{D})}}_{\text{Posterior Odds}} = \underbrace{\frac{\mathrm{P}(\mathrm{D}|H_1)}{\mathrm{P}(\mathrm{D}|H_0)}}_{\text{Bayes Factor}} \cdot \underbrace{\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)}}_{\text{Prior Odds}}$$

BF can be interpreted as the degree to which one "changes one's mind" (i.e. goes from prior to posterior odds) based on the evidence (D). However, it is cautioned that: BF does include the prior in its construction –your mind changes more, the more you already believe! The use of Bayes Factor (BF) alongside the approach introduced in this paper is motivated by:

1. It is the ratio of two Likelihoods. Thus, the relative size of the effect matters.

2. The number of observations are usually the same for the denominator and the numerator and, thereby, its influence is cancelled out.

3. Evidence for both H_1 and H_0 considered (not just H_0 as in classical p-value testing). Bayes Factor quantifies the evidence of Data for H_1 versus H_0

4. More free parameters are automatically penalised. This can be intuitively explained as follows: A more general hypothesis is consistent with a broader set of realisations of the data and, thus, its probability distribution is relatively more 'spread-out'. A more constrained

hypothesis makes tighter forecasts. Should the data be consistent with this tighter distribution, P(D|H) will be relatively higher as its probability distribution is more concentrated. Thus, Bayes Factor has Occam's Razor built in.

Once Bayes Factor has been calculated it can be used for a comparison between two hypotheses, BF_{12}. The hurdle rates suggested by Jeffreys (1961) are presented in Table A.1 below:

Table A.1

Evidence Categories for the Bayes Factor BF_{12} (Adjusted From Jeffreys, 1961)

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for M_1
30	—	100	Very strong evidence for M_1
10	—	30	Strong evidence for M_1
3	—	10	Moderate evidence for M_1
1	—	3	Anecdotal evidence for M_1
	1		No evidence
1/3	—	1	Anecdotal evidence for M_2
1/10	—	1/3	Moderate evidence for M_2
1/30	—	1/10	Strong evidence for M_2
1/100	—	1/30	Very strong evidence for M_2
	<	1/100	Extreme evidence for M_2

Wagenmakers (2007) demonstrates a method whereby Bayes Factor can be estimated from standard regression output making its practical application relatively easy. This is done through use of the Bayesian Information Criterion (BIC):

$$BIC = -2\ln(L) + k.ln(n)$$

Where: L = the likelihood function

k = the number of explanatory variables

n = the number of observations

The BIC is a measure of model fit (like R^2) and used primarily for comparing models with the same dependent variable. Note that small values for the BIC indicate a better fit. Wagenmakers (2007) shows that, in the case of regression analysis, the BIC can be calculated as:

$$BIC = n.\ln(1 - R^2) + k.\ln(n)$$

In other words the BIC is a function of the R^2 of the model. The BF can then be approximated as a function of the difference in explanatory powers of two models: $\Delta BIC = BIC(H_0) - BIC(H_1)$:

$$BF = \frac{P(\mathrm{D}|H_1)}{P(\mathrm{D}|H_0)} \approx e^{\Delta BIC/2}$$

Finally, the resulting BF value can be interpreted as in Table A.1. An accessible and more detailed tutorial to applying Bayes Factor is provided by Masson (2011). It is hope that this exposure will increase the popularity of this approach to be used alongside that with the adjustment to p values suggested in this paper. This approach is not a complete solution to the hypothesis testing problem as a prior is not specified and the unit information prior is implicitly applied. This is an uninformative prior and in this way it has a commonality with classical statistics but with the benefits of catering for effect size, number of observations, evidence for H_1 and parsimony as listed above. Its strengths and weaknesses complement the approach suggested in this paper.