A Crisis of Confidence:

Adjusting p value Hurdles when Testing Multiple Hypotheses

PAUL VAN RENSBURG*

ABSTRACT

A crisis of non-replication has become associated with hypothesis testing that applies the conventional p value hurdle of 0.05. This paper extends the Minimum Bayes Factor approach of Edwards, Lindman and Savage (1963) to include an empirically estimated prior that can be applied in multiple hypothesis testing.

.

*van Rensburg is the Frank Robb Professor of Finance at the University of Cape Town (paul.vanrensburg@uct.ac.za) . Thanks to Prof EJ Wagenmakers, University of Amsterdam and Prof Cambell Harvey, Duke University for comments on an earlier draft, Prof Zhang, University of Ohio for additional information about the multiple test conducted in Hou, Xue and Zhang (2018); Warwick Lucas and participants at the 2019 Southern African Finance Association Conference in Cape Town (Best Presentation Award) as well as seminars in the Departments of Mathematical Statistics and Finance and Tax at the University of Cape Town. This paper has been accepted for presentation at the 2020 Frontiers of Factor Investing Conference in Lancaster, England (currently postponed due to Covid-19)

CONTEMPORARY SCIENCE is characterised by "big data" and the low cost of computing power. This allows multiple hypothesis testing to be rapid, cheap and indiscriminate. The empirical documentation of correlation has dominated over deliberation over causation. 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 approximately 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 the discipline of finance, a "zoo" of factors explaining the cross-section of equity returns has been documented (Cochrane 2010). However, replication of these findings has been poor. For example, Mclean and Pontiff (2016) consider 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 anomalies. Using updated (but overlapping) data they find a 64% non-replication rate at the 95% level of confidence.

Indeed, across almost all disciplines, non-replication has been well in excess of the 5% "expected level" when applying a p value threshold of 0.05. Among the consequences of this "replication crisis" is 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, an 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 to ascertain the correct adjustment for hurdle p values in multiple hypothesis testing. Due to the broad relevance of the topic, it is written in a manner intended to be accessible to academics across all disciplines despite an example in the field of finance being used to illustrate its recommendations. In order to appropriately adjust for multiple tests, it is also necessary to reconsider hypothesis testing at the level of the individual test. Most of the arguments presented here have been published in prior research. However, there is far from a consensus as to how to interpret and apply Bayesian statistics " ..there must be at least as many Bayesian positions as there are Bayesians" (Edwards, Lindman and Savage, 1963, 195). As a step towards a convergence of consensus there is value in putting together pieces of pre-existing arguments so as to form, what is motivated to be, the most coherent combination. Clarity of explanation and ease of implementation are also required to facilitate practical application.

The paper is organised as follows. Section I reviews prior approaches to adjusting p values in multiple hypothesis testing. The approaches are classified, almost chronologically, into Family-Wise Error Restriction Methods, False Discovery Rate Restriction methods and Bayes Factor approaches where each successive approach attempts to rectify weaknesses in the former. It becomes apparent that, in order to appropriately adjust for multiple tests, it is first necessary to reconsider hypothesis testing at the level of the individual test and here a Bayes Factor approach is motivated for. However, it requires the selection of an alternate hypothesis and a prior. The Minimum Bayes Factor approach is reviewed as a natural alternative that uses the maximum likelihood alternate. The BIC approach is also reviewed as this does not require the specification of either prior (selecting the unit information prior) or alternate. Both BF methods are procedures that are applicable at the individual as well as the multiple hypothesis testing level. Essentially, by providing a suitable prior for the MBF an adjustment for multiple hypothesis testing is suggested in this paper. Section II considers the most fundamental weakness of classical Null Hypothesis Statistical Testing (NHST):p values are the probability of observing the data given the null hypothesis is true, $P(D | H_0)$ and not $P(H_0 | D)$ -which is actually a test of the null-hypothesis. Bayes Rule is used to switch these conditional probabilities and this requires the input of a base rate or "prior", $P(H_0)$. The case for a 'positivist' Bayesian approach is made whereby priors are empirically estimated or analytically derived. In Sections III and IV, it is demonstrated how, in multiple hypothesis testing, the prior can be empirically estimated from the total number of tests conducted (M), the power of the tests (1-ß),and the portion observed to be statistically significant at the threshold α, P(D)). This allows the calculation of a hurdle p value, α*= P(D | H$_0$), that is consistent with a prespecified P(Ho | D). Section V applies this approach to a case of multiple tests in the field of asset pricing i.e. Hou, Xue and Zhang's (2017) retesting of 447 previously documented equity market anomalies. This illustrative example provides a template for the adjustment of hurdle p values in multiple testing that can be practically applied in a broad range disciplines. Section VI summarises and concludes.

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

A. Family Error Wise Rate Restriction

The Bonferroni (1936) approach aims to keep the "family wise error rate" (i.e. the probability that all M null hypotheses in the group being tested are simultaneously true) at the α level. The adjusted significance 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:

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

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 multiple hypothesis testing

This adjustment likely to be extremely punitive in situations where a large number of tests have been conducted. Applying 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:

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

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.

First, these adjustments takes no account of Type II errors (rejecting H_0 when it is true) and the resulting decrease in the power of the tests after the adjustment. Adjustments intended to be conservative for the Type I errors (not rejecting the H_0 when it is false) 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, in the Bonferroni (1936) and Holm (1979) measures 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 conflated with the fact that the base rate in the out-of-sample is test will be lower the more that the initial selection of M is based on overfitting and selection bias (see Section V). Thus, the problem of the inobservability of M can be avoided by applying these metrics to out–of-sample batch tests and this problem is also not specific to the FWER restriction methods.

More fundamentally, Pergneger (1998) has argued that the family-wise error rate is the incorrect metric to be constraining. Rubin (2017, 6) argues this point as follows "… consider a gambler who purchases 100 lottery tickets. Although this mass purchase increases the probability that the gambler will win the lottery, it does not increase the probability that any one of her lottery tickets will be the winning ticket. In the same way, a researcher who undertakes single tests of 100 different null hypotheses will have a relatively high probability on incorrectly rejecting **one** of these hypotheses, but she will not increase the probability of rejecting **each** hypothesis". Rubin proceeds to make the case that the family-wise error rate is only applicable in multiple tests of the same hypothesis.

These adjustments to p values commit the "standard error of science" i.e. using $P(D \mid H_0)$, rather than $P(H_0 \mid D)$, as its hurdle metric. 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. $P(D \mid H_0)$ under a one tailed test. However, $P(D \mid H_0)$ is not particularly useful to know in isolation. It seems to be commonly interpreted as the much more useful $P(H_0 \mid)$ which is, in fact a measure of whether the null hypothesis in true. Any adjustment to p values, be it at the individual or multiple test level, that does not account for the need to switch the above conditional probabilities is fundamentally flawed. This critique applies to all members of the FWER class of approach.

B. False Discovery Rate Restriction

Unlike the Bonferroni and Holm adjustments which aim to avoid a single false discovery, the Benjamini and Hochberg (1995) adjustment applies a False Discovery Rate (FDR) hurdle which they label "q". The FDR is the expected ratio of False Positives to All Positives conditioned on their being at least one positive result (Storey, 2003) The FDR is discussed in more detail in Section II and shown to be equivalent to $P(H_0 \mid D)$ and, in this manner, these approaches avoid the 'standard error of science' weakness that characterised the FWER methods..

Ordering observed p values from smallest to largest and defining i_{max} to be the largest index in this ranking such that:

$$p_i \leq \frac{i}{M} q$$

The Benjamini and Hochberg (1995) rule is that H_0 is rejected for all tests $i \leq i_{max}$. They provide a proof that such a threshold p value ensures a conservative FDR threshold of q. Benjamini and Yekutieli (2001) also adopt a FDR hurdle approach and demonstrate the following adjustment to be robust to dependency between the test statistics:

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

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

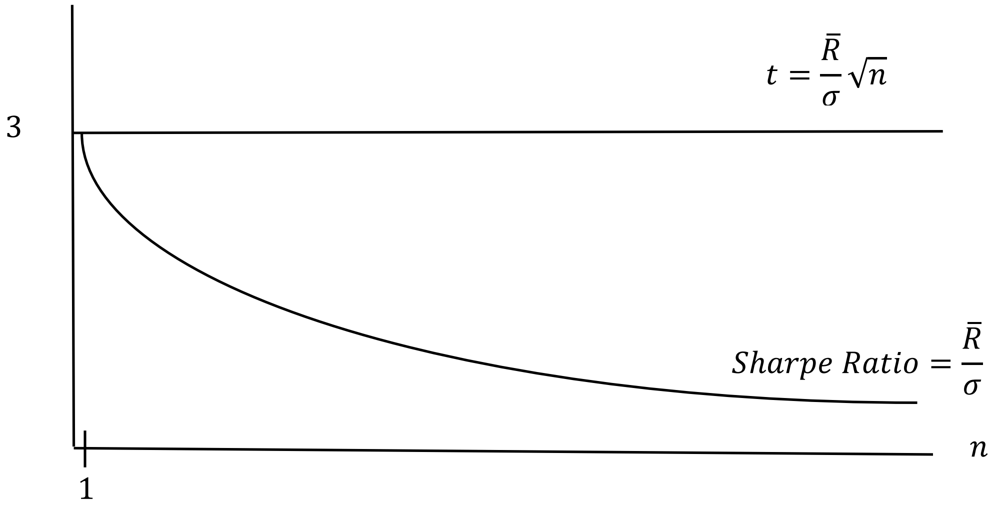
Again, each test has its own hurdle rate. Ranking from highest to lowest p values, the null hypothesis is rejected for all of those tests below and including the first case where this hurdle is met. Especially when M is large, it is generally more lenient than the two former approaches. Harvey, Liu and Zhu (2015) employ an equally weighted composite of the Bonferroni, Holm and Benjamini and Hochberg measures to suggest an adjusted t value of 3 for current asset pricing tests. However, in later work, Harvey (2017, 2) emphasizes "..that making a decision based on t>3 is not sufficient either" and a Bayesian approach is suggested (see the following section).

There are, indeed, a number of remaining problems that characterise the above FDR methods. Like the FWER restrictions methods discussed previously they also do not take account of the power of the test. The above FDR restriction are also based merely on a ranking of p values and, thereby, inherit a number of the p value's problems.

One of the most serious problems with p values is that they have the problem of misrepresenting effect size. To Illustrate, Harvey et al's (2016) 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.

$\bar{R}$ refers to mean returns in excess of the risk free rate and σ is the standard deviation of excess returns.

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 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 indicate increasing statistical significance as it does so. This is related to the bias introduced by the "stopping problem". In the field of medicine, for example, the cost of obtaining each additional observation through a clinical trial may be expensive in terms of cost, time and, possibly, the health of the subjects. Observations are collected until n is big enough to find a significant result and then the study is completed.

The 'stopping problem' is but one symptom of the effect of sampling design (see Wagenmakers 2007 for an excellent discussion of this problem from which this paper draws). Consider the lady drinking tea example mentioned in the previous section and assume she tasted a 6th cup but got it wrong obtaining 5/6 correct (x). Under the null of the binomial distribution with 6 observations the likelihood function for this result is:

$$L(p|x) = \binom{6}{1} p^5(1-p)$$

Which is 6/64 =0.0938 under the null that p=0.5. As the p value is $P(H_0 | p \geq x)$ it requires the addition of the likelihood of obtaining the more extreme value of 6/6 which is 1/64. Thus, the p value associated with the observation x=5/6 is 7/64=0.11 for a one tailed test. This would not meet the threshold level of 0.05.

However, the experiment design Sir Fischer was applying was not a 6 cup test but to continue testing until she got one wrong and then end the experiment i.e. the negative binomial distribution was applied under the null. The density function for observing s successes before f failures is:

$$L(p|n) = \binom{n-1}{f-1} p^s(1-p)^f$$

Where f=(n-s) mistakes. In this case under the null of p=0.5 it is $\binom{5}{0} 0.5^6(0.5)^1$=1/64=0.016. Again, to calculate a p value the cumulative probability of more extreme cases needs to be added:

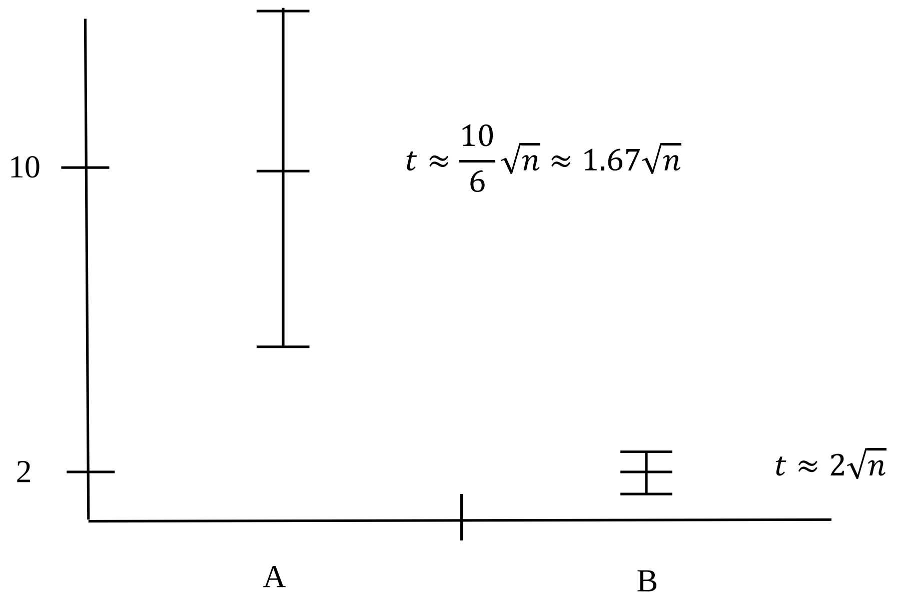
$$\sum_{n=7}^{\infty} \binom{n-1}{0} 0.5^n = 1/64$$

Note that this accounts for the effects of trials that may have but actually never happened under the negative binomial distribution. A 'total' p value of 2/64=0.032 is calculated which would meet the threshold level of 0.05. The key point of this example is that in both cases the evidence is the same (5/6 correct) and consistent with both experiment designs. However, depending on which design was intended a different p value is obtained.

This example also illustrates another problem with p values i.e. that it reflects the probability of more extreme results than that of the 'border case' that is observed. While an α threshold is by definition an inequality, this does not imply that the metric that attempts to meet it need also be. In the binomial example, the p value sums the probabilities of 5/6 and 6/6 based on the evidence of 5/6. Credit is given for two possible outcomes of the experiment when only one can happen. The likelihood of this sum is less than that of the border case, given that the border case is based on a single outcome of the experiment while the p value is based on two. The inequality specification of the p value thereby results in making observations appear more extreme than they actually are and, thereby, more likely to reject the null. This is explained in more detail for the continuous case in the following section.

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! As illustrated, while not a formal hypothesis test, confidence intervals can be a useful supplementary metric for the NHST researcher to assess effect size.

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.



C. Bayes Factor Approaches

Unlike in NHST, the use of Bayes Factor (BF) requires the selection of an alternate hypothesis, H_1, as well as the Prior Odds ratio associated with it, $P(H_0)/P(H_1)$. Computational difficulties arise when a diffuse alternative is used. For example, when H_0: μ=0, and the alternate is any value other than zero. Then the likelihood needs to be integrated over the distribution of the priors for each value other than zero to get the posterior distribution. Early work such as Edwards et al (1963) looked at conjugates – distributions whose properties are retained over integration in order to obtain a closed form solution. However, this only applies in selected cases. Also to facilitate application, Edwards et al (1963) consider under what conditions "stable estimation" is possible i.e. where a uniform prior may be used as a satisfactory approximation. Alternatively, the same increase in computational power that facilitates large scale multiple hypothesis testing also allows use of numerical methods such as Markov Chain Monte Carlo and the Gibbs Sampler. However, on a practical level, researchers would prefer to avoid these computational difficulties and this has deterred the application of Bayesian methods.

The computational trouble of integration is eliminated if a specific rather than diffuse H_1 is selected. The maximum likelihood value of H_1 is a natural candidate. The choice of H_1 in the BF specification is selected to be what is the most likely estimate based on what has been observed. In other words, if a mean value of say 5% is observed for x then H_1 : μ=5%. This is applied in the Minimum Bayes Factor (MBF) approach of Edwards, Lindman and Savage (1963) which is also recommended by Goodman (1999) and Harvey (2017) *inter alia*. When calculating Bayes Factor with H_0 in the numerator i.e. $\frac{P(D|H_0)}{P(D|H_1)}$, it will be the BF with the minimum value out of all of the possible specifications of H_1. It can be interpreted as the alternate most likely to reject H_0. If no relationship is found using the MBF, using another alternate will not alter this interpretation. Using H_0 in the numerator also facilitates comparison with NHST p values where small values indicate a rejection of H_0.

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

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)}.\ \frac{P(H_0)}{P(H_1)}$$

Clearly evidence for both H_0 and H_1 are considered. This can also be stated as the Posterior Odds is equal to the BF, times the Prior Odds. BF can be interpreted as the ratio which one should "change one's mind" from the Prior Odds. It is based on the evidence (D) and is independent of the Prior and as such can be interpreted as an 'objective' measure of comparison between two hypotheses. For example, using the hurdle rates suggested by Jeffreys (1961) a BF below 1/30 would be strong evidence for the alternate and above 30 for the null. A BF of 1/20 would roughly correspond to the NHST conventional level of 5% (albeit on a different metric) . A notable feature of the BF hurdle tables is that, in contrast to those of NHST, they are not contingent on the numbers of degrees of freedom in the test. In the case where there is only one value for $H_{1,}$ BF, is the likelihood ratio. If a uniform prior is used then the BF is also the posterior odds.

The division entailed in calculating BF allows a lot of NHST's difficulties to be "cancelled out". For example, the same sample design (e.g. binomial or reverse binomial) is used is considering evidence for both H_0 and H_1. The number of observations are also the same for the denominator and the numerator. (The number of observations does improve the standard errors of the estimated difference between the two and hence its t statistic. However, the effect that a given t statistic's p value converges downwards to that of the Z score as n increases - being equally beneficial to H_1 and H_0 - is cancelled out). This division also solves the problem of needing to convert a likelihood into a probability as it is possible to find the ratio between two lines despite them having an area of zero. Edwards et (1963) point out that the division entailed in calculating likelihood ratios is an application of the "likelihood principal" of Barnard (1947) and Fisher (1957) i.e. that multiplying a likelihood by a constant does not alter its influence. This is because the constant is both in the numerator and dominator of the likelihood function. Changing sampling designs and stopping rules would reflect in such a constant. This has the important implication that "According to the likelihood principle data stands on its own feet" (Edwards et al, 1963, 239).

Under the assumption of a Gaussian distribution of the true mean and standard error of an estimate x of the mean Edwards et al show that the MBF for H_0:μ=0 can be o be simplified as:

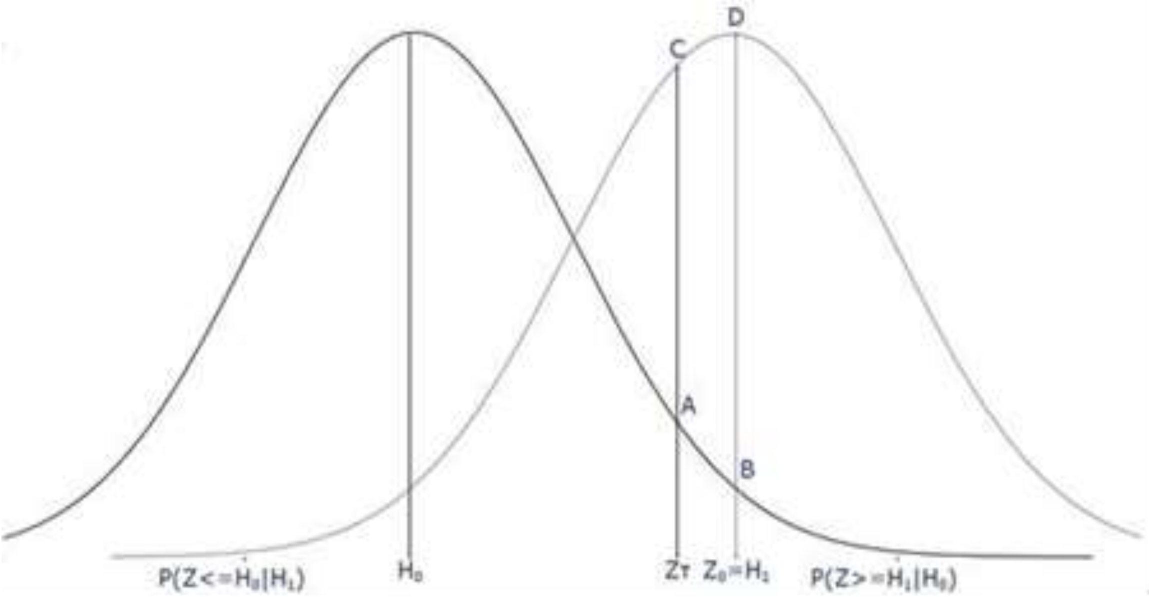
$$MBF = e^{-(\frac{Z^2}{2})}$$

Where Z ~N(0,1) is the standardised Z-score of x. Z=(x-µ)/σ where µ is the hypothesised true value of x and σ is its standard error. A comparison of the normal pdf under H$_0$: $\mu = 0$ i.e. $(P(H_0|D) = P(x|\mu = 0, \sigma)$=$\frac{1}{\sigma\ \sqrt{2\pi}} e^{-\left(\frac{(x-0)}{\sigma}\right)}/2$ with that under H$_0$: $\mu = x$ is that only the term in the exponent changes from (x-0) to (x-x=0). This allows the cancellation of the term $\frac{1}{\sigma\ \sqrt{2\pi}}$ which scales the Gaussian function such its area sums to one and this facilitates the simplification (see Appendix I in Goodman (1999)).

The conventional t statistic may be used as an approximation for Z which makes this version of the MBF easy to calculate using a familar metric the NHST toolkit. It also provides a method of 'translation of p-values into t-statistics (for a given number of degrees of freedom) and, thereby, into corresponding MBFs. Note that for a given p-value the corresponding t-statistic will increase as the number of observations decreases. Thus, a given p-value from a smaller sample will actually result in a lower (more significant) MBF.

Calculating the MBF for Z=1.96 (which is consistent with a two tailed p value of 0.05) obtains a value of 0.146. This value is the ratio of the odds of H$_0$ versus H$_1$. The BF can be converted from an odds ratio to a probability (p) by p=odds/(1+odds) which is 0.127. Thus, the MBF probability of the null is about 2.5 times as large as the p value of 0.05 in a two tailed test and double this for a single tailed test. This discrepancy arises due to the fact that the NHST p value represents the probability of the inequality µ ≥ x rather than the border value µ=x that was actually observed. This is comparison is illustrated in Figure 3 which shows the probability density functions under H$_0$ (on the left) and the maximum likelihood observed value of H$_1$ (on the right).

Figure 3 Graphical Representation of the Minimum Bayes Factor (MBF) Hypothesis testing Approach. The Gaussian probability distribution of Z-scores on the left represents the case under H_0 while that on the right the case under H_1 (where H_1 is Z_o, the observed Z-score in the MBF approach). Z_T is the threshold Z-score to be deemed statistically significant based on the selected α value in a two tailed test.



The area under the distribution assuming H_0 that is to the right of the observed value of the Z-score, Z_0, represents the p value of a single tailed test i.e. $P(Z \geq H_1 | H_0)$. In a two tailed test, the area of the same magnitude under H_1 and to the left of H_0 i.e. $P(Z \leq H_1 | H_0)$ is added to this probability, doubling its value. In textbooks this latter area is routinely depicted as the leftmost tail under H_0 but, in actual fact, it represents an area under a different distribution i.e. that of the maximum likelihood H_1. This distribution is also used in calculating the power of the test which is the area under it to the right of Z_T. Thus despite NHST's not explicitly specifying an alternate, the maximum likelihood alternate plays a key role in the calculation of both the power and two tailed p value metrics. As a MBF approach explicitly selects an H_1, it knows in which direction the effect tested for is and, thus, only a single tailed test would make sense (Edwards et al 1963).

In Figure 3, The length of the vertical line B represents the likelihood of the observed value, Z_0, under H_0 while D is its likelihood under H_1. The probability of any particular event such as Z_0 occurring is vanishingly small in the continuous case. Thus, NHST has used the integration of all values more extreme than Z_0 as an approximation to solve this problem. However, it is possible to calculate and interpret relative likelihoods when an alternate has been specified. The ratio B/D is the likelihood ratio of the observed value under H_0 relative to the maximum likelihood H_1. A and C represent the analogous likelihood values at the p value threshold. The likelihood ratio at Z_T is similarly represented by the ratio A/C. However, when the inequality $P(Z \geq Z_T | H_0) = \alpha$ is (incorrectly) used in a one-tailed test the likelihood becomes this area relative to the area $P(Z \geq Z_T | H_1) = 1 - \beta$. Thereby, a relatively elegant expression for BF = $\alpha/(1-\beta)$ i.e. alpha divided by the power of the test, is obtained at this threshold. Clearly, the MBF approach takes the power of the test into account unlike the FWER and FDR restriction methods reviewed earlier. However, this measure is imprecise in its use of equalities and this ratio of areas will be less than that of the correct ratio A/C, making the test result seem less likely under the H_0 than it really is.

One of the possible weaknesses of the MBF approach is the need to specify a prior. A uniform prior, however, is a useful 'base case' to consider. An alternate BF approach that has minimal reliance on an empirically observed prior is that of Raftery, (1995) and Wagenmakers (2007). They show how BF can be approximated from standard regression analysis output, making it easy to practically apply. This is done through use of the Schwarz (1978) or "Bayesian Information Criterion" (BIC) of the model estimated:

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

Where: L = the maximised value of the likelihood function

k = the number of parameters estimated

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 smaller 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, $\Delta BIC = BIC(H_1) - BIC(H_0)$:

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

In this case, the model associated with H_1 may, for example, have an additional explanatory variables(s) than that associated with H_0. Indeed, H_0 may just be a regression model with only an intercept term reflecting only the base rate mean of the dependent variable. The resulting estimated BF value can be interpreted as discussed earlier in this section. An accessible and more detailed tutorial to applying the BIC approach is provided by Masson (2011).

Both a strength and weakness of this approach is that it does not require the specification of a prior as the unit information prior is implicitly applied. This prior essentially uses the observed mean and standard deviation in the dependent variable and accords it the information content of a single observation. Rafterty (1998, 1) argues that "*Clearly a prior that represents the available information should be used, although the unit information prior often seems reasonable in the absence of strong prior information*". Raftery (1998) further argues that due to its 'spread-out' nature the unit information prior will be conservative towards rejecting the null and can used as a "*baseline reference analysis*".

Despite the appeal of the BIC approach for individual hypothesis testing, due to its ease in calculation and not allowing the explicit selection of a prior, it is not clear how this approach can be adapted for multiple hypothesis testing. In contrast, the MBF approach can be facilitated by estimating an empirical prior that can be used in multiple hypothesis testing. Section II below reviews the use of Bayes Law to switch conditional probabilities from $P(D \mid H_0)$ to $P(H_0 \mid D)$. Section IV decomposes the proportions of expected test results under the permutations of True/False and Positive/Negative. This decomposition is used to obtain an estimate of the prior in Section V. The approach is illustratively applied in Section VI.

II. The Bayesian Flip

The following analysis make two important simplifying assumptions. First it is assumed that the test is dichotomous in outcome i.e. it provides a binary True or False result. This avoids the problem of having a diffuse alternate and the need to choose a specific for H_1. The alternate is also, by default, the maximum likelihood H_1. Second, it is assumed that that the 'precise' value for the p value is equivalent to the inequality value of the p value i.e $p=0.05$ is equivalent to $p\leq0.05$. In other words, when $p\leq0.05$ is obtained, an event with a 5% probability under H_0 has happened. The avoids the complexities entailed due to the fact that the p value reflects the probability of more extreme results than that of the 'border case' that is observed (see Section I.B).

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 into the analysis. Incorporation of the prior has been the most controversial feature of applying Bayes Law. It has been criticised as a means for subjective biases to enter the analysis. This need not be the case. The "Positivist Bayesian" approach suggested in this paper can use (i) empirical analysis i.e. it subsumes an "empirical" Bayes approach, 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. With some simplification, 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 approximately 5:1 (the $\mathrm{P}(\mathrm{H}_0 | \mathrm{D})$ is 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. Given these odds, even for an individual in-sample test, it is no wonder that there will be pervasive non-replication of this result. The p value is not a measure of replicability nor a test of the null hypothesis.

Importantly, estimating the base rate is just as valid an empirical analysis as the conventional estimation of $\mathrm{P}(\mathrm{D}|H_0)$. The prior is not the probability before getting *any* data. It is the prior before getting the *additional* data used by the test concerned. In this sense, the term "base rate" maybe preferred to that of "prior". An element of what is typically referred to as "descriptive statistics" (e.g. the empirically observed proportion of people with the disease in our example) is actually an essential input into the inference process. Furthermore, as more data accumulates and follow up tests are conducted, the initial base rate becomes less relevant. Bayes Theorem provides a mechanism to rationally update the prior based on new evidence. The BF can be multiplicatively updated for N independent tests:

$$BF_N = \prod_{i=1}^{N} BF_i$$

In our disease example, given the initial positive result, a second independent test for the disease would commence from a 17% rather than 1% base rate. In this case, if a second positive result is found, the $\mathrm{P}(\mathrm{H}_0 | \mathrm{D})$ 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 $\mathrm{P}(H_1) = 50\%$, the incorrect inference of 100% would be obtained. This uniformed prior ignores the information provided by the assumption that a standard pack of cards is used.

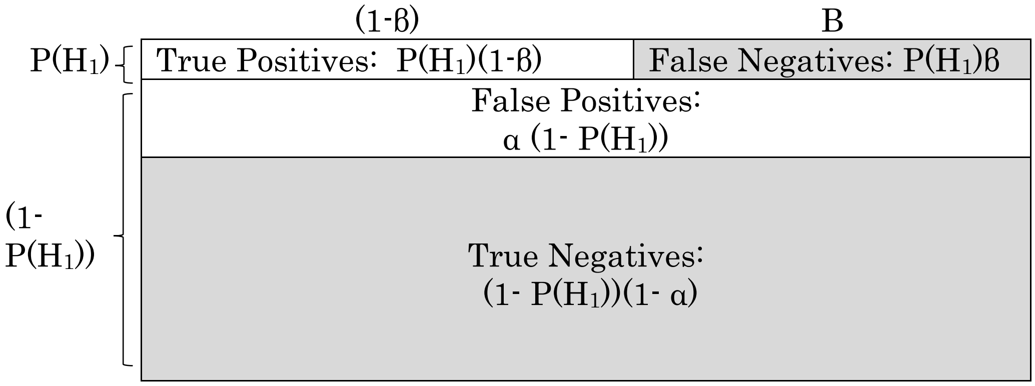
Examples of analytically deriving a prior also occur when considering test design. For example, in 'extreme performer' research – trying to identify those shares with the 10% best returns over the following year, by methodological construction, in the absence of additional information the prior can be inferred to be 10%.

IV. Adjusting the p value and Supplementary Metrics

Figure 4 considers the breakdown of research findings against true relationships. Here for simplicity, a dichotomous variable (having the disease or not i.e. True/False) is matched with by another dichotomous metric (a two-tailed test result that is found to have a p value of α or lower i.e. Positive/Negative). In the case of an single in-sample test, such as the probability of disease example used earlier, $P(H_1)$ would represent the base rate probability of having the disease without knowing the test result. Given the α level and power of the test (β) this allows calculation of the expected number of cases for the permutations of True/False and Positive/Negative.

Figure 4. Test Results and True Relationships

The total area of the rectangle below represents the full probability space of possible results and sums to unity. Where P(H$_1$) represents the base rate, α refers to the level of significance used and (1-ß) to the power of the test. The unshaded area represents the portion of positives identified by the test and the shaded area the negatives.


(1-ß)
B
P(H$_1$)
True Positives: P(H$_1$)(1-ß)
False Negatives: P(H$_1$)ß
False Positives: α (1- P(H$_1$))
(1- P(H$_1$))
True Negatives: (1- P(H$_1$))(1- α)

The top row of Figure 4 represents the proportion of cases where the null hypothesis is false, P(H$_1$). Some will be associated with 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-ß) where ß 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 the null hypothesis is true. 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.

Defining $P(D)$, such that the data (D) is a Positive test result and substituting α for $\mathrm{P}(D|H_0)$ allows us to solve for $\mathrm{P}(H_0|\mathrm{D})$ using Bayes Law:

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

From the breakdown in Figure 4, P(D) is the sum of the False Positives and the True Positives:

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

Given that $P(H_0) = 1 - P(H_1)$, this allows the following substitution:

$$\mathrm{P}(H_0|\mathrm{D}) = \frac{\alpha(1-(P(H_1))}{\alpha\big(1-P(H_1)\big)+P(H_1)(1-\beta)}$$

The influence of the power of the tests (1-ß) can be explicitly seen here as influencing the True Positive $P(H_1)(1-\beta)$ portion of the denominator. A weaker power leads to less True Positives and a higher P(H0 | D). It is also evident by substitution that P(H0 | D) can be more simply expressed as the ratio of false positives to all positives:

$$\mathrm{P}(H_0|\mathrm{D}) = \frac{False\ Positives}{All\ Positives} = \mathrm{FDR}$$

Unlike the p value –this measure is scaled by the probability of getting a positive result. As mentioned in Section I.B, in the field of finance this ratio has been termed the "False Discovery Rate" (FDR) (see Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001 Harvey Liu and Zhu, 2016;, Harvey, 2017 inter alia). Wacholder et al (2004) independently derive a "False Positive Report Probability" (FPRP) that is equivalent to this expression. In the context of the "replicability crisis" the FDR also has the relevant interpretation as a "Non Replication Rate" i.e. the probability that there is no True relationship when a Positive test result is observed. The most important feature to recognise about this metric, however named, is that is equivalent to $\mathrm{P}(\mathrm{H}_0|\mathrm{D})$ – by definition the proper hypothesis testing metric.

Ioanndis (2005) derives the "complement" to the FDR/FPRP 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 via Bayes Theorem that, PPV ≡ $P(H_1|D)$ ≡ 1- $P(H_0|D)$.

The approach adopted in this study (is that a threshold is applied to $\mathrm{P}(H_0|\mathrm{D})$ and then the p value that is consistent with this threshold (α^*) is solved for. Typically, we would want to keep $\mathrm{P}(\mathrm{H}_0|\mathrm{D})$ low, analogous to the 0.05 p value level used in NHST. Substituting 0.05 as the hurdle $\mathrm{P}(H_0|\mathrm{D})$ into Bayes Law and rearranging with the variable to be solved for on the left hand side:

$$\alpha^* = 0.05\frac{\mathrm{P(D)}}{\mathrm{P}(H_0)}$$

As can be seen, lower the base rate i.e. a larger value for $\mathrm{P}(H_0)$, the stricter the threshold α^* becomes. This equation cannot be directly solved algebraically due to the interdependence of $\mathrm{P(D)}$ 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 allows us to solve for the p value (α^*) such that $\mathrm{P}(H_0|\mathrm{D}) = 0.05$.

There also exist supplementary metrics which, although not tests of the null-hypothesis, provide useful insight beyond that of the FDR. The "Miss Ratio" (RMISS) is the rate at which test will miss finding a true relationship (Genovese and Wasserman, 2002; Sarkar, 2004; Harvey and Liu, 2019) :

$$RMISS = \frac{False\ Negatives}{All\ Negatives}$$

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

Analogous to the FDR, RMISS is set to zero if there are no Negative results. As can be seen the lower the hurdle α and the base rate $P(H_1)$, the higher the proportion of True Negatives and the lower the Miss Ratio. Given that the proportion of True Negatives is a positive number, an increase in the power of the test $(1 - \beta)$ will decrease the Miss Ratio. It is probably good practice to report the power of the test(s) concerned when interpreting the RMISS.

Harvey and Liu (2019) introduce a useful metric RRATIO, the "Ratio of False Discoveries over Misses":

$$RRATIO = \frac{False\ Positives}{False\ Negatives}$$

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

This metric allows insight into the trade-off between Type I and Type II errors. for For example, in the identification of a dread disease that would benefit from being treated early, a Type II error (missing that something's there) is likely to be much more costly that of a Type I (a stressful false alarm). In contrast, making a false criminal conviction (Type I error) is generally deemed more costly to society than letting a guilty person get way (Type II error). Hence the criminal threshold "beyond a reasonable doubt" rather than the civil threshold of the "balance of probabilities". Harvey and Liu (2019) suggest solving for the α threshold level such that a desired RRATIO is achieved. In a similar manner, Storey (2003) analytically defines an investor's loss function as a weighted average of RDR and RMISS and solves for a significance threshold that minimises its value. If, for example, there is an excess of False Positives the alpha level is dropped. There exists an interaction effect whereby β will increase as a result and both will contribute to the RRATIO decreasing.

V. An Empirical Prior for Multiple Hypothesis Testing

At the level of multiple hypothesis testing there arise new difficulties that are more familiarly associated with the replication crisis. There are the concerns regarding the false positives that we anticipate are going to arise for a given hurdle p value within the batch of tests conducted. Due to selection bias, overfitting and the correlations between tests, the true number of independent tests that are represented in the batch is unobservable. As a result, when applied to multiple out-of-sample hypothesis testing, applying the FDR has the practical difficulty that $P(H_1)$ is not directly observable. 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.

At the level of multiple hypothesis testing it is possible to use the same framework as before where the probability space (area of Figure 4) constitutes M (individually properly specified) tests rather than N observations (as was the case for the individual in-sample tests). $P(H_1)$ is defined here as the probability of a True relationship being present across the batch of tests being conducted. As argued in Section II, the number of *out-of-sample* tests (M) is observable. The proportion of positive results for a batch of out-of-sample tests at a given α and β, $P(D)$ is, also observable. It is possible to work back from this latter observation and solve for an estimate of $P(H_1)$. 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, in this case, P(D) is new empirical input that is gleaned from the batch of tests being conducted. Unlike in the individual test case, it is now possible to observe the results of all of the other tests in the batch and see how many were found to be significant. 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% of the results being Positive. With some simplification, 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 α level approach unity and zero respectively. However, as stated, it has the weakness of not considering the power of the tests concerned. As the power of the tests weaken, the estimated base rate 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. Thus, the power term is in the denominator of this metric and as it weakens the estimate for P(H_1) will increase to cater for this this effect.

A key precaution in estimating the base rate 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, of course converge. When applied to a batch of M tests an estimate of P(D), $\hat{P}$, is obtained with a standard error of: $\sqrt{\frac{\hat{P}(1-\hat{P})}{M-1}}$. The estimate is likely to be robust to correlation between p values but its standard error is not. A refinement may to replace M with the number of orthogonalised strategies when estimating the standard error. 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.

To compare with prior work, unlike the Bonferroni (1936) and Holm (1979), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) approaches which only take α and M into account, this adjustment also takes the power of the test and proportion of positive results within the batch as inputs to adjust the hurdle p value. The analytic approach taken here differs from the bootstrapping approach applied in Harvey and Liu (2019) and Fama and French (2010). The analytic solution offers computational ease without further econometrics. It is, however, more sensitive to the violation of the somewhat restrictive assumptions on which it is based. The Positivist Bayesian approach described in Section II would disagree with Harvey and Liu (2019,8) "… the choice of (prior) is inherently subjective". A Positivist Bayesian approach similarly differs from the "personal probability" approach of Edwards et al (1963, 197) which they illustrate as follows "For you now, the probability P(A) of an event A is the price you would be willing to pay in exchange for a dollar to be paid to you in case A is true. Thus, rain tomorrow has a probability 1/3 for you if you would pay just $.33 now for in exchange for $1 payable to you in the event of rain tomorrow". It is also necessary for these probabilities to be consistent such that you cannot be trapped into accepting a combination of bets that assures a loss. In order to ensure consistency it is necessary to be willing to take both sides of each bet otherwise risk aversion could lead a universal lowering of probabilities i.e. you would require a payoff rate higher than the perceived probability of payment to take on risk. However, subjective opinions are nevertheless influenced by well documented behavioural biases such as underestimation of uncertainty. There may be 'other interests' such as owning an umbrella factory that may influence your required payoffs. The vaguarities of personal opinion aside, an important objective of the scientific method is to avoid the situation where two people may have different subjective beliefs and one cannot be considered more correct than the other. In this case, a Bayesian approach based on an empirically estimated prior "lets the data speak" in such a manner that two researchers presented with the same data would get to the same conclusion. Facilitating such communication between researchers and, thereby allowing for the possibility of 'replication' is the primary motivation for the Positivist rather than Subjectivist view. Aside from this difference, this paper broadly agrees with and follows Edwards et al (1963). Harvey (2017) and also the very well written Goodman (1999a, 1999b). As a result, it is in the MBF context that multiple hypothesis testing is considered in this paper and an adjustment to p value hurdles is made by empirically estimating a prior that it can use.

This adjustment for multiple testing is illustratively applied using 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). For the purposes of this example, the simplifying assumption of a power level of 80% is applied across all tests (as the power of a test is readily measurable this could be adjusted on a case by case basis by the researcher who has access to the data). It is assumed that 15% of results are found to have a p value of 0.05 in the out of sample period.

Table I

Illustrative example of adjusting the level of the p value in multiple testing

In this example, α refers to the hurdle p value used in the batch of tests which is set to 0.05. ß refers to the type II (False Negative) error rate. P(D)) is the proportion of relationships found to be significant (the Data being defined as Positive test result) in the batch of M tests. P(H) is the probability of an hypothesis being true. The upper panel Table 1.a estimates the prior (base rate) probability of an effect being present, $P(H_1)$. This is an input into Table 1.b. which calculates the FDR or $P(H_0 | D)$ at the p value of 0.05. Essentially, reversing the process, Table 1.c applies Bayes Theorem to calculate an adjusted hurdle p value consistent with $P(H_0 | D) = 0.05$. Outputs are formatted in **bold**.

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

α	ß	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': $\alpha^* = \mathrm{P}(\mathrm{D}|H_0) = 0.05\frac{\mathrm{P(D)}}{\mathrm{P}(H_0)}$

	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 Estimates the prior $P(H_1)$ given inputs of α, ß and the proportion of tests found to be significant in an out of sample multiple test batch, P(D)).

Table 1.b. Calculates $P(H_0 | D)$ 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 0.1067. The chance of a False Positive is 0.0433. 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 dividing by P(D) results in the Posterior distribution summing 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 $P(H_0 | D)$ or the FDR. From this metric it can be said that the null hypothesis can, in this case, be rejected at the 71% level of confidence (rather than the 95% level routinely incorrectly inferred from the p value). As pointed out in Section III, $P(H_1 | D)$ (=0.71 in our example) corresponds to the Positive Predictive Value (PPV) of Ioanndis (2005). The BF (which is also the likelihood ratio in this case) can be calculated as $P(D | H_0)$/ $P(D | H_1)$=0.05/0.80=0.065. Converting from odds format to probability format we obtain 0.065/(1+0.65)=0.0588. In this case, is higher than the p value of 0.05. This is due the fact that stated in probability format the MBF is α/(1-β+α) and this is >α if β>α,, as it is in this case.

Table 1.c. Calculates the $P(D | H_0)$ i.e. adjusted p value, given inputs of target $P(H_0 | D)$, β and $P(H_1)$. Here the rightmost column of table 1.c has 0.05 inputted for the desired hurdle for the $P(H_0 | D)$. 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 i.e. $P(H_0 | D)$ is lowered to 0.05. The p value calculated to be associated with this hurdle is 0.0065. In this example, this is the hurdle that should be used for hypothesis testing in an out of sample batch test.

Table 2 calculates the supplementary metric of the Miss Ratio. Here the data, D, is defined as a negative test result. Here the proportion of False Negatives is 0.0267 while True Negatives are 0.8233. The total proportion of Negatives is 0.85. Thus about 3% of Negatives are likely to be false and 97% correct. The RRATIO of False Positives to False Negatives is 0.0433 to 0.0267 which is 1.625. Thus, the test is more likely to find false effects that miss true ones. Depending on their relative cost this may result in an adjustment to the p value. If we assume that the cost is even i.e. the desired RRATIO=1 then the hurdle p value would need to decrease to 0.036 (under the conservative but incorrect assumption that the power of the test remains constant). As the power of the test weakens the drop in the alpha value will be less than this. In contrast to the FDR, the Miss Ratio rewards lower powered tests with easier threshold levels. If a p value threshold of 0.0065 consistent with a FDR of 5% is applied then the RMISS remains relatively unchanged. However, the number of false positives drops dramatically resulting an RRATIO of 0.18. In cases where the cost of relative cost missing an effect is assessed as being larger than this, it may motivate a relaxing of the p value hurdle for practical application (Harvey and Liu, 2019). Clearly, the supplementary metrics of RMISS, RRATIO and an awareness of test power, provide a much richer 'dashboard' is provided to the researcher than the myopia of overusing p values.

Table II

Calculating the Miss Ratio (RMISS)

This table is analogous to Table I with the difference that P(D)) is the proportion of relationships found to be *insignificant* (the Data being defined as Negative test result). As before, α refers to the hurdle p value used in the batch of tests which is set to 0.05. ß refers to the type II (False Negative) error rate. Outputs are formatted in **bold**.

Calculating: $P(H_0|D)$ where D = Negative Test Result

	Likelihood P(D∣H)	Prior P(H)	Raw Posterior P(D∣H)*P(H)	Posterior P(H∣D)
H_1 (is an effect):	0.20	0.13	0.0267	**0.03**
H_0 (no effect):	0.95	0.87	0.8233	**0.97**
P(D):			0.85	

VI. Conclusion

Reviewing the problems in prior approaches to adjusting the p value hurdle for multiple hypothesis testing has necessitated a reconsideration of the foundations of individual hypothesis testing methods. The most serious problem with FWER restriction methods such as that of Bonferonni (1936) and Holm (1978) is that do not cater for the fact on the individual test level the p value is not a test of H_0. This is corrected for in the attempts at FDR restriction such as Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). However, these approaches being based on a ranking of p values inherit a number of its problems that are addressed by the 'division' entailed in a Bayes Factor approach where a specific alternate is required. Following Edwards, Lindman and Savage (1963), Goodman

(1999) and Harvey (2017) a MBF approach is supported where the observed mean value is selected as the maximum likelihood alternate. However, this approach also requires the specification of a prior. This paper advocates a “fully blown” but “Positivist” Bayesian approach which uses empirical observation and logical relationships, rather than subjective beliefs, to estimate the base rate or prior. Replication of conclusions is thereby facilitated, which is clearly germane in the context of the ‘replication crisis’. For the individual hypothesis tests where the prior is observable or can be analytically derived it can be used directly to estimate the $P(H_0)$ of the individual test. In such a vein, the appealing and easy to estimate BIC approach of Raftery (1995) and Wagenmakers (2007) uses the mean observed value of the explanatory variable as a prior but only accords it the explanatory power of one observation. While the BIC can be recommended as a the individual test case, the MBF approach has greater flexibility in setting a prior and this is useful in the multiple hypothesis testing context. Here, a method of empirically estimating the prior is presented and applied in an illustrative example. As demonstrated, this approach is easy to apply, has a closed form solution and makes an incremental contribution by demonstrating how to adjust p value hurdles for multiple hypothesis testing.

REFERENCES

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

Barras L, O Scaillet and R Wermers (2010), False Discoveries in Mutual Fund Performance: Estimating Luck in Estimated Alpha, Journal of Finance, 65, 179-216

Bayarri MJ and JO Berger (1998), "Quantifying Surprise in the Data and Model Verification", in Bernardo JM, JO Berger, AP Dawid and AFM Smith (ed.), Bayesian Statistics 6, Oxford University Press, Oxford

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

Benjamini Y and D Yekutieli (2001), "The Control of the False Discovery Rate In Multiple Testing Under Dependency", The Annals of Statistics", 29, 4, 1165-118

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

Debondt WFM and R Thaler (1985), "Does the Stick Market Overreact?", Journal of Finance, 40(3), 793-805

Edwards W, H Lindman and LJ Savage (1963)," Bayesian Statistical Inference for Psychological Research", Psychological Review, 70, 193-242

Fama EF and KR French (2010), "Luck versus Skill in the Cross-Section of Mutual Fund Returns", Journal of Finance, 65, 915-1947

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

Genovese C and L Wasserman (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure", J.R. Stat. Soc. Ser. B Stat. Methodol. 64, 499-517

Goodman SN (1993), "P Values, Hypotheses Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate" American Journal of Epidemiology, 137 (5), 485-496

Goodman SN (1999a), "Towards Evidence-Based Medial Statistics" 1: The P Value Fallacy, Ann Intern. Med., 130, 995-1004

Goodman SN (1999b), "Towards Evidence-Based Medial Statistics" 2: The Bayes Factor, Ann Intern. Med., 130, 1005-1013

Goodman SN (2008), "A Dirty Dozen, Twelve P value Misconceptions", Seminars in Haematology, 45, 135-40

Harvey CR (2017), American Finance Association Presidential Address: "The Scientific Outlook in Financial Economic", Journal of Finance, 72, 1399-1440, SSRN abstract=2893930

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 and Y Liu (2020), "False and Missed Discoveries in Financial Economics", Journal of Finance (forthcoming).

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

Raftery AE (1998), “Bayes Factors and BIC: Comment on Weaklim”, Technical report no. 347, Department of Statistics, University of Washington, Seattle, WA

Rubin, M (2017), “Do p values lose their meaning in exploratory analysis? It depends on how you define the familywise error rate”, Review of General Psychology, 21, 269-275

Sarkar SK (2004), “FDR-Controlling Stepwise Procedures and their False Negative Rates”, Journal of Statistical Planning and Inference, 125, 119-137

Schwarz GE (1978), “Estimating the Dimension of a Model”, Annals of Statistics, 6, No 2, 461-464

Storey JD (2003), “A Direct Approach to False Discovery Rates”, J.R. Sata. Soc. Ser. B Stat. Methodol., 64, 479-498

Storey JD (2002), “The Positive False Discovery Rate: A Bayesian Interpretation and the q-value”, Annals of Satatistics, 31, 2013-2035

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