Significance Testing in Empirical Finance: A Critical Review and Assessment

Jae H. Kim*
Department of Finance
La Trobe University, Bundoora, VIC 3086
Australia

Philip Inyeob Ji
Department of Economics
Dongguk University, Seoul
Republic of Korea

Abstract

This paper presents a critical review on the practice of significance testing in modern finance research. From a survey of recently published articles in four top-tier finance journals, we find that the conventional significance levels are exclusively used with little consideration of the key factors such as the sample size, power of the test, and expected losses. It is found that statistical significance of many surveyed papers becomes questionable, if Bayesian method or revised standards of evidence were instead used. We also observe strong evidence of publication bias in favour of statistical significance. We propose that substantial changes be made to the current practice of significance testing in finance research, in order to improve research credibility and integrity. To this end, we discuss alternative ways of choosing the optimal level of significance, in consideration of the key factors relevant to finance research, with Monte Carlo evidence and empirical application.

Keywords: Level of significance; Lindley paradox, Massive sample size; Meehl's conjecture; Publication bias; Spurious statistical significance.

Draft: March 2014

* Corresponding Author. Tel: +613 94796616; Email address: J.Kim@latrobe.edu.au

We would like to thank Geoff Cumming, Steve Easton, Tom Engsted, Edward Podolski-Boczar, Abul Shamsuddin, Mervyn Silvapaulle, Tom Stanley, and Xiangkang Yin for their constructive comments on an earlier version of the paper.

1. Introduction

Significance testing is widely and extensively conducted in finance research. It is an essential tool to establish statistical evidence of an association or relationship among the financial variables of interest. In academic research, significance testing plays an important role in testing empirical validity of financial theories or hypotheses, both new and old. In business and government, outcomes of significance testing aid the key stakeholders in their corporate and policy decisions. Hence, the way in which significance testing is conducted has huge bearings on knowledge advancement and social welfare. For example, there has been a controversy that the recent global financial crisis is partly attributable to rating agencies which incorrectly over-rated sub-prime mortgages, based on deficient mathematical models or faulty statistical models with inadequate historical data (Donnelly and Embrechts, 2010)[1]. Since significance testing is a building block for statistical or mathematical models, it should be carefully conducted bearing in mind the potential consequences of making incorrect decisions.

For years, abuse and misuse of significance testing have been a subject of criticism in many disciplines of science (see Morrison and Henkel; 1970; Ziliak and McCloskey; 2008, p.57). Recently, in medical research, Ioannidis (2005) has expressed concerns that most current published findings are false, partly because researchers merely chase statistical significance. From psychology, Cumming (2013) calls for substantial changes in the way that statistical research and significance testing are being conducted. Keuzenkamp and Magnus (1995) and McCloskey and Ziliak (1996) provide the critical reviews on the practice of significance testing in applied economics. Their main criticisms include (i) arbitrary choice of the level of significance; (ii) little consideration of the power (or Type II error) of test; (iii) confusion

[1] See also p.32 of Financial Stability Forum Report available at http://www.financialstabilityboard.org/publications/r_0804.pdf.

between statistical and substantive importance (economic significance); and (iv) the practice of "sign econometrics" and "asterisk econometrics" with little attention paid to effect size. Despite these continuing criticisms, it appears that the practice of significance testing has not improved noticeably. For example, in their updated survey of the articles published in the *American Economic Review,* Ziliak and McCloskey (2004) report little improvement since the publication of their earlier survey in 1996, although Hoover and Sigler (2008) and Engsted (2009) strongly criticise and refute this claim. Recently, Ioannidis and Doucouliagos (2013) question the credibility of empirical research in economics and business studies, and discuss a range of key parameters that affect the credibility including sample size, effect size, and replicability.

In empirical finance to date, the practice of significance testing has not been given proper attention that it deserves[2]. The purpose of this paper is to fill this gap. We conduct a survey of recently published papers in top-tier finance journals to shed light on the current practice of significance testing in finance. We have identified the following salient features. First, the use of large or massive sample size is prevalent. As Neal (1987) points out, there is danger that statistical significance can be over-stated or even spurious in this case. Second, the conventional levels of significance (1%, 5%, and 10%) are exclusively used, with little considerations given to the key factors such as sample size, power of the test, or expected losses from incorrect decisions. There is no reason to believe that these conventional levels are optimal or should be preferred, especially when the sample size is massive. Third, we find clear evidence of publication bias in favour of statistical significance, with the proportion of the studies with statistical significance unreasonably high.

[2] Petersen (2009) report a survey on the use of robust standard error estimators in finance; and Engsted (2009) provides a selected review of the studies in empirical finance in relation to discussion of economic significance, in response to the criticisms made by McCloskey and Ziliak.

In finance, noting the effect of large sample size on significance testing, Connolly (1989, 1991) proposes the use of the Bayesian method of hypothesis testing previously developed by Leamer (1978) and Zellner and Siow (1979). However, it has been largely ignored in modern finance research. Specifically, Leamer (1978) recommends that the level of significance be adjusted as a decreasing function of sample size, which is not generally followed in finance (and also in other areas). Recently, from a survey of the studies published in psychological journals, Johnson (2013) argues that the level of significance be set at 0.001 or 0.005 as a revised standard for evidence by reconciling the Bayesian and classical methods. In this paper, we find that the outcomes of significance testing in many studies included in our survey are reversed if the Bayesian alternatives were instead used or a much lower level of significance than the conventional one was adopted as revised standard for evidence.

We also discuss the methods of choosing the optimal level of significance with explicit consideration of sample size, power (or Type II error) or expected losses from incorrect decisions, following Leamer (1978). We provide Monte Carlo evidence and an empirical application that the optimal choice of level of significance based on Leamer's (1978) line of enlightened judgement can provide substantially different inferential outcomes from those based on the conventional levels of significance. With a large or massive sample size, we recommend the use of the Bayesian method of Zellner and Siow (1979) or level of significance be set at a much lower level following Johnston (2013). When the small sample size is small and the power of the test is low, the level of significance should be set at a level much higher than the conventional ones, for a sensible balance between Type I and II errors. Our survey also reveals that replication is not extensively conducted in finance research and that the data and computational resources are not actively shared among finance academics. We also find that, while the use of robust standard error estimators is widespread in finance

research, little effort is being made to identify the error structure and conduct more efficient estimation of effect size for improved performance of significance testing.

We conclude that finance researchers should rethink the way they conduct significance testing in their statistical research. In particular, mindless use of the conventional level of significance should be avoided, especially when the sample size is large or massive. As mentioned earlier, grave concerns have been raised about the credibility in published studies in many fields of science. Coupled with the problems which occur when large sample size is combined with fixed levels of significance, the publication bias and limited replicability of finance research suggest that the same concern may be raised in the field of finance. Ioannidis and Doucouliagos (2013) provide a detailed discussion as to how research credibility can be improved; while Ellis (2010) and Cumming (2013) propose the guidelines for improved statistical research, which are highly suggestive to empirical finance. The rest of the paper is organized as follows. In the next section, we discuss the background of statistical issues. Section 3 provides the details and a summary of our survey. In Section 4, we present the analysis of our survey results with Monte Carlo evidence and empirical example. Section 5 presents further discussion, and Section 6 concludes the paper.

2. Background

In this section, we provide the background of significance testing in the context of finance research, paying attention to the level of significance, sample size, power of the test, and expected losses. We also introduce the Bayesian methods of hypothesis testing of Zellner and Siow (1979), and the methods of choosing the level of significance based on Leamer's (1978) line of enlightened judgement.

Level of Significance and Sample Size

The level of significance (α) is the probability of Type I error of a hypothesis test, which is the probability of rejecting the true null hypothesis. It is usual convention to set the level of significance at 0.05. This choice is nearly universal, while 0.01 and 0.10 levels are also widely used. The common practice of setting $\alpha = 0.05$ is attributable to Sir R. A. Fisher's argument that one in twenty chance is a reasonable criterion for unusual sampling occurrence (Moore and McCabe, 1993; p.473). However, there is no scientific ground on this choice, and no reason to believe that these conventional choices are optimal by any criterion (Morrison and Henkel, 1970; p.307; Lehmann and Romano, 2005; p.57). From their survey of empirical studies published in *Journal of Econometrics*, Keuzenkamp and Magnus (1995) conclude that "the choice of significance levels seems arbitrary and depends more on convention and, occasionally, on the desire of an investigator to reject or accept a hypothesis". They also note that Fisher's theory of significance testing is intended for small samples, stating that "Fisher does not discuss what the appropriate significance levels are for large samples". It is highly unlikely that the level of significance that Fisher has adopted for his small sample analysis is appropriate for modern finance research in which large or massive data sets are widely used.

It is well known that sample size is an important factor for the outcomes of significance test. Kish (1959)[3] states that "in small samples, significant, that is, meaningful, results may fail to appear statistically significant. But if the sample size is large enough, the most insignificant relationships will appear statistically significant." Labovitz (1968)[4] also points out that sample size is one of the key factors for selecting level of significance, along with the power or probability of Type II error (β) of the test (acceptance of a false null hypothesis). More specifically, Leamer (1978) suggests that the level of significance be adjusted as a decreasing

[3] Reprinted in Morrison and Henkel (1970; p.139)
[4] Reprinted in Morrison and Henkel (1970; p.168)

function of sample size for sensible hypothesis testing. Klein and Brown (1984) make the same point in the context of model selection in the linear regression, with an observation that a smaller model is rejected too easily in large samples under a fixed level of significance.

Bayesian Methods of Hypothesis Testing

The Bayesian method of significance testing is based on the posterior odds ratio in favour of the alternative hypothesis (H_1) over the null (H_0), which is defined as

$$P_{10} \equiv \frac{P(H_1 \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_0)} \frac{P(H_1)}{P(H_0)}, \tag{1}$$

where $P(H_i)$ is the prior probability for H_i ; D indicates the data; $P(D|H_i)$ is the likelihood under H_i; and $P(H_i \mid D)$ is posterior for H_i, while $B_{10} \equiv P(D|H_1)/P(D|H_0)$ is called the Bayes factor. The evidence favours H_1 over H_0 if $P_{10} > 1$.

Consider the regression model of the form[5]

$$Y_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} + u_t$$

where Y is the dependent variable, X's are independent variables and u is a normal error term. Let F be the F-statistic for H_0: $\beta_1 = \ldots = \beta_p = 0$. Leamer (1978), based on the P_{10} obtained under the improper diffuse prior, derives the Bayesian critical value which increases with sample size: that is, the evidence favours H_1 when

$$F > \frac{T-k-1}{p}(T^{p/T} - 1), \tag{2}$$

where T is the sample size and $p \leq k$.

Concerned with the use of improper prior that may favour smaller models, Zellner and Siow (1979) derive the posterior odds ratio by using a proper diffuse prior, which is given by

[5] The notation β is used for both Type II error and the regression parameters, following the convention.

$$P_{10} = \left[\frac{\pi^{0.5}}{\Gamma[0.5(k_1+1)]} \frac{(0.5v_1)^{0.5k_1}}{[1+(k_1/v_1)F]^{0.5(v_1-1)}} \right]^{-1}, \quad (3)$$

where $\Gamma()$ is the gamma function and $v_1 = T - k_0 - k_1 - 1$, while k_0 and k_1 are the number of X variables under H_0 and H_1 respectively. Note that the above posterior odds ratio is derived under the even prior odds where $P(H_0)=P(H_1)$ in (1), and P_{10} in (3) is identical to B_{10}. When $P(H_0)\neq P(H_1)$, the value of P_{10} can be adjusted, depending on the researcher's prior beliefs, as Connolly (1991; p64) points out. To interpret the Bayes factor (B_{10}) or the posterior odd ratio (P_{10}), Kass and Raftery (1995) classify that the evidence against H_0 is "not worth more than a bare mention" if $2\log(B_{10}) < 2$; "positive" if $2 < 2\log(B_{10}) < 6$; "strong" if $6 < 2\log(B_{10}) < 10$; and "very strong" if $2\log(B_{10}) > 10$, where log is the natural logarithm.

Lindley Paradox and Revised Standard of Evidence

It is well known that the Bayesian method and the classical hypothesis testing can provide conflicting inferential outcome when the sample size is large (Lindley, 1957). This occurs largely because the classical method of hypothesis testing is conducted at a fixed level of significance, while the Bayesian methods imply the critical values as an increasing function of sample size, as in (2) and (3). In finance, Neal (1987) and Connolly (1989, 1991) acknowledge the effect of large sample size on significance testing and propose the Bayesian methods as an alternative. Connolly (1991) finds that earlier empirical evidence for the weekend effect in stock return is reversed if the Bayesian method is used. He attributes this to the interaction between large sample size and the fixed level of significance in the previous studies. Neal (1987; p.524) states that the large sample size inflates the test statistic and may generate spurious significant result. Keef and Khaled (2011) also adopt the Bayesian critical values given in (2) for the regression with a massive sample size. As we shall see in the next

section, none of the studies in our survey use the Bayesian methods, which is an indication that they are rarely adopted in modern finance research.

By relating nearly 800 *p*-values of the t-test reported in psychology journals with their respective Bayes factors, Johnson (2013) finds that *p*-value of 0.005 and 0.001 correspond to strong and very strong evidence against H_0, while the *p*-values of 0.05 and 0.01 reflect only modest evidence. Based on this, Johnson (2013) argues that the standards for statistical evidence in the classical method of hypothesis testing be revised and claim that the level of significance be set at 0.005 or 0.001. This is a reconciliation of the classical and Bayesian methods of significance testing, with a point that the conventional levels of significance are too lenient as a standard for evidence.

Line of Enlightened Judgement (Leamer, 1978)

It is well known that the probability of Type I error α is inversely related to the probability of Type II error (β). Leamer (1978) calls the line plotting the values α of against β values "the line of enlightened judgement". As an example, suppose we test for H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$ where β_1 is a slope coefficient in a linear regression model. Let b_1 be the least-squares estimator for β_1. For simplicity, it is assumed that b_1 follows a normal distribution with a fixed variance. We assume that $\beta_1 = 0.1$ represents the minimum value under H_1 where the effect size is economically significant. That is, if $\beta_1 < 0.1$, we regard the effect of X_1 to Y as being economically trivial or substantively unimportant. This is what Ziliak and McCloskey (2004) called the "minimum policy oomph", which is the minimum threshold of economic significance (see Thorbecke, 2004). Using this value, the power of the test can be calculated when the null hypothesis is violated to an economically meaningful extent (see, MacKinnon, 2002; p.634).

As an illustration, we consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u$$

where $X \sim$ iid N(0,1) and $u \sim$iid N(0,1). Setting the true value of β_1 to 0, we test for H_0: $\beta_1 = 0$ the usual t-test against H_1: $\beta_1 = 0.1$ at 5% level of significance. For $\alpha = 0.05$, we calculate $\beta = P(Z_1 < 1.64)$, where $Z_1 = (b_1\text{-}0.1)/s$ and s is the standard error of b_1. The line of enlightened judgement can be obtained by plotting the values of β corresponding to a grid of α values between 0 and 1[6]. Figure 1 provides an example of the line when the sample size 100, 500, and 2000. As the sample size increases, the line shifts towards the origin, as the value of β declines (or power increases) for each value of α.

According to Leamer (1978; Chapter 4), there are two ways of choosing the level of significance in relation to the line of enlightened judgement. The first is to minimize the expected loss from hypothesis testing; and the second is to minimize the maximum of the losses associated with Type I and II errors. To illustrate, let L_1 be the loss from Type I error and L_2 be the loss from making Type II error. The expected loss from hypothesis testing is $p\alpha L_1 + (1\text{-}p)\beta L_2$, where $p \equiv P(H_0)$. For simplicity and without loss of much generality, we assume $p = 0.5$ and $L_1 = L_2$, which means that Type I and II errors are of equal importance and that H_0 and H_1 are equally likely to be true. Then, the minimization of the expected to loss is equivalent to minimization of $\alpha+\beta$.

The 45 degree straight line from the origin in Figure 1 corresponds to the minimum value of $\alpha+\beta$, for each sample size considered. The horizontal line at $\alpha = 0.05$ corresponds to the

[6] Even though classical hypothesis testing conventionally sets α to a fixed value such as 0.05, the value of α is varied from 0 to 1 map its relationship with β and the sample size.

values of β when α is set to the conventional level of 0.05, regardless of sample size. Let (α^*,β^*) denote the values of α and β that minimize $\alpha+\beta$. When the sample is 100, (α^*,β^*) = (0.32, 0.32) with $\alpha^*+\beta^* = 0.64$, and the critical value implied by $\alpha^* = 0.32$ is 0.46. If α is set at the conventional level of 0.05, $\beta = 0.76$ and $\alpha+\beta = 0.81$. When the sample size is 500, (α^*,β^*) = (0.12, 0.12) and $\alpha^*+\beta^* = 0.24$ with the corresponding critical value of 1.15, while setting α = 0.05 gives $\alpha+\beta$ =0.30. When the sample size is 2000, (α^*,β^*) = (0.012, 0.012) and $\alpha^*+\beta^*$ = 0.024 with the critical value to be used 2.26. If $\alpha = 0.05$ is chosen in this case, $\alpha+\beta$ =0.052.

As well expected, the choice of α minimizing the value of $\alpha+\beta$ provides smaller expected loss than when α is set arbitrarily to 0.05. When the sample size is 2000, the latter case is more than twice riskier than the former. In the former case, the critical values of the test increase with the sample size; whereas, in the latter case, they are fixed and the expected loss is bounded by the value of α. This example illustrates how the level of significance can be chosen optimally in practice, based on the sample size and power of the test. It is also clear from this example that, when the sample size is massive (say, in the range of 20000 to 30000), the level of significance should be set at a much lower level than the conventional ones.

To minimize the maximum of the expected losses associated with Type I and II errors, the level of significance is chosen so that $\alpha/\beta = L_2/L_1$ (see Leamer; 1984). This means that the optimal choice (α^*,β^*) is made based on the line from the origin with the slope of L_2/L_1. Figure 2 illustrates the case where the value of L_2/L_1 is 3 and 1/3. If the loss from Type I error is higher (lower) relative to that from Type II error, then the lower (higher) level of significance should be chosen, given the sample size. For example, if the consequence of rejecting the true null hypothesis is a catastrophic event such as the global financial crisis, the

value of L_2/L_1 would be close to 0 as long as the value of L_2 is moderate. This means that, according to the Minimax rule, one should select a very low level of significance, imposing a tall bar for the rejection of the null hypothesis. In this case, the use of the conventional level of significance (e.g. 0.05) can be extremely risky.

According to Skipper et al. (1967)[7], the level of significance should be chosen with full awareness of the implications of Type I and Type II errors. Ziliak and McCloskey (2008) also state that "without a loss function, a test of statistical significance is meaningless" and that "significance testing without considering the potential losses is not ethically and economically defensible". This means that, when losses from these errors are known or assessable, they should be taken into account in choosing the level of significance. In finance research, it may be possible that losses from incorrect decision are estimated. For example, an investor can estimate the possible (relative) losses, when the investment decision is made based on the significance testing of an asset pricing model.

3. Survey of Finance Research Papers

To gather information as to how significance testing is being conducted in finance research, we conduct a survey of the papers published in four journals in 2012: *Journal of Finance* (*JF*), *Journal of Financial Economics* (*JFE*), *Journal of Financial and Quantitative Analysis*, and *The Review of Financial Studies* (*RFS*). To simplify the analysis and comparison, we restrict our attention to the papers which use the linear regression method (time series, cross-sectional, and panel). The total number of published articles is 320. We exclude (i) purely theoretical papers with no empirical contents; (ii) the papers based on purely time series methods; (iii) and those that use non-linear (e.g., probit, logit or GMM) or purely Bayesian

[7] Reprinted in Morrison and Henkel (1970; see p.160)

estimation methods. The papers that do not provide sufficient information (sample size, t-statistic, or p-value) are also excluded. After these exclusions, the total number of papers included in our analysis is 161. There are papers which report the *p*-values only: for these papers, where possible, the t-statistics are calculated or estimated from the reported *p*-values. It is usually the case that an empirical paper reports a number of regression results, due to robustness check, subsample analysis, or sensitivity analysis. To avoid analysing qualitatively similar results from the same study, only the details associated with the baseline or the most representative regression are recorded for each study. We record the t-statistic associated with the key variable of interest in the baseline regression.

Our survey has revealed the following main points:

- Large or massive sample size is widely used. The mean sample size is 161200 (median is 4719), with the maximum of 6.96 million and the minimum of 11. Nearly 42% (31%) of the studies use the sample size greater than 10000 (20000);
- The conventional levels of significance (1%, 5%, 10%) are exclusively used. None of the surveyed papers adopt other levels. There are five papers which do not explicitly mention the chosen level of significance, although it appears that they implicitly assume the conventional levels for their analysis;
- All the surveyed papers use the classical null hypothesis testing method, and none makes use of the Bayesian methods of Leamer (1978) or Zellner and Siow (1979);
- Only one paper discusses the power of the test for significance testing;
- None of the papers has considered the potential losses from making incorrect decisions;
- Only 22% of the papers discuss or present at least one diagnostic test for autocorrelation or heteroskedasticity;
- Only three papers use the generalized (or weighted) least squares (GLS) for estimation;

- Only two of the surveyed papers replicate previous studies;
- All surveyed papers use the t-statistics or *p*-value, with none reports confidence interval; and
- The number of studies with statistically insignificant results is unreasonably low (more details in the next section).

4. Analysis and Discussion

In this section, we provide an analysis of the survey results and compare alternative methods discussed in the previous section using Monte Carlo experiment and empirical application.

Comparison of Conventional and Bayesian Methods: Survey

Figure 3 presents the scatter plot of the absolute value of t-statistics against the natural log of the sample sizes from our survey. The relationship between the two is clearly positive, with a larger variation of t-statistics when the sample size is larger. This may be the reflection of the property that statistical significance is inflated with increasing sample size. The horizontal line corresponds to 1.645, which is two-sided critical value at the 10% level of significance. It appears that only three studies report statistically insignificant results at the 10% level, which means that nearly 98% of the studies report the results with statistical significance. This proportion may make statistical sense only when the null hypothesis is false for all studies, which is a strong evidence for publication bias in favour of statistical significance. This means the studies with statistically significant results have higher chance of publication than those not; and that many new and important findings may be severely disadvantaged if their results are statistically insignificant (see Sterling, 1959; and Kendall, 1970; p.89). Confronted with this bias, finance researchers may increasingly be motivated to gather additional data points in order to secure statistical significance and higher chance of publication.

Figure 4 presents the plot of the posterior odds ratio given in (3) (in the likelihood ratio scale of $2\log(P_{10})$) against the natural log of sample size. The horizontal lines correspond to 2 and 10, which are the criteria points for the evidence against H_0 according to Kass and Raftery (1995). According to the Bayesian criteria, it turns out that 42% of the surveyed papers report the relationships which are "not worth more than a bare mention", while those with more than "positive" relationship is 58%. Only 32% of the studies correspond to "very strong" evidence against the null hypothesis. This means that statistical significance of many surveyed papers can be challenged if the posterior odds ratio were instead used. We note that similar results are obtained if Leamer's (1978) Bayesian critical values given in (2) are used.

One may argue that the above observation is a realization of the Lindley (1957) paradox mentioned earlier. However, it is evident from Figure 3 that the conflict between the two methods also occurs when the sample size is relatively small. For the studies with the sample size is less than 1000, around 39% of the studies show conflict between the two methods. Hence, the Lindley paradox is not confined to the studies with massive sample size. If one adopts the levels of significance such as 0.001 (two-tailed critical value: 3.29) or 0.005 (two-tailed critical value: 2.81) following Johnson (2013), only around 52% or 67% of the studies are found to be statistically significant respectively. These proportions are close to that of "more than positive" relationship based on the posterior odds ratio, which is around 59%. This indicates that a level of significance in the vicinity of 0.001 or 0.005 is more appropriate for empirical research in finance, as a revised standard of evidence, similarly to the findings of Johnson (2013).

It is found that more than 36% of the surveyed papers report the t-statistics greater than four in absolute value. This proportion is extremely high, given that the probability of the standard normal distribution (or a Student-t distribution) taking such a value is practically zero. This may be explained in the context of Meehl's (1978) conjecture from psychology which states that, "in non-experimental settings with large sample sizes, the probability of rejecting the null hypothesis of nil group differences in favour of a directional alternative is about 0.50" (Gigerenzer, 2004). Waller (2004) empirically confirms this conjecture using a sample of size 81000. On this point, Gigerenzer (2004) points out that "the combination of large sample size and low *p*-value is of little value in itself".

Comparison of the Alternative Methods: Monte Carlo Comparison

We compare the properties of alternative methods of selecting the level of significance in repeated sampling, paying attention to the effect of increasing sample size. We conduct a Monte Carlo experiment using the regression model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$X_i \sim N(0,1)$ are independent variables fixed over Monte Carlo trials and $u \sim N(0,1)$ is random error term, setting $(\beta_0, \beta_1, \beta_2) = (0, 0.01, 1)$. H_0: $\beta_1 = 0$; H_1:$\beta_1 > 0$ is tested using the usual t-test. We assume that the value of "minimum policy oomph" of β_1 is 0.05. That is, although the true value of β_1 is set to 0.01, it is not economically sensible to reject H_0 because the effect size of 0.01 is economically trivial. We consider two cases for losses $L_2/L_1 \in \{5, 1/5\}$. The above exercise is repeated 1000 times.

Table 1 present the probability of rejecting H_0. If α is set to 0.05 regardless of sample size, the null hypothesis is rejected with increasing frequency. This means that, as the sample size grows, we are more and more likely to reject the null hypothesis in favour of statistical

significance of the effect size, which is economically trivial. With the Bayesian methods, the rejection probability is substantially lower, especially as the sample size increases. By minimizing $\alpha+\beta$ (expected loss), the probability of rejection of the null hypothesis decreases as the sample size gets massive. It is worth noting that, as the sample size increases, the role of relative losses is becoming less important in choosing the level of significance, in contrast with the case when the sample size is smaller where it makes substantial difference.

Comparison of the Alternative Methods: An Example for Small Sample Case

The above Monte Carlo experiment examines the case when the sample size is large or massive. There are also many situations in finance where significance testing is conducted with a small sample where the power of the test is likely to be low. According to MacKinnon (2002; p.633), researchers pay surprisingly little attention to the power when they conduct hypothesis testing, which is confirmed by the survey of Ziliak and McCloskey (2004). When the power is low, the level of significance much larger than the conventional levels may be more appropriate (Kish, 1959)[8]. As an illustrative example, we estimate the CAPM equation for the GM (General Motors) stock using monthly data from 1998 to 2008. Suppose that a financial analyst is making an investment decision as to where the GM stock should be included in her portfolio. If its CAPM beta is equal to or less than one, then the stock will be included. The analyst believes that the minimum threshold value for the CAPM beta value to be economically greater than one is 1.10.

Table 2 presents the estimation and significance testing results. The beta estimates are stable over different sample periods, slightly bigger than 1.10 for all cases. However, if the t-test is conducted at 5% level of significance for the null hypothesis that CAPM beta is equal to one

[8] Reprinted in Morrison and Henkel (1970; p.139)

against the alternative that it is greater; the null is accepted for all cases since the t-statistics are smaller than the critical value of 1.645. Leamer's (1978) Bayesian critical values and posterior odds ratios also favour the null hypothesis. This is despite the analyst's belief that CAPM beta estimate represents a strong effect size, with its value greater than the threshold for economic significance.

Figure 2 plots the line of enlightened judgement when the whole sample size is used. The 45 degree line corresponds to the level of significance minimizing the $\alpha+\beta$, indicating the optimal level of significance of 0.40. If the level of significance is chosen to minimize the expected loss, the critical values are 0.25, 0.22, and 0.11 for each sample period, corresponding to the chosen level of significance of 0.40, 0.41, and 0.45. This means that the analyst rejects the null hypothesis in favour of the alternative hypothesis, and decides not to include the GM stock in her portfolio. The blue line in Figure 2 corresponds to the case where $L_2/L_1=1/3$, where the chosen level of significance is 0.21 with the critical value of 0.82. In this case, the analyst accepts the null hypothesis. Hence, if the loss from making Type I error is around three times greater than that from Type II error or higher, the investor decides to include the GM stock in her portfolio.

This example illustrates how investment decision can be different under alternative methods of choosing the level of significance. Note that the same results hold for different values of minimum policy oomph, such as the CAPM beta value of 1.2 or higher, as the last column of Table 2 shows. This example has strong implications to finance research in practice, as Ziliak and McCloskey (2004; p.8 and p.15) argue that "without a loss function a test of statistical significance is meaningless" and that "significance testing without considering the potential losses is not ethically and economically defensible". As Engsted (2009) point out, the use of

loss function may be "in theory appealing but in practice often inapplicable". However, mindless and mechanical use of the conventional level of significance can be avoided when researchers consider potential consequences of incorrect decisions.

This example also demonstrates that the use of conventional level of significance is not optimal when the power of the test is low, and may lead to incorrect decision. In this case, a higher level of significance may be more appropriate as seen above. We note that this claim is not new, as Winer (1962: I3) states that "when the power of the tests is likely to be low under these levels of significance, and when Type I and Type II errors are of approximately equal importance, the 0.3 and 0.2 levels of significance may be more appropriate than the .05 and .01 levels" (quoted from Skipper et al., 1967)[9].

5. Further Discussions

Statistical Significance vs. Economic Significance

The focus of this paper has been the practice of significance testing in empirical finance. However, economic significance is also an important factor in establishing the existence of an association. It is well known that statistical significance does not necessarily indicate that an association is substantively important. Kendall (1970; p.89) argues that many conceptually important and significant results in social research may not attain statistical significance. Ziliak and McCloskey (2008) report that confusion between statistical significance and economic significance is widespread in economics, from the textbook level to academic research; while Hoover and Sigler (2008) provide evidence against their claims. Engsted (2009) further refutes the claims made by Ziliak and McCloskey (2008) with case studies and examples in finance research, pointing out that they focus too much on "real error" totally

[9] Reprinted in Morrison and Henkel (1970; p.157)

neglecting "statistical error". The latter is also important in distinguishing between economic significance and the effects of random sampling error, for which significance testing plays an important role (see Horowitz, 2004). We find that, overall, the authors in our survey discuss economic importance and effects size carefully, although we occasionally find the studies where such confusion is made or the effect size is not properly analysed. In this paper, we do not attempt to quantify these features, since the lessons from the ongoing controversies over the economic significance vs. statistical significance are highly suggestive to researchers in empirical finance.

Other Statistical Issues

Cumming (2013) argues that researchers should move away from the dichotomous decisions based on the *p*-value or t-statistic, by reporting confidence interval and analysing effect size. As we have found from our survey, finance researchers are strongly in favour of *t*-statistic or *p*-value over confidence interval as a means of evaluating statistical significance. Ziliak and McCloskey (2004; p.534) note that economists seldom report confidence interval, contributing to widespread neglect of effect size and economic significance. It is also alarming that only a small proportion of the papers in our survey adopts regression diagnostics and conducts GLS estimation. While the use of robust standard error estimators is widespread[10], our survey finds that many researchers in empirical finance do not attempt to improve estimation of effect size and mitigate the "real error", through extensive diagnostic testing and efficient model estimation. In what follows, we present Monte Carlo evidence to demonstrate that the use of robust standard error estimator does not provide the best outcome for significance testing.

[10] Petersen (2009) finds that more than 50% of the studies he surveyed use a form of robust standard error estimators (similarly to the results from our survey), and points out that they are often incorrectly used or adopted with little statistical justifications.

As a means of comparison, the interval score of confidence interval proposed by Gneiting and Raftery (2007; p.370) is used. For a 100(1-α)% confidence interval [L,U], it is given by

$$S_\alpha(L,U;x) = (U-L) + \frac{2}{\alpha}(L-x)I(x<L) + \frac{2}{\alpha}(x-U)I(x>U)$$

where I is an indicator function which takes 1 if the condition insider the bracket is satisfied and 0 if otherwise; and x is the value of the parameter of interest. If the interval covers x, the score is its length; if otherwise, a penalty term is added, which is how much the interval misses x scaled by $2/\alpha$. In the event that the interval slightly (clearly) misses x, a small (heavy) penalty is imposed. The interval score measures the quality of the probabilistic statement implied by a confidence interval. As an alternative measure, we calculate the coverage rate of confidence interval, which is the proportion of confidence interval covering the true parameter value x in repeated sampling. Note that the coverage is a dichotomous measure similar to the decision based on t-statistic or p-value.

We consider a simple case of a cross-sectional regression with heteroskedastic error term of the form

$$Y = 1 + \beta_1 X + u; Var(u) = \sigma^2 X ,$$

where X~N(0,1) is non-random independent variable and u~N(0,0.25^{2}X) is the random error tem. Setting β_1=1, we calculate the mean interval score and coverage rate for β_1 over 1000 Monte Carlo trials. Table 3 presents the results of the confidence intervals based on OLS, White's, and GLS standard error estimators. The OLS-based confidence interval under-covers the true value, especially when the sample size is large, which means they are often too short, rejecting the H_0: $\beta_1 = 1$ too often in repeated sampling. Those based on White's and GLS estimators show much better performance, with their coverage rates fairly close to (1−α). It appears that White's standard error estimator provides the results as desirable as the

GLS, when a dichotomous measure is used as a means of comparison. However, if the interval score is used, the confidence interval based on White's estimator performs almost identically to the OLS-based confidence interval, while the GLS-based confidence interval provides much smaller interval score than the other two. The results are evident when α is set to 0.05 as well as 0.25. This indicates that the use of robust standard error estimator adds little gain if the performance of significance testing is evaluated using a quality-based measure instead of a dichotomous one. The GLS-based confidence interval is the clear winner in terms of the interval score[11].

Replicability of the Published Results

Replicability and reproducibility of scientific studies are essential for their credibility and stronger empirical validity. It is recognized as the essential part of the scientific method which "allows science to move forward" (Anderson et al., 2008; p.100). However, there are widespread concerns about non-reproducibility of scientific studies. In particular, Anderson et al. (2008) and Ioannidis and Doucouliagos (2013) observe low replicability of economics and business research, despite the revolutions in data and software archives. The degree of replicability of empirical finance research is not fully known: to the best of our knowledge, there has been no dedicated study that has evaluated the degree of reproducibility of empirical finance studies. In our survey, we have found only two studies which explicitly replicated the previous studies.

In economics, a number of journals have explicit editorial policy on replicability or data availability. *Journal of Applied Econometrics* has a dedicated replication section where the readers are encouraged to replicate or reproduce the findings of published in nine top journals

[11] We also conduct a similar experiment in the context of panel regression following Petersen (2009), but the results are similar to those reported in Table 3. For simplicity, the details are available from the corresponding author on request.

in economics. *American Economic Review, Journal of Applied Econometrics, Journal of Money, Credit and Banking, Journal of International Economics, Review of Economics and Statistics*, *Journal of Business and Economic Statistics*, *Journal of Socio-Economics*, and *International Journal of Forecasting*, among others, either require or encourage the authors to make their data sets and computational resources available. Anderson et al. (2008) provide an excellent survey of the data/code archives in economics journals and discuss the benefits they will bring for more and better research.

From our survey, we have found that nearly all studies clearly indicate the data sources. However, it is rarely the case in finance that the exactly same data set can be obtained from databases, due mainly to data cleaning issues and treatment of missing values, among others. We have contacted the authors (by email) of 50 randomly selected papers in our survey and asked if they could share the (cleaned-up and regression-ready) data for the purpose of replication. Only four have responded with the data sets, with detailed information and computational resources. Twelve authors have declined to share the data, with the ten giving adequate reasons such as the copyright issues, contract with the data vendor, and confidentiality. However, the rest have not responded to date. This suggests that data and computational resources are not actively shared, a sign that replicability of the published studies in finance is rather limited.

Why Does This Matter?

Ziliak and McCloskey (2008) provide a number of practical examples where incorrect decisions based on statistical significance incur social costs and impact human lives. In particular, Ziliak and McCloskey (2004; p.530) state that "a statistically insignificant coefficient in a financial model may give its discoverer an edge in making a fortune; and a

statistically significant coefficient in the same model may be offset in its exploitation by transactions costs". This point indicates that the decisions based heavily on statistical significance (but lightly on economic significance) may often be incorrect and can be highly costly. In a recent study based on a survey of academic economists, Soyer and Hogarth (2012) provide evidence that regression statistics and statistical significance create illusion of predictability. Their survey shows that the surveyed economist provide better predictions when they are presented with simple visual representation of the data than when they make predictions purely based on statistical measures such as R^2 or *t*-statistic (see also Armstrong; 2012; and Ziliak, 2012). This means that heavy reliance on statistical measures can create false sense of significance or association.

As Hoover and Siegler (2008; p.1) point out, "properly used, significance tests are a valuable tool for assessing signal strength, for assisting model specification, and determining causal structure". The main point of our paper is that the way significance testing being conducted in modern finance research is conducive to spurious statistical significance and incorrect decisions. We have presented the survey results for the current practice; and discussed alternative methods which can overcome these problems. In defence of a mathematical model that was partly blamed for the outbreak of the global financial crisis, Donnelly and Embrechts (2010) argue that it is misuse of mathematic models by end-users that has caused the crisis, not the mathematical model itself. We should also realize that abuse and misuse of significance testing has a potential for disaster; and that there are ways that we can improve the practice of significance testing in empirical finance.

6. Concluding remarks

Abuse and misuse of significance testing have been widely criticised for many years (see, for example, Morrison and Henkel, 1970). There have been numerous calls for change, but they are largely ignored in modern statistical research in many fields of science. As a result, serious questions have been raised about the credibility of statistical research in many areas of science (see, for example, Ioannidis; 2005). Recently, there are renewed calls to improve and ensure research integrity and credibility, by making substantial changes in the way statistical research is being conducted (Ellis, 2010; Cumming, 2013; Ioannidis and Doucouliagos, 2013). Although an integral part of empirical research, the current practice of significance testing in finance has not received proper attention to date.

From a survey of the articles recently published in four top-tier journals in finance, we have found that large or massive sample size is widely used with a fixed level of significance; and that publication bias in favour of statistical significance is evident. In the former case, statistical significance can be seriously over-stated or spurious (Neal, 1978); while, in the latter case, many important new studies may heavily disadvantaged for publication if they were unable to produce statistically significant result (Sterling, 1970). Using a Bayesian method and a much lower level of significance as a revised standard for evidence, we find that statistical significance is questionable in many papers in our survey. We also discuss how the optimal level of significance can be chosen, with explicit consideration of the key factors such as sample size, power of the test, and expected losses from incorrect decisions, following Leamer (1978).

Our findings strongly suggest that finance researchers move away from mindless use of the conventional levels of significance, and choose the level carefully taking account of these key

factors. As Engsted (2009; p.401) point out, the use of conventional level "mechanically and thoughtlessly in each and every application" is meaningless. When sample size is large or massive, the use of the Bayesian method of Zellner and Siow (1979) is recommended; or the level of significance should be set at a much lower level than the conventional levels following Johnson (2013). When the sample is small, the level of significance should be chosen (usually at a level much higher than the conventional ones), considering power of the test or expected losses from incorrect decisions. We also find that the degree of replicability of empirical finance research is rather limited.

We conclude that finance researchers should substantially change the way they conduct significance testing in their academic research, bearing in mind potential social cost that their research outcomes can bring about, either directly or indirectly. Given the concerns raised in other areas of science recently, the "new statistics", the term coined by Cumming (2013), suitable for finance research should be established, in order to improve its credibility and integrity. Altman (2004) argues that the problems can only be corrected by changing the institutional parameters related to publication. Anderson et al. (2008) call for implementation of mandatory data/code archives and explicit replication policies by journals. Amid the controversies over the outbreak of the global financial crisis attributing its cause to faulty statistical models, we should always be "scientifically critical, socially honest and adhere to the highest ethical principles, especially in the face of temptation", as Donnelly and Embrechts (2010) suggest.

References

Anderson, R. G., Green, W.W., McCullough, B.D., and Vinod, H.D., 2008, The role of data/code archives in the future of economic research, Journal of Economic Methodology, 15, 1, 99-119.

Armstrong, J. S. 2012, Illusions in regression analysis, International Journal of Forecasting 28, 689-694.

Altman, M. 2004, Statistical significance, path dependency, and the culture of journal publication: Comment on "Size Matters", Journal of Socio-Economics 33, 651-663.

Connolly, R. A. 1991, A posterior odds analysis of the weekend effect, Journal of Econometrics 49, 51-104.

Connolly, R. A. 1989, An Examination of the Robustness of the Weekend Effect, The Journal of Financial and Quantitative Analysis, Vol. 24, No. 2, pp.133-169.

Cumming, G., 2013, The New Statistics: Why and How, Psychological Science, DOI: 10.1177/0956797613504966

Donnelly, C. and Embrechts, P., 2010, The devil is in the tails: actuarial mathematics and the subprime mortgage crisis, www.math.ethz.ch/~baltes/ftp/CD_PE_devil_Jan10.pdf.

Ellis, P. D. 2010, Effect sizes and the interpretation of research results in international business, Journal of International Business Studies 41, 1581-1588.

Engsted, T. 2009, Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak, Journal of Economic Methodology, 16, 4, 393-408.

Gigerenzer, G. 2004, Mindless statistics: Comment on "Size Matters", Journal of Socio-Economics 33, 587-606.

Gneitung, T. and Raftery, A. E., 2007, Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, Vol. 102, 359-378.

Horowitz, J. L., 2004, Comment on "Size Matters", Journal of Socio-Economics 33, 551–554.

Hoover, K. D. and Siegler, M. V., 2008, Sound and fury: McCloskey and significance testing in economics, Journal of Economic Methodology, 15:1, 1-37.

Ioannidis, J.P.A., 2005, Why most published research findings are false. PLoS Medicine 2: e124.

Ioannidis, J.P.A , and Doucouliagos , C. 2013, What's to know about credibility of empirical economics? Journal of Economic Surveys (2013) Vol. 27, No. 5, pp. 997–1004.

Johnson, V. E., 2013, Revised standards for statistical evidence, Proceedings of the National Academy of Sciences, www.pnas.org/cgi/doi/10.1073/pnas.1313476110.

Kass, R. E. and Raftery, A. E., 1995, Bayes Factors, Journal of the American Statistical Association, Vol. 90, No. 430, pp. 773-795.

Kish, L. 1959, Some statistical problems in research design, American Sociological Review 24, 328-338.

Keef, S. P. and Khaled, M. S., 2011, Are investors moonstruck? Further international evidence on lunar phases and stock returns, Journal of Empirical Finance 18, 56-63.

Kendall, P. 1970, Note on Significance Tests, Chapter 7, The Significance Test Controversy: A Reader, edited by D. E. Morrison and R. E. Henkel. Aldine Transactions, New Brunswick, NJ.

Keuzenkamp, H.A., Magnus, J., 1995. On tests and significance in econometrics. Journal of Econometrics 67 (1), 103–128.

Klein, R. W., Brown, S. J. 1984, Model selection when there is "minimal" prior information, Econometrica 52, 1291-1312.

Labovitz, S. 1968, Criteria for selecting a significance level: a note on the sacredness of 0.05, The American Sociologist 3, 200-222.

Leamer, E. 1978, Specification Searches: Ad Hoc Inference with Nonexperimental Data, Wiley, New York.

Lehmann E.L., Romano, J.S., 2005, Testing Statistical Hypothesis, 3rd edition, Springer, New York.

Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.

MacKinnon, J. G., 2002, Bootstrap inference in Econometrics, Canadian Journal of Economics 35(4), 615-644.

McCloskey, D., Ziliak, S., 1996. The standard error of regressions. Journal of Economic Literature 34, 97–114.

Meehl, P.E., 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46, 806–834.

Moore D.S. McCabe, G.P. 1993, Introduction to the Practice of Statistics, Second Edition, W.H. Freeman and Company, New York.

Morrison, D. E., Henkel, R. E. 1970, Significance tests in behavioural research: pessimistic conclusions and beyond, Chapter 31, The Significance Test Controversy: A Reader, edited by D. E. Morrison and R. E. Henkel. Aldine Transactions, New Brunswick, NJ.

Neal, R., 1987, Potential Competition and Actual Competition in Equity Options, The Journal of Finance, Vol. 42, No. 3, pp. 511-53.

Petersen, M. A. 2009, Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches, The Review of Financial Studies 22, 435-480.

Skipper J. K. JR., Guenther, A. L., Nass, G. 1967, The sacredness of .05: a note on concerning the use of statistical levels of significance in social science. The American Sociologist 2, 16-18.

Soyer, E., Hogarth R. M. 2012, The illusion of predictability: How regression statistics mislead experts, International Journal of Forecasting 28, 695-711.

Sterling, T. D., 1959, Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa, Journal of the American Statistical Association 54, 30-34.

Thorbecke, E. 2004, Economic and Statistical Significance: Comment on "Size Matters", Journal of Socio-Economics 33, 571–575.

Waller, N.G., 2004. The fallacy of the null hypothesis in soft psychology. Applied and Preventive Psychology 11, 83–86.

Winer, B. J., 1962, Statistical Principles in Experimental Design, New York, McGraw-Hill.

Zellner, A. and Siow, A. 1979, Posterior odds ratio of selected regression hypotheses, http://dmle.cindoc.csic.es/pdf/TESTOP_1980_31_00_38.pdf.

Ziliak, S. T., 2012, Visualizing uncertainty: On Soyer's and Hogarth's ''The illusion of predictability: How regression statistics mislead experts'', International Journal of Forecasting 28, 712-714.

Ziliak, S. T., and McCloskey, D.N., 2004, Size matters: the standard error of regressions in the American Economic Review, Journal of Socio-Economics 33, 527-546.

Ziliak, S. T., and McCloskey, D.N., 2008, The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. The University of Michigan Press.

Table 1. Probability of rejecting H_0: $\beta_2 = 0$ when the true value of β_2 is economically trivial.

T	$\alpha = 0.05$	Leamer	$P_{10} > 1$	Min $\alpha + \beta$	$L_2/L_1=5$	$L_2/L_1=1/5$
100	5.7	1.6	2.1	44.5	76.2	17.6
1000	9.6	1.4	1.1	32.1	55.4	14.7
5000	17.1	1.8	1.6	14.5	23.3	7.9
10000	26.8	2.7	2.3	8.3	13.3	5.3
20000	43.9	4.8	4.4	1.8	3.4	0.9
30000	53.5	7.8	7.2	0.6	1.2	0.0

The entries are the probability.
Leamer: Leamer's (1978) Bayesian critical values; P_{10}: posterior odds ratio in favour of H_1 (Connolly, 1991); L_i: the expected losses under H_i.

Table 2. Comparison of alternative methods for the CAPM application

Period	Beta	t-statistic	Critical Values			P_{10}
			Leamer	Min $\alpha + \beta$		
				beta*=1.1	beta*=1.2	
1998-2008	1.15	0.74	2.21	0.25 (0.40)	0.51 (0.31)	0.09
2004-2008	1.19	0.43	2.02	0.11 (0.45)	0.23 (0.41)	0.12
1998-2002	1.12	0.54	2.02	0.22 (0.41)	0.45 (0.33)	0.12

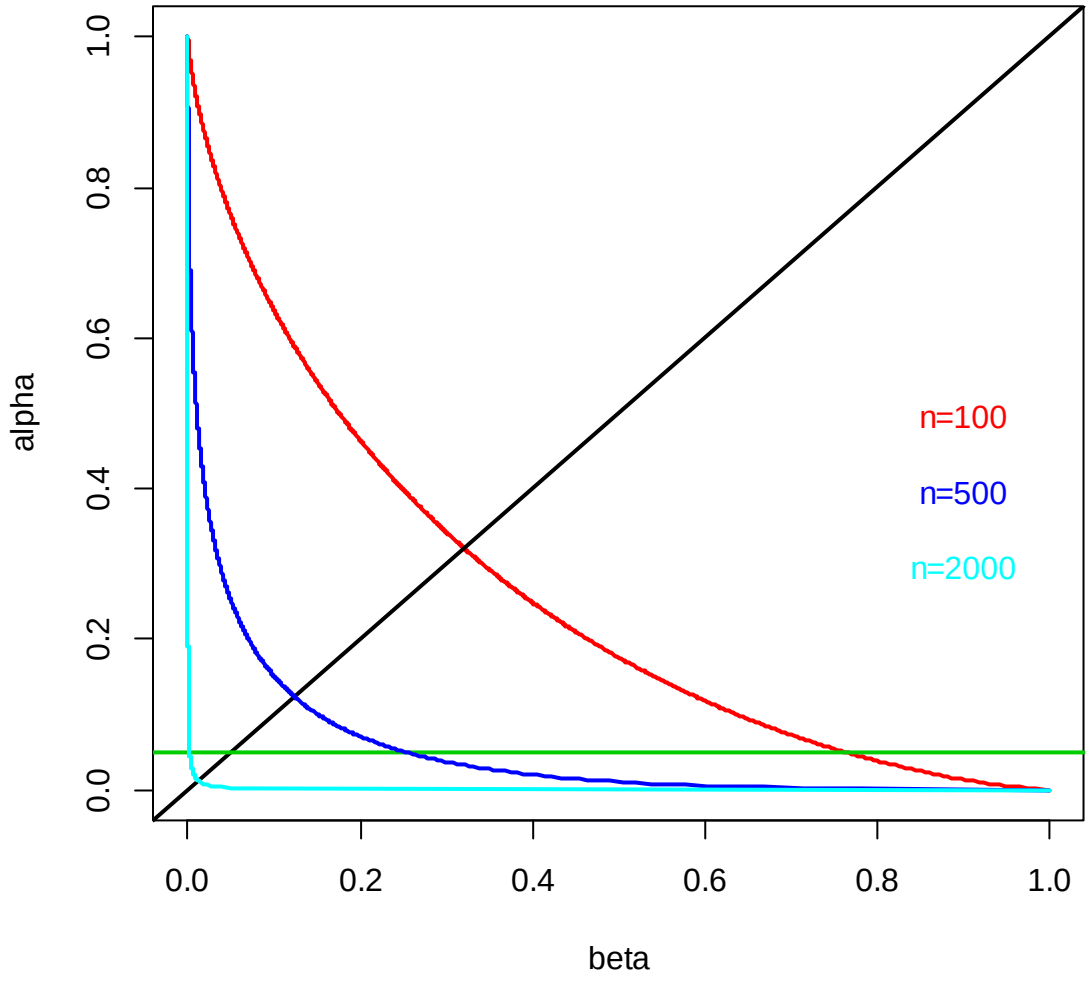
beta*: minimum oomph, the value below which the effect is economically trivial; P_{10}: posterior odds ratio in favour of H_1 (Connolly, 1991); Leamer: Leamer's (1978) Bayesian critical values
The figures in the bracket under the critical values for Min $\alpha + \beta$ are the chosen levels of significance.

Table 3. Performance of alternative standard error estimators

	Coverage			Score		
	OLS	White	GLS	OLS	White	GLS
α=0.05						
100	93.7	95.1	95.1	0.81	0.80	0.53
1000	94.2	96.5	96.3	0.24	0.24	0.15
2000	94.6	96.9	95.7	0.16	0.17	0.06
5000	90.5	93.7	94.6	0.12	0.12	0.03
10000	91.3	93.8	94.5	0.09	0.08	0.02
α=0.25						
100	71.1	74.2	75.6	0.67	0.57	0.48
1000	72.3	76.3	77.8	0.21	0.17	0.14
2000	71.7	76.0	76.4	0.14	0.12	0.06
5000	69.4	73.4	72.5	0.09	0.08	0.03
10000	68.9	72.7	73.2	0.07	0.06	0.02

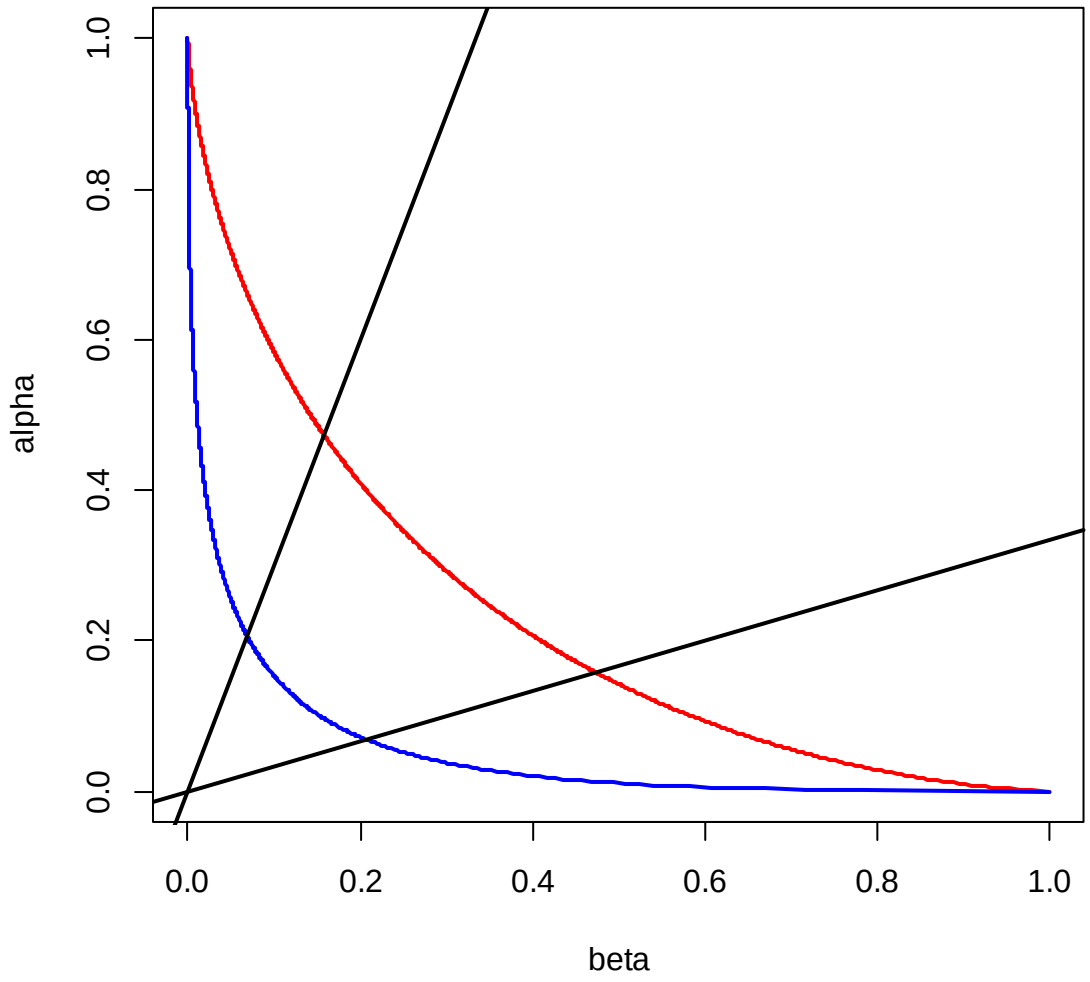
Coverage: the proportion of the confidence interval covering the true value
Score: interval score of Gneitung and Raftery (2007)

Figure 1. Line of enlightened judgement under different sample sizes



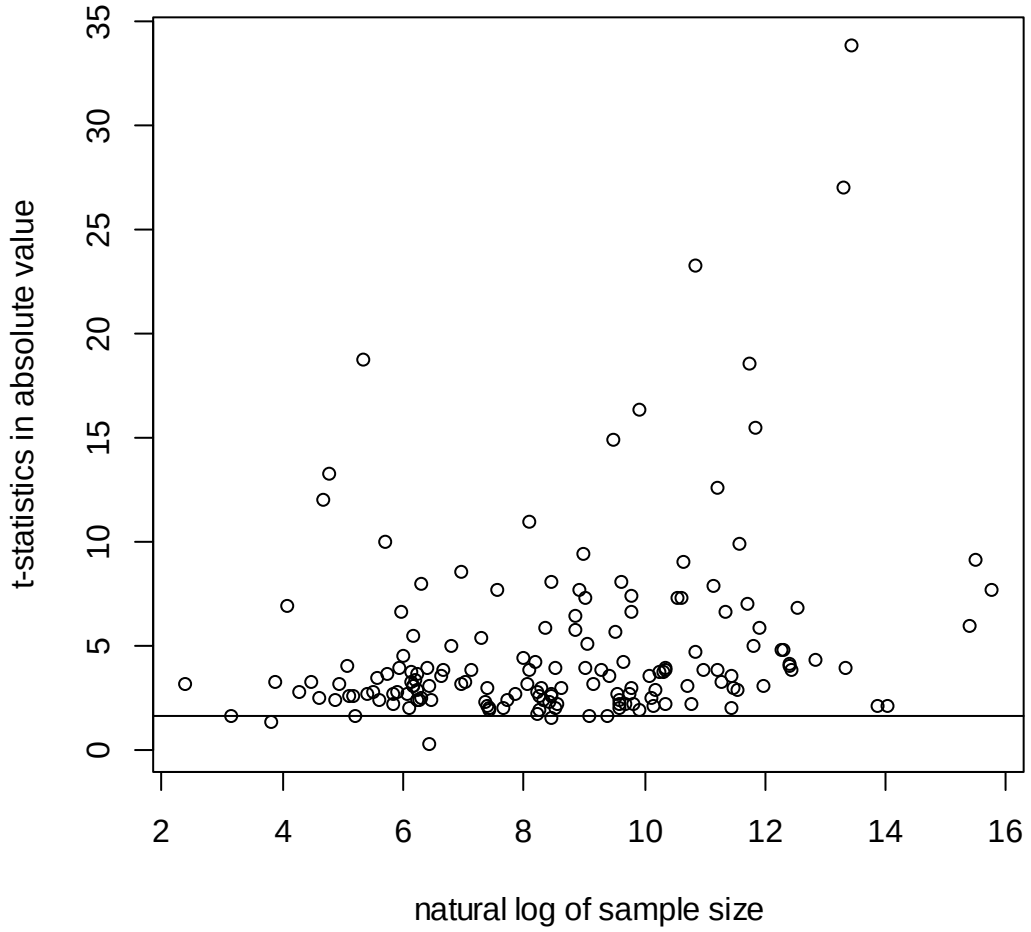
alpha = probability of Type I error (level of significance)
beta = probability of Type II error
n = sample size
The 45 degree line connects the points of Min $\alpha + \beta$, the horizontal line is at 0.05.

Figure 2. Line of enlightened judgement under different expected losses from Type I and II errors



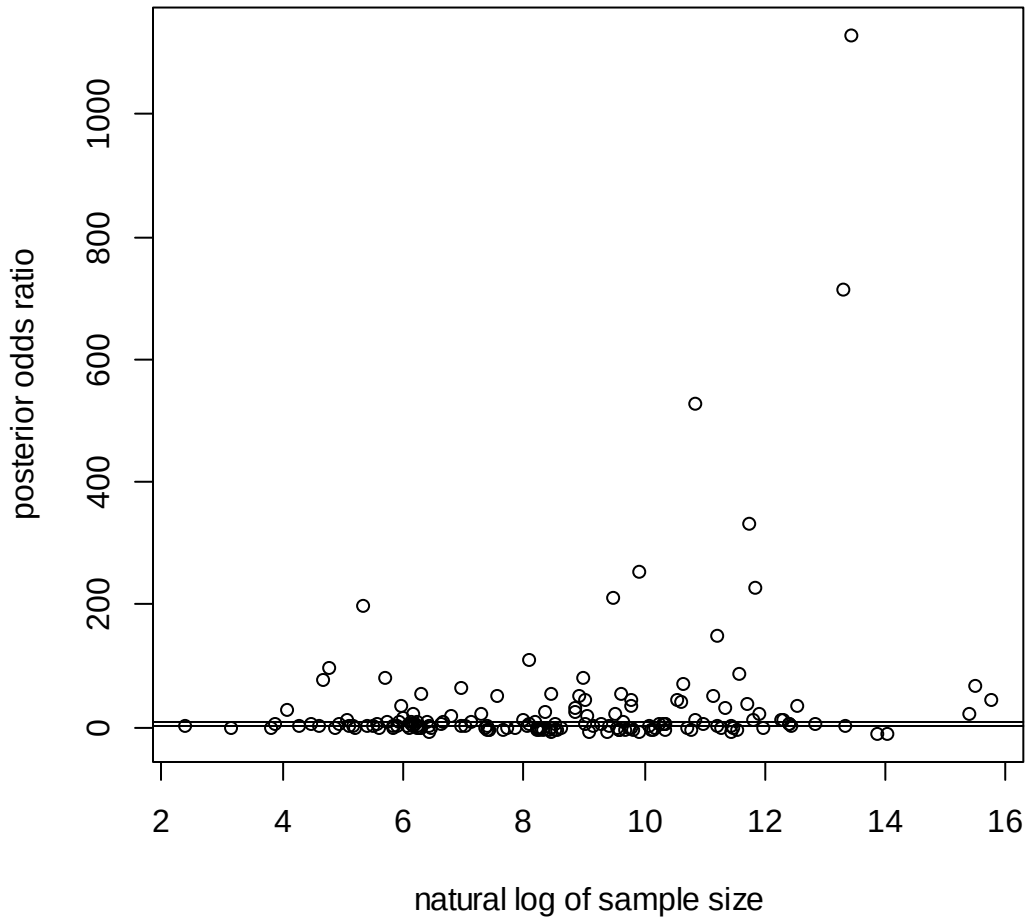
alpha = probability of Type I error (level of significance)
beta = probability of Type II error
The slope of the black straight lines is L_2/L_1, the relative expected losses from Type II and I errors. .

Figure 3. Plot of t-statistics (in absolute values) against sample size (in natural log)



Horizontal Line = 1.645

Figure 4. Plot of posterior odds ratios (in likelihood scale) against sample size (in natural log)



2log(Bayes factor)	Evidence against H_0	Proportion
Less than 2	Not worth more than a bare mention	0.42
Between 2 and 6	Positive	0.18
Between 6 and 10	Strong	0.07
Greater than 10	Very strong	0.32

The horizontal lines in the graph correspond to 2 and 10 (Kass and Raftery, 1995).