ELSEVIER

Information and Software Technology 39 (1997) 285–295

INFORMATION AND SOFTWARE TECHNOLOGY

Statistical power and its subcomponents – missing and misunderstood concepts in empirical software engineering research

James Miller[1], John Daly[2], Murray Wood, Marc Roper, Andrew Brooks

Department of Computer Science, University of Strathclyde, Livingstone Tower, Richmond Street, Glasgow G1 1XH, UK

Received 17 January 1996; revised 31 July 1996; accepted 6 August 1996

Abstract

Recently we have witnessed a welcomed increase in the amount of empirical evaluation of Software Engineering methods and concepts. It is hoped that this increase will lead to establishing Software Engineering as a well-defined subject with a sound scientifically proven underpinning rather than a topic based upon unsubstantiated theories and personal belief. For this to happen the empirical work must be of the highest standard. Unfortunately producing meaningful empirical evaluations is a highly hazardous activity, full of uncertainties and often unseen difficulties. Any researcher can overlook or neglect a seemingly innocuous factor, which in fact invalidates all of the work. More serious is that large sections of the community can overlook essential experimental design guidelines, which bring into question the validity of much of the work undertaken to date.

In this paper, the authors address one such factor — Statistical Power Analysis. It is believed, and will be demonstrated, that any body of research undertaken without considering statistical power as a fundamental design parameter is potentially fatally flawed. Unfortunately the authors are unaware of much Software Engineering research which takes this parameter into account. In addition to introducing Statistical Power, the paper will attempt to demonstrate the potential difficulties of applying it to the design of Software Engineering experiments and concludes with a discussion of what the authors believe is the most viable method of incorporating the evaluation of statistical power within the experimental design process.

Keywords: Statistical power; empirical evaluation

1. Introduction

Empirical research in software engineering is a difficult undertaking. Perhaps the most critical points are in the formulation of the hypothesis and the framework for evaluation of the hypothesis. Frequently this framework is based around statistical significance testing of the Neymann–Pearson type. In fact in our informal review of the software engineering subject-based empirical literature, it is difficult to find articles adopting other approaches. At first glance, this type of significance testing seems very straightforward, but if the researcher is going to derive meaningful conclusions from their work, then this technique has a number of parameters which must be carefully controlled — such as setting of significance levels, the choice of test.

One such parameter is the statistical power of the test being undertaken. Any test without sufficient statistical power is effectively meaningless, as the experiment simply does not have enough information to allow the researcher to draw any reliable conclusions using statistical significance testing. Our informal review of the software engineering empirical literature failed to find many articles which report the statistical power of the described experiment; in fact we failed to find any articles which suggested that statistical power had been considered when establishing the evaluation framework. From this one could conclude that the field is in crisis — how many of the reported works are valid? Unfortunately, any meaningful post-analysis is impossible due to the lack of details concerning statistical power and its sub-components. One can only guess at the impact this regrettable omission has had on the conducted work.

In the following sections, the importance of statistical power will be addressed. Section 2 will discuss statistical significance testing and its relationship to statistical power and section 3 details how to use and calculate the

[1] (email: james@cs.strath.ac.uk)

[2] Daly is now with the Fraunhofer Institut (IESE), Sauerwiesen 6, D-67661 Kaiserslautern, Germany (email: daly@iese.fhg.de)

PII S0950-5849(96)01139-1

statistical power for future empirical undertakings. Much of the following discussion uses examples applicable to parametric equality of means testing; for examples of other statistical procedures, see Cohen [1] or Kraemer and Thiemann [2].

2. Statistical significance testing

Significance testing of the Neyman–Pearson type is the form of rejecting or accepting a null hypothesis (denoted H_0), where the null hypothesis is stated simply for the purpose that it may be rejected allowing the researcher to accept the alternative hypothesis (denoted H_1) and conclude that an effect exists. For example, an experiment concerned with programmer productivity might have a null hypothesis

H_0: the mean programmer productivity of group A (treatment) is the same as that of group B (control)

with the alternative hypothesis stated as

H_1: the mean programmer productivity of group A (treatment) is greater than that of group B (control).

From the many articles read by the authors, it is clear that the majority of researchers within software engineering use this type of significance testing as their primary means to detect the presence of an effect within the phenomena being empirically investigated.

Statistical power analysis, an inherent part of significance testing, is defined as: the probability that a statistical test will correctly reject a false null hypothesis [3], i.e. the chance that if an effect exists it will be found.[3] For example, a power level of 0.4 means that if an experiment is run ten times, an existing effect will be discovered only four times out of the ten experimental runs. An adequate power level (that is, one at which the cost of running the experiment is deemed to be worth the chance of not detecting an existing effect) is usually quoted at 0.8, i.e., the chance you will not detect an existing effect is one in five (this is explained in more detail in Section 2.1.) Any researcher not undertaking a power analysis of their experiment has no idea of the role that luck or fate is playing with their work and consequently neither does the Software Engineering community.

In theory, integrating statistical power analysis into an experimental design is a relatively straightforward process, involving only the following three components and the required power level (unfortunately, as we will see later the evaluation of the effect size component is not straightforward):

- **The significance criterion** (α): the chosen risk of committing a Type I error, that is the probability of incorrectly rejecting the null hypothesis (H_0), when performing significance testing. The directionality of the test (the test can be directional or non-directional) is also of importance. Power can be increased at the expense of a larger probability of committing a Type I error, e.g., raising α from 0.05 to 0.10, or by using a directional statistical test (these concepts are explained in detail in Section 2.1.1.).
- **The sample size** (N): the larger the number of subjects, the smaller the error, the greater the accuracy, and therefore the higher the power of the test.
- **The effect size** (γ): the degree to which the phenomenon under study is present in the population. If all other factors are constant, then the larger the effect size, the greater the probability the effect will be detected and the null hypothesis is rejected.

The power level and these three determinants are related in such a manner that given any three values, the fourth can be easily calculated. Ideally, the researcher should estimate or anticipate the effect size, set the significance criterion, and specify the power level desired. The number of subjects needed to meet these specifications can then be derived from the appropriate statistical tables, such as the ones presented by Cohen [1].

The above is the ideal scenario, but how often is it used? Unfortunately any post-analysis of the Software Engineering literature is nearly impossible. Inadequate reporting or consideration of the experimental design implies that significant information is normally missing. How often do we see the effect size reported? This makes post-calculation of the statistical power impossible. Some researchers in other disciplines have attempted a post-analysis of their respective literature by examining each experiment at three typical power levels (small, medium and large). For example, Baroudi and Orlikowski [4] report in their study of the Management Information Systems literature for the period 1984–89, covering 57 articles, that if it is assumed that all the studies have:

- a small effect size, then 99% of the studies have inadequate power.
- a medium effect size, then 66% of the studies have inadequate power.
- a large effect size, then 34% of the studies have inadequate power.

Although it is difficult to draw any reliable conclusions from this analysis, it certainly paints a worrying picture for the MIS discipline. Any analysis of the Software Engineering literature will suffer from the same reporting problems, and hence we are left to guess at the impact of ignoring this critical design parameter on the existing Software Engineering empirical research literature. This picture is not particularly surprising as articles such as Tiller [5], Basili and Reiter [6], and Pfleeger [7] which present a general overview on experimental design and analysis, and introduce controlled experiments, fail to detail statistical power or the importance it holds prior to, and after, the running of

[3] The statistical power has also an indirect implication on the study's ability to accept the true null hypothesis.

an empirical study. Even MacDonell [8] who reviews the lack of experimental rigour in software complexity measurement, only briefly touches statistical power, stating

"they simply do not have the power to identify false hypotheses"

probably leaving the reader confused and uncertain of the importance of the issue. Researchers are, therefore, unlikely to consider their experiments' power levels which subsequently, as demonstrated by Baroudi and Orlikowski [4], are likely to be inadequate. As a consequence, articles may produce results which have a reasonable chance of being reported as:

- inconclusive, that is no significant findings were demonstrated in the study,
- incorrect due to accepting the false null hypothesis,
- of no real interest due to the effect size being particularly small, that is the phenomenon under study does not hold any real degree of importance: it has no clinical significance.

Yet these failings can be and, indeed, should be avoided. Researchers who are prepared to spend time performing empirical work should not waste their efforts by poor experimental planning or poor statistical analysis. Conventional wisdom suggests that the concept of statistical power analysis must be seriously considered.

2.1. The significance criterion (α)

It is essential that the researcher guards against the two types of errors which can occur during statistical significance testing. First, Type I error: the probability of incorrectly rejecting the null hypothesis (H_0). Second, Type II error: the probability of incorrectly accepting H_0. Essentially, a Type I error is committed when an effect is thought to have been found even though one does not exist. Conversely, a Type II error is committed when an existing effect remains undetected. The risk associated with committing a Type I error is represented by α and, similarly, a Type II error is represented by β. Furthermore, the power of a statistical test, defined as the probability that the statistical test will correctly reject the null hypothesis, is represented by $1 - \beta$.

Typically, α is set at a prudently low level of 0.05 to guard against Type I error, i.e. there is a 1 in 20 chance of incorrectly rejecting H_0. However, the β value is often ignored by researchers. If the β value is preset researchers can ensure that their statistical tests will have sufficient power to detect whether the phenomenon being examined exists. It is for this reason the β value should not be overlooked (remembering power $= 1 - \beta$). α and β, however, are not independent. Hence, with α set the value of β (and thus power) will be constrained. Setting the α at a vanishingly small level of 0.001, given an arbitrary effect size and number of subjects, may reduce the power level to 0.1 and, consequently, β error to 0.9. From this example, two points are worth mentioning:

- the power of such a test is exceedingly small and any researcher would have to think twice about their experimental plans if they calculated such numbers.
- the implication of relative seriousness of Type I to Type II error is β/α which is $0.90/0.001 = 900$ to 1, i.e., false rejection of H_0 is 900 times more serious than erroneously accepting it.

There are times when such conditions do occur, for example Baroudi and Orlikowski [4] cite a paper by Mazen et al. (p. 89) where the risk of incurring a Type II error far outweighed that for a Type I. Mazen et al. discuss the ill-fated Challenger Space Shuttle, where NASA officials had to make a choice between two assumptions

"The first assumption was that the shuttle was unsafe to fly because the performance of the O-ring used in the rocket booster was different from that used on previous missions. The second was that the shuttle was safe to fly because there would be no difference between the performance of the O-rings in this and previous missions. If the mission had been aborted and the O-ring had indeed been functional, Type I error would have been committed. Obviously the cost of the Type II error, launching with a defective O-ring, was much greater than the cost that would have been incurred with Type I error."

Perhaps a more realistic example, in terms of software engineering, is to set $\alpha = 0.05$, power $= 0.8$ thus producing a β error of 0.20, i.e., false rejection of H_0 is 4 times more serious than erroneously accepting it. Presetting of the criterion factor at this level has, according to Baroudi and Orlikowski [4], Sawyer and Ball [3], and Stevens [9], become widely accepted as the norm.

Many researchers, however, fall into the trap of setting their α value after the experiment. It is common to read the findings of statistical tests reported by researchers at the level of * ($\rho < 0.05$) as significant, ** ($\rho < 0.01$) as very significant and *** ($\rho < 0.001$) as extremely significant (see for example [10,11]). According to Slakter et al., this is "a statistical nonsense" [12]. Standard statistical procedures demand that α must be preset and not changed after the experiment has been performed.

2.1.1. Direction/non-direction of a statistical test

Another element of the significance criterion is the directionality or non-directionality (one tailed/two tailed) of the statistical test. For example, if the researcher is comparing the means from two groups of subjects, A and B say, the phenomenon under study can be defined in two ways:

- the phenomenon exists if and only if the means of A and B differ. No direction — for example, the mean of A is larger than the mean of B, is given so deviations in either direction from the null hypothesis constitute as evidence

against it. Because either tail may contribute to α this is termed a two-tailed test.

- the phenomenon exists if and only if the means of A and B differ in a direction specified in advance, for example the mean of A is larger than that of B. In this case, evidence against the null hypothesis comes from only the direction specified, hence the term one-tailed test.

When the experimental results are in the predicted direction, and all other things are equal, a non-directional two-tailed test will have less power than a directional one-tailed test. This is because, although there is a rejection area equal to $\alpha/2$ in each tail of a two-tailed test, one of these tails is meaningless in the case of predicted direction results. It is important to note, however, that this concept only holds when the sample result is in the predicted direction. It is also important to note that if the direction of the effect is different from that hypothesised, all that can be said is that the data did not support the hypothesis. In the case of Lucas and Kaplan (cited by Korson [13] p. 20) who performed a structured programming experiment, however, the hypothesis was changed after the experiment because the results were not in the predicted direction. This approach is incorrect.

2.1.2. Parametric and non-parametric tests

A parametric statistical test requires the estimation of one or more population parameters. For example, in the t and F tests the calculated within-group sample variance presents an estimate of the actual within-population variance [14]. A non-parametric test, however, does not involve such an estimation. Furthermore, a parametric test requires assumptions about the distribution curve of the population, for example, the sample of the population should be normally distributed for the t and F tests; having said this, the t test and F test are extremely robust and moderate normality deviation does not seriously influence their validity, and hence decisions about their validity of application are not straightforward. In advance of a parametric test, it is common practice to apply a normality test. Brooks [15] provides one source of tables for skewness and kurtosis to apply such tests.

The obvious advantage that non-parametric tests hold over parametric ones is they do not require the sample population to be normally distributed. This does not, however, make a non-parametric test superior. Non-parametric tests do not have the same statistical power of their counterparts. When the sample is normally distributed, the statistical power of the non-parametric test will be less than the corresponding parametric test (Power-Efficiency) and as a consequence, a Type II error is more likely to be committed. Briand [16] also warns against the inappropriate use of non-parametric statistics when conducting Software Engineering experiments and provides sample figures comparing Pearson's product moment correlation against Spearman's rank correlation. Selecting a parametric test can also be erroneous, an experiment with a small data set can only produce a normality test with relatively low power, and hence has a greater risk of inappropriately selecting a parametric test.

In conclusion, in many circumstances the choice between parametric and non-parametric analysis is not straightforward, a non-parametric test should only be used when the parametric assumptions are not met, or when it is wished to be particularly conservative on the side of Type I errors. In general, there is some loss of power, but this loss is normally small. (For a more general discussion of the relative performance, in terms of power, of parametric and non-parametric metrics, see Gibbons [17].)

2.2. The sample

The sample size, represented as N, is an important feature of an empirical study. Given the effect size and the significance criterion are constant, the power level of the test is directly dependent upon the sample size. As N increases, the probability of error decreases, thus the greater the precision and the higher the chance of rejecting the false null hypothesis, assuming that the sample is a representative cross-section of the entire population.

2.2.1. The sampling procedure

There exists doubt as to whether any sample, no matter how large, can be extrapolated to allow the conclusion 'this can be applied to the population as a whole'. Regardless of the characteristic under investigation, the software engineering field has no defined sampling frame (i.e. description of the entire population) for its practitioners, and hence we cannot know if the sample is truly representative of the underlying population. For example, Brooks [18] reports differences from 4 to 1 to 25 to 1 across experienced programmers with equivalent backgrounds, and Curtis [19] reports 25 to 1 or 30 to 1 differences in performance among programmers. These results can have a major impact on any experiment investigating programmer performance. But we have no way of knowing how representative (of the entire programmer population) the samples used in these studies were. This problem is compounded if these performance estimates are not qualified by a description of the population from which they were obtained. Careful experimental design is required to control this sampling problem. Random sampling the population for subjects is probably the best option to obtaining a cross-section of the general populace. It is, however, extremely costly both in terms of time and money. An alternative to random sampling is availability sampling [20]. The researcher conducting the study collects data from subjects who are willing to participate in it; since the decision is left to the subject, however, it is difficult to know how random or representative the sample population is. Consequently, although this type of sampling has the advantages of economy of time and money, the findings of a study using this technique are less able to generalise their results to larger populations. The widespread use of availability

sampling, within software engineering experimentation, is a major source of concern.

2.2.2. The sample size

Given appropriate sampling, as N increases the power level increases. It is therefore imperative to calculate the sample size needed for the desired power level. If this is not carried out the researcher tends to end up with a sample size of convenience, something which should be avoided if at all possible as often the power level will be inadequate. Although it is possible to increase the power level of the experiment by increasing the homogeneity of the sample, it is important to realise that what is gained in statistical power is lost in generalisability.

Once the required sample size has been calculated (see Section 2.4.) and the subjects recruited it is important to separate into groups of approximately equal numbers, N_i say. If this is not exercised, the skewed distribution of subjects results in a lower power of statistical test because a subset of subjects will contribute nothing to the study [21]. The harmonic mean:

$$N = \frac{(2 \times N_1 \times N_2)}{(N_1 + N_2)} \tag{1}$$

is used to calculate how many subjects this subset includes. For example, Baroudi and Orlikowski [4] offer the following example: a researcher has 108 subjects, distributed across two groups; one group receives training ($N = 86$ cases) and the other does not ($N = 22$ cases). The harmonic mean for this study would be 35 subjects.[4] The harmonic means should roughly equal half the total number of subjects in the study. Thus, in the above example, with a harmonic mean of 35, the study is equivalent to one with equal group size of 35 rather than 54. The skewed distribution of the subjects between groups has meant that 38 subjects have not been fully utilised. If the subjects had been divided equally into two groups, statistical power would then have been maximised for that number of subjects. The problem of not equally dividing the subjects into groups has been encountered in software engineering, for example Shneiderman et al. [22] and Sinha and Vessey [11]. Admittedly, however, it is not always possible to split subjects into even groups. In such cases it is always better to use the 'extra' subjects rather than simply to discard them. They could, for example, help identify relationships during an inductive analysis which may have otherwise gone unnoticed [23].

2.3. The effect size (γ)

The effect size is the degree to which the phenomenon under study is present in the population. Thus, the larger the effect size the greater the degree a phenomenon is likely to be detected and H_0 rejected. In comparison to the significance criterion and sample size, however, it is a poorly understood concept. Sawyer and Ball [3] found in their marketing research paper that effect size was not considered to the same extent as the concepts of sample size and significance criterion. Similarly, Baroudi and Orlikowski [4] found the same trend arising in the area of MIS. This is a cause of concern, as the effect size, as stated by Baroudi and Orlikowski, plays a critical role

"... in the determination of the power of a statistical test [which] is fundamental to adequate interpretation and application of research results."

Reporting the effect size allows other researchers in the field to judge the importance of the study's results, while at the same time allowing comparison to the findings of previous studies. Moreover this information will facilitate meta-analyses and cost-effective planning for future research in related areas [2].

Unfortunately, the effect size is not a measure easily predicted, especially if the area of research is new, or if little experimental work has been performed (as in most areas of software engineering). Usually, effect size of a phenomenon can be estimated from previous empirical results, but in the case of software engineering, due to a lack of empirical studies, the best option for a reasonable estimation is by expert judgement — this approach is explored in Section 3.2. Post power analysis of the experimental data will allow an estimation of the effect size index, γ, which can be used as the effect size index for calculating the power level of a replicated experiment. Potential researchers considering embarking upon replicating a study should note that this effect size will, in general, differ from the original experimenter's estimate and furthermore is highly error-prone, being deduced from a single source. (This problem is illustrated further in Section 3.1.) γ is expressed in the measurement unit of the dependent variable by dividing it by the standard deviation of the measures in their respective populations. For a directional test using two independent samples the formula is:

$$\gamma = \frac{\mu_B - \mu_A}{\sigma} \tag{2}$$

where μ_A and μ_B are the means of the populations, σ is the standard deviation of either population, assuming they are equal, and the alternative hypothesis is $\mu_B > \mu_A$. For a non-directional test using two independent samples the formula is:

$$\gamma = \frac{|\mu_A - \mu_B|}{\sigma} \tag{3}$$

where μ_A and μ_B are the means of the populations, σ is the standard deviation of either population, assuming they are equal, and the alternative hypothesis is $\mu_A \neq \mu_B$. If $\sigma_A \neq \sigma_B$ for Eqs. (2) and (3) then the definition of the effect size index will be slightly modified. Since there is no longer a common within-population variance, γ is defined as in Eqs. (2) and (3), but instead of σ as the denominator, the root

[4] $(2 \times 86 \times 22)/(86 + 22) = 35$.

mean square of σ_A and σ_B is required, i.e., the root mean square of the two variances (σ'):

$$\sigma' = \sqrt{\frac{(\sigma_A^2 + \sigma_B^2)}{2}} \quad (4)$$

The use of Eq. (4) produces an average within-population standard deviation. This standardises the difference between the means and the calculation of the power remains unaffected.

This process is the standard formulation for parametric equality of means tests. Formulations for non-parametric tests are relatively straightforward, but are obviously in terms of the entire population distribution rather than the parameters (means, variances) of the distribution, see Cohen [1] for other formulations. Again care must be taken to make sure the correct formulation is chosen (see Parametric and Non-parametric Tests Section 2.3.).

An alternative method requires deciding whether the effect size of the phenomenon under study is small, medium or large, as proposed by Cohen [1]. Once decided, the effect size index is set to one of the following: $\gamma = 0.2$ for a small effect, $\gamma = 0.5$ for a medium effect, and $\gamma = 0.8$ for a large effect, where the measurement is expressed in standard deviations, i.e., 0.2 of a standard deviation and so on. This common conventional frame of reference is, however, rather generalised and Cohen recommends its use only when no better basis is available for accurately estimating γ.

Accurately estimating γ is not easy, however. For a given dependent measure and a given difference between the treatment and control conditions on that measure

"the effect size will be larger or smaller depending on the relative values of the difference between the means, on the one hand, and the variance, on the other." [23]

As a result, factors which influence either the variance or the mean in relation to one another can produce a large change in the overall effect size. Fig. 1 displays such a scenario. In part (i) the variance of the population is large and the means differs by $\mu_t - \mu_c$ ($\gamma = 0.5$, a medium effect). In part (ii) the variance is much smaller, although the mean difference has remained the same. This has led to a much increased effect size ($\gamma = 1.2$, a large effect). In part (iii) the variance remains the same as in (ii), but the difference in means is smaller, in turn, producing a smaller effect ($\gamma = 0.5$). Essentially, poor estimation of either the standard deviation or difference in means will produce a poor estimation of the effect size.

Finally, it is important to try to achieve an effect size which will provide statistical significance and allow conclusions to be made that have clinical significance — practical meaningfulness. Cohen summarises these two points rather nicely,

"Small effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and

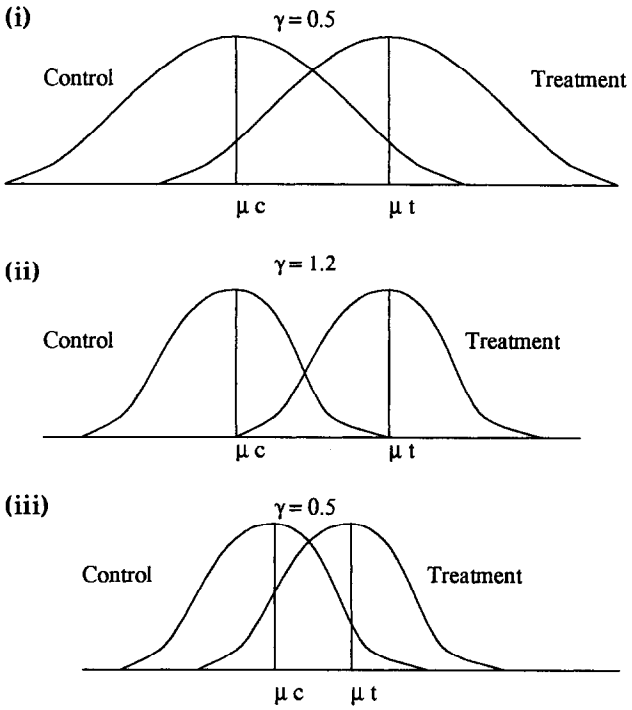

Fig. 1. Effect size is dependent on the relative magnitude of the difference between the means and the variance.

experimental bias and lack of fidelity is a bootless task, yet not so large as to make them fairly perceptible to the naked observational eye." [7]

An example is a test of significance where $\rho < 0.05$ is not necessarily less significant than a case where $\rho < 0.001$. For example, the experiment with the hypothesis that using a particular design guideline saves one month over a twelve month project produces a test of significance, $\rho < 0.05$, compared to a similar experiment and hypothesis which provides a one day reduction over the same time-scale, test of significance $\rho < 0.001$ leaves few doubts as to which result is of more significance. The first hypothesis provides clinical significance to a high degree: a substantial saving in time has been found to exist when using one particular guideline over another. In the second case, however, is there clinical significance? Hardly: nobody will be very much interested in a result which saves only one day over a twelve month period, even if it is almost certainly correct. In conclusion, it is very important to realise a small ρ value does not imply a large effect size or a strong relation [12].

2.4. Power determination

The fact that the phenomenon being empirically investigated exists, far from guarantees that a statistically significant result will be produced. Statistical power analysis is a method of increasing the probability that an effect is found in the empirical study: a high power level means a statistical test has a high probability of producing a statistically significant result. In other words, a high power means if an effect

exists, there is a high probability that it will be found; a Type II error is unlikely to be committed. Similarly, if an effect does not exist, the researcher has a solid statistical argument for accepting the null hypothesis — this is not the case, if the study has low power.

The power of the statistical test becomes a particularly important factor when H_0 is not rejected; that is the effect being tested for is not found. The lower the power of the test the less likely H_0 is accepted correctly. Consequently, when H_0 is not rejected and the statistical test is of low power, the only conclusion that can be made, because the results produced are ambiguous, is that the effect examined has not been demonstrated by the study. Studies with a high power, on the other hand, offer the advantage of an interpretation of the results when there is insignificance. There exists strong support for the decision not to reject the null hypothesis, something a low powered, statistically insignificant study cannot give.

A fact of greater concern is the inconsistency which can arise across the many existing studies due to low power. For example, in a paper by Robins [24], who was interested in determining why so much inconsistency existed in the area of clinical depression, of the 87 studies he examined, he found that only 8 of these had adequate power. The high power studies all reported a significant relationship, while the low power studies tended not to support the relationship. As a consequence, what seemed to be a large number of studies with inconsistent findings was actually a small number of studies which provided consistent, meaningful and reliable conclusions.

3. Calculating statistical power

The calculation of the power level requires the values of the significance criterion (α), the sample size (N) and the effect size (γ). Once these values are available, power can be easily calculated using δ where

$$\delta = \gamma f(N) \tag{5}$$

which combines the effect size and sample size into a single index, that can then be used, along with the α value, to obtain the power level from the appropriate tables. For the comparison of two means, the most frequently performed test in the behavioral sciences (and extremely common in subject-based Software Engineering experiments), the value of δ[5] is calculated by:

$$\delta = \gamma\sqrt{\frac{N}{2}} \tag{6}$$

where N is the harmonic mean of the two samples. Having calculated δ, set α and chosen the appropriate version of the statistical test (one- or two-tailed), the statistical power of the system can be found by simply consulting the relevant table in a suitable statistics book (for example Table H, p. 309 of [14]). As an example, if we require a power level of 0.8 and have set α to 0.05, then if we are conducting a two-tailed test we require δ to be at least 2.8 (2.5 for a one-tailed test).

Often when the power level is calculated, the estimated power is inadequate. In this case, the experimental design must be altered to increase the power, unfortunately the experimenter has only one parameter to manipulate to achieve this increase in power — the sample size.[6] Hence if the power level calculated is inadequate, the number of subjects required to meet the desired power level can be obtained from the formula:

$$N = 2\left(\frac{\delta}{\gamma}\right)^2 \tag{7}$$

where the value of δ is retrieved from the appropriate table using the desired power level and the specified α value (see for example Table I, p. 310 of [14]). This relationship is discussed in detail in the next section.

3.1. Statistical power analysis: an example

To illustrate how to calculate the power level of a statistical test, a statistical power analysis of Korson's [13] first experiment, for use in replication of this study is detailed.[7] The experiment was designed to test if a modular program used to implement information hiding, which localises changes required by a modification, is faster to modify than a non-modular but otherwise equivalent version of the same program. Korson used two groups, each with $N = 8$ cases. The test was performed at $\alpha = 0.05$ (one-tailed), and the mean of each group were as follows: $\bar{X}_A = 19.3$ with $S_A = 8.1$, and $\bar{X}_B = 85.9$ with $S_B = 47.8$.[8] First, calculate the root mean square of the two variances using Eq. (4)

$$S' = \sqrt{\frac{(8.1)^2 + (47.8)^2}{2}} = 34.28$$

Second, calculate the effect size index using Eq. (2)

$$\gamma = \frac{85.9 - 19.3}{34.28} = 1.94$$

Third, using Eq. (6), calculate

$$\delta = 1.94 \times \sqrt{\frac{8}{2}} = 3.88$$

[5] The calculation of δ changes with each statistical test, fortunately the calculation of this intermediatory value can be omitted; see the last paragraph of this section.

[6] Alternatively, the researcher could increase their α level, but in general this is a risky practice and should only be considered as a last resort.

[7] The reader should note that the following calculations are error-prone, this is discussed later in this section.

[8] In standard statistical notation, μ and σ are used when discussing the population, $\bar{X}$ and S are used when discussing a sample of the population.

which allows the power level of the test, 0.98, to be derived from the appropriate table.

It is essential to note that the above calculation doesn't form an estimation of what Korson considered to be the statistical power in his experiment. We have no way of knowing or even guessing this, unless the author reports the relevant information. The above post-analysis is not an acceptable alternative. Like a statistical test's α level, statistical power must be estimated before the experiment is carried out, post-analysis is statistical nonsense within an experiment. Any well conducted experiment can start off with an acceptable estimation of statistical power, and subsequently find no significant effect. A subsequent post-analysis of statistical power will often claim that: 'The experiment had insufficient power'. This is not the case: at the start of the experiment the power was believed to be sufficient and hence the experiment is not invalid. The lack of a significant result and resulting low post-analysis could be caused by a myriad of reasons, such as an unconsidered variable in the experimental design or a breakdown in the experimental procedure or a myriad of other causes or effects, obviously including the possibility that only a weak relationship exists between the items under investigation. Hence simply to claim that the power was inappropriate is a naive statement. So why the post-analysis? As part of an on-going body of research it was decided to replicate this experiment. (Sadly the replication of experimental work is not often carried out within the Software Engineering community, this is another major deficiency, see Brooks et al. [25] for a discussion on the benefits of replication.) In reviewing this work, the authors found a number of potential defects, which the authors believe would have influenced the results. Despite these reservations, the authors undertook the above analysis as their power estimation. The authors subsequently decided not to alter their estimation of the effect size due to the difficulty in quantifying the impact of the defects — again this illustrates the dangers of deriving an effect estimate from a single source. Note this estimation took place before our replication. Given the large power rating, the authors were happy to conduct the experiment with the same sample sizes as the original. In fact one of our groups had 9 people and the other group had 8, all other design parameters remained consistent with the original experiment. Full details of the findings can be found in [26].

Unfortunately the results of our replication differed significantly from the other experiment, hence it is difficult to draw reliable conclusions from either piece of work. Further replications are required to resolve the debate between the two experiments. An undertaking anyone replicating these experiments must attempt is an estimation of the statistical power of their replication. This new replication has evidence from two sources: the original experiment and the first replication. An obvious approach is a post-analysis of the two experiments. A post-analysis of the replication follows the same procedure as before.

Two groups were used, one with $N = 8$ cases, one with $N = 9$ cases. First, calculate the harmonic mean from Eq. (1)

$$N = \frac{2 \times 8 \times 9}{8 + 9} = 8.47$$

The test was performed as above with $\alpha = 0.05$ (one-tailed), and the means of each group were as follows: $\bar{X}_A = 48.0$ with $S_A = 25.4$, and $\bar{X}_B = 59.1$ with $S_B = 27.0$. Calculating the root mean square of the two variances using Eq. (4)

$$S' = \sqrt{\frac{(25.4)^2 + (27.0)^2}{2}} = 26.2$$

Then calculate the effect size index using Eq. (2)

$$\gamma = \frac{59.1 - 48.0}{26.2} = 0.42$$

Finally, using Eq. (6), calculate

$$\delta = 0.42 \times \sqrt{\frac{8.47}{2}} = 0.86$$

This derives a power level for the test of 0.22 from the appropriate table.

Calculating the required number of cases for the conventional power level of 0.8 using the appropriate table, it is found that for a power level of 0.8, and the test with α set at 0.05 (one-tailed), the value of $\delta = 2.49$. Now using Eq. (7), calculate the required number of subjects

$$N = 2 \times \left(\frac{2.49}{0.424}\right)^2 = 68.98$$

Hence, 69 cases are required for each group (138 cases in all).

Any group undertaking a new replication now has an enigma. What should they use as their power estimation — the post-analysis of the original, the post-analysis of the replication or some sort of merging of the two analyses. Their first thought should be to look for differences between the experiments or any indications that the experiments revealed any problems or limitations in the experimental design. In this case, the replication attempts to show that an uncontrolled parameter affected the results (an ability effect) causing the difference and calling into question the completeness of the experimental design. The group must arrive at a value judgement of upon which study to place the greater weight. We would advise caution during this process, the old adage 'safety in numbers' is a good guiding principle — it is better to overestimate the required sample size than to conduct an experiment with insufficient power. This example illustrates further the inadequacy of deriving a power estimate from a single source — multiple, even if inaccurate, sources are nearly always a safer proposition.

In fact, the above power calculations are unnecessarily detailed, for illustrative purposes. Given the effect size, the desired power, the α level, whether the test is one- or

two-tailed and a decision on which test to apply, an experimenter needs simply to consult the correct table in Cohen [1].

3.2. Evaluating effect size

Undoubtedly the most difficult component in producing a statistical power estimation is evaluating the effect size. Any researcher embarking upon an evaluation has two major categories of approaches: judgemental and normative. Currently, due to the limited number of empirical studies within Software Engineering, normative approaches are difficult to apply. Normative approaches rely on either other related empirical studies or the establishment of an empirical norm for the subject of the experimentation, see Jeffrey et al. [27] or Smith and Glass [28] for good examples of the use of normative techniques in other disciplines. The most likely use of a normative approach within Software Engineering is when conducting a replication (see above for an example), and this usage only just qualifies as a normative approach unless the experimenter is replicating a study replicated by many other researchers. Hence, the remainder of this section focuses on judgemental approaches.

Probably the most common method of evaluating the magnitude of an effect is by guesswork — researchers finding an interesting result, which they hope is large enough, decide to follow up their initial findings. Although intuitive guesses are undoubtedly superior to the uncritical acceptance of all statistically significant effects as important or the assumption that results that are not statistically significant are unimportant, it does not provide a solid foundation for reliable scientific investigation.

The judgemental approach to the estimation and evaluation of effect sizes can simply be regarded as a consensus opinion of experts within a field of experimentation. Since experts have a realistic set of expectations about what constitutes significance (within their field), one may ask them to determine the degree of impressiveness of research results and data. The difficulty of this task may be compounded within Software Engineering as the experts will often not fully understand the concepts of significance and effect size, and hence their opinion may only address these concepts in a relatively indirect manner. Hence the researcher is more likely to extract qualitative rather quantitative opinions on these topics.

Sechrest and Yeaton [29] in their excellent paper give many examples of judgemental approaches leading to successful conclusions. They also report research results showing that experts have shown a significant ability to provide accurate estimates of effect sizes into the Psychological Sciences. The authors believe that these concepts are sufficiently unknown that any attempt to replicate directly these findings for the Software Engineering community would currently fail. Hence a more indirect approach to obtaining effect size information is required, such as the use of formal structured interviewing and formal questionnaire surveys of experts to elicit their opinions on the concept under investigation. The authors have recently undertaken both approaches as a form of expert knowledge elicitation and effect size estimation.

The authors were interested in the effect of object-oriented software construction on the maintenance process. Rather than simply embarking upon an experiment, they chose to conduct a series of formal structured interviews [30], followed by a formal questionnaire survey [31] to gain insight into what the experts (i.e. practitioners) believed were the concepts displaying large effects (i.e. causing major effects in terms of positive or negative alterations to the maintenance process).

The structured interviews were carried out with 13 experienced object-oriented developers. The interviews elicited information about the perceived advantages and disadvantages of the paradigm with regard to the maintenance phase of the life-cycle. Subsequent analysis was undertaken by transcribing and summarising each interview and tabulating each subject's answer to each question. From this the authors were able to identify several hypotheses with relatively large effects. This was followed up by a questionnaire study, undertaken to increase the authors' confidence in the experts' opinions by taking a larger sample. The questionnaire survey was completed by 275 object-oriented practitioners. Again the responses were tabulated and analysed to find concepts with large effects. The authors are confident that this process has identified several concepts with large effects sizes within this domain, and would recommend this approach to anyone considering conducting experimental investigations. This initial procedure has been concluded by a series of experiments where the authors were able to show statistically a direct cause and effect relationship between object inheritance and maintainability [32].

The main difficulty with this procedure is that the investigators are gathering qualitative statements from the experts about the perceived effect size. This in turn must be translated by the investigators into a normal effect size metric using their own experience at interpreting the qualitative statements. The authors would recommend that inexperienced investigators simply translated this information into one of Cohen's [1] three categories: small (0.2), medium (0.5) or large (0.8); again investigators must show caution in the aspect of overestimating the effect size during this translation process.

Although the process is extremely inexact, the authors would urge every empirical investigator to undertake this type of process. If such a process is not undertaken, the investigator is in grave danger of producing meaningless metrics, even if they produce statistical significant results. Once an effect estimation has been produced, the statistical tests have been defined and the α level set, the investigator is ready to calculate the last piece in their power puzzle — the sample size. The following section gives a brief overview of the relationship between the effect and sample sizes.

Table 1
Sample size required for variable effect size

Sample size	Effect size (Power 0.8, α 0.05)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
One-tailed	1240	310	138	78	50	35	25	19	15	12
Two-tailed	1568	392	174	98	63	44	32	25	19	16

3.3. A quick numeric guide to sample size with respect to effect size

When any researcher is embarking upon an experiment it is vitally important that they ensure that the experiment has sufficient power. In achieving this desired state, the researcher will often have only one parameter to manipulate — the sample size. Often the directionality and type of the statistical tests, the α value and the effect size are all defined by the experiment, leaving the researcher to select a suitable sample size to ensure a meaningful experiment. Although defined by the experiment, the effect size must be estimated. Regardless of the estimation vehicle, this estimate is likely to have a certain degree of error and any researcher who places their belief in this exact number is jeopardising their experiment. Now armed with their estimation it is a simple case of referring to the relevant tables as outlined above. As an illustrative example, Table 1 shows sample size (or more accurately harmonic mean) against effect size for common effect sizes and one- and two-tailed tests. The tables assume that a power of 0.8 is required, that α level is set to 0.05 and that the experiment plans to use parametric equality of means test, and displays the required sample size rounded to the nearest digit.

4. Conclusions

This paper has discussed the importance of statistical significance testing to empirical software engineering. In particular it has focussed on one aspect of significance testing — statistical power. It has attempted to demonstrate that consideration of statistical power must be an essential component of any experimental design. Any experimenter which ignores this factor, even if they obtain a significant result, is not in a position to claim that their study will necessarily have any impact upon the real world, because the experimenter does not know if the experiment contained sufficient information to ensure the statistical tests were significant in terms of having a real, sizable impact on the concept under investigation. Producing and conducting experiments which demonstrate clinical significance is much more demanding than simply achieving a statistically significant result via some experimental procedure.

The paper has also shown that although theoretically straightforward, the calculation of statistical power is, in practice (especially with Software Engineering and other immature empirical disciplines) a difficult, inexact and error-prone process. The authors would urge researchers not to be put off by this fact. For Software Engineering empirical research to become a mature discipline, it is vital that the use of statistical power becomes standard practice. The authors concede that this places an extra burden upon empirical researchers, but as discussed by Robins [24], see Section 2.4., the alternative is a subject in disarray with researchers being unable reliably to compare experiments on the same topic or even with the same hypothesis.

The principal difficulty in calculating the statistical power of an experiment is in estimating the size of the effect under investigation. Currently we would recommend that a judgemental approach to effect size estimation is the most viable procedure into Software Engineering. But hopefully as the field matures, normative approaches can be adopted and the field can reap the benefits available to mature empirical disciplines such as the medical sciences and social psychology.

References

[1] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Academic Press, first edition, 1969.
[2] H. Kraemer and S. Thiemann, How many subjects? Sage Publications, first edition, 1987.
[3] A. Sawyer and D. Ball, Statistical power and effect size in marketing research, Journal of Marketing Research, 18(3) August 1981 275–290.
[4] J. Baroudi and W. Orlikowski, The problem of statistical power in MIS research, MIS Quarterly, 13 March 1989 87–106.
[5] D. Tiller, Experimental design and analysis, in N. Fenton, editor, Software Metrics — a Rigorous Approach, pp. 63–78, Chapman and Hall, 1991.
[6] V. Basili and R. Reiter, A controlled experiment quantitatively comparing software development approaches, IEEE Trans. on Software Engineering, SE-7(3) May 1981 299–311.
[7] S. Pfleeger, Experimental design and analysis in software engineering, Annals of Software Engineering, 1 (1995) 219–253.
[8] S. MacDonell, Rigor in software complexity measurement experimentation, Journal of Systems Software, 16(2) (1991) 141–149.
[9] J. Stevens, Power of the multivariate analysis of variance tests, Psychological Bulleting, 88(2) November 1980 728–737.
[10] V. Gibson and J. Senn, System structure and software maintenance performance, Communications of the ACM, 32(3) 1989 347–358.
[11] A. Sinha and I. Vessey, Cognitive fit: An empirical study of recursion and iteration, IEEE Trans. on Software Engineering, 18(5) (1992) 368–378.
[12] M. Slakter, Y. Wu and N. Suzuki-Slakter, *, **, and ***; statistical non-sense at the .00000 level, Nursing Research, 40(4) July/August 1991 248–249.
[13] T. Korson, An Empirical Study of the Effects of Modularity on Program Modifiability, PhD thesis, College of Business Administration, Georgia State University, 1986.
[14] J. Welkowitz, R. Ewen and J. Cohen, Introductory Statistics for the Behavioral Sciences, Academic Press, second edition, 1976.
[15] A. Brooks, D. Clarke and P. McGale, Investigating stellar variability by normality tests, Vistas in Astronomy, 38 (1994) 377–399.
[16] L. Briand, K. El Emam and S. Morasca, On the application of measurement theory in software engineering, Empirical Software Engineering, An international journal, 1(1), 1996.

[17] J. Gibbons, Nonparametric Statistical Inference, McGraw-Hill, 1971.
[18] R. Brooks, Studying programmer behavior experimentally: The problems of proper methodology. Communications of the ACM, 23(4) April 1980 207–213.
[19] B. Curtis, Measurement and experimentation in software engineering. Proc. of the IEEE, 68(9) September 1980 1144–1157.
[20] G. Keppel, W. Saufley and H. Tokunaga, Introduction to Design and Analysis, W. H. Freeman and Company, first edition, 1992.
[21] J. Medler, P. Schneider and A. Schneider, Statistical power analysis and experimental field research: Some examples from the national juvenile restitution evaluation, Evaluation Review, 5(6) (1981) 834–850.
[22] B. Shneiderman, R. Mayer, D. McKay and P. Heller, Experimental investigations of the utility of detailed flowcharts in programming, Communications of the ACM, 20(6) (1977) 373–381.
[23] M. Lipsey, DESIGN SENSITIVITY Statistical Power for Experimental Research, SAGE Publications, first edition, 1990.
[24] J.C. Robins, Attributions and depression: Why is the literature so inconsistent? Journal of Personality and Social Psychology, 54(5) (1988) 880–889.
[25] A. Brooks, J. Daly, J. Miller, M. Roper and M. Wood, Replication of experimental results in software engineering. Research report EFoCS-17-94, Dept. of Computer Science, University of Strathclyde, Glasgow, 1995.
[26] J. Daly, A. Brooks, J. Miller, M. Roper and M. Wood, Verification of results in software maintenance through external replication. In Proc. of the IEEE Int. Conf. on Software Maintenance, September 1994, pp. 50–57.
[27] R.W. Jeffrey, R.R. Wing and A.J. Stunkard, Behavioral treatment of obesity: The state of the art. Behaviour Therapy, 9 (1978) 189–199.
[28] M.L. Smith and G.V. Glass, Meta-analysis of psychotherapy outcome studies. American Psychologist, 32 (1977) 752–760.
[29] L. Sechrest and W.H. Yeaton, Empirical bases for estimating effect size, in R.F. Boruch, P.M. Wortman and D.S. Cordray, editors, Re-analyzing program evaluations: Policies and practices for secondary analysis of social and educational programs, Jossey-Boss, 1979.
[30] J. Daly, A. Brooks, J. Miller, M. Roper and M. Wood, Structured interviews on the object-oriented paradigm. Research report EFoCS-7-95, Dept. of Computer Science, University of Strathclyde, Glasgow, 1995.
[31] J. Daly, J. Miller, A. Brooks, M. Roper and M. Wood, Issues on the object-oriented paradigm: A questionnaire survey. Research report EFoCS-8-95, Dept. of Computer Science, University of Strathclyde, Glasgow, 1995.
[32] J. Daly, J. Miller, A. Brooks, M. Roper and M. Wood, The effect of inheritance on the maintainability of object-oriented software: An empirical study. In Proc. of the IEEE International Conf. on Software Maintenance, 1995, pp. 20–29.