Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SEARCHING FOR EFFECTS IN BIG DATA: WHY P-VALUES ARE NOT ADVISED AND WHAT TO USE INSTEAD

Marko A. Hofmann

ITIS
University of the Federal Armed Forces Munich
Werner-Heisenberg-Weg 39, Neubiberg 85577, Germany

ABSTRACT

p-values of null hypothesis significance testing have long been the standard and decisive measure of deductive statistics. However, for decades, top statistical methodologists have argued that focusing on *p*-values is not conducive to science, and that these tests are regularly misunderstood. The standard replacement or at least complement proposed for *p*-values by those critics are confidence intervals *and* statistical effects sizes. Regrettably, analyzing and comparing huge data sets (from data mining or simulation based data farming) with two measures is awkward. As a single-value measure of first interpretation for the scanning of Big Data this article proposes statistically secured effect sizes either based on exact, mathematically sophisticated confidence intervals for effect sizes or simplified approximations. It is further argued that simplified secured effect sizes are among the most instructive single measures of statistical interpretation completely perspicuous for the layman.

1 INTRODUCTION AND MOTIVATION

Data mining and simulation based data farming provide analysts with huge sets of data. Their paramount task is to find important effects in such Big Data. Effects that are visible by differences between data parameters like means and variances. Most of these data sets are samples of further unknown populations. Hence, all reasoning is based on inferential statistics. Deducing from samples and transferring the conclusions to populations has to be made with respect to possibility of random effects. The standard single measure for a *first assessment* of theses effects are *p*-values based on significance tests (see item 3 in the enumeration of section 2 for an exact definition of *p*-values). Table 1 exemplifies the situation for four deliberately chosen pairs ($i = 1,2$) of samples (represented by the sample mean $\bar{x}_i$, standard variation s_i and size n_i).

Table 1: Four exemplary data sets ready for interpretation based on *p*-values from adequate t-tests

No.	$\bar{x}_1$	s_1	n_1	$\bar{x}_2$	s_2	n_2	*p*-value
1	495	4	50	500	5	50	< 0.0001
2	495	4	10^3	500	10	25	0.0199
3	495	100	10^5	496	100	10^5	0.0253
4	495	4	10	500	6.5	10	0.0529

The data is based on the assumption that $(X_1, X_2, \ldots, X_{n_1}) \sim N(\mu_1, \sigma_1^2)$ and $(Y_1, Y_2, \ldots, Y_{n_2}) \sim N(\mu_2, \sigma_2^2)$ are in each case two independent samples of sizes n_1 and n_2 from two approximately normally distributed populations with the sample means $\bar{x}_1$ and $\bar{x}_2$, and the sample variances s_1^2 and s_2^2 . Our main interest is whether $\delta := |\bar{x}_1 - \bar{x}_2|$, the difference between the means of both samples, is practically important and unlikely a coincidence. What can be concluded from table 1?

- The first sample pair is an unproblematic "textbook" case. Both samples have the standard size considered to be large enough for the majority of methods in inferential statistics. Variances are similar, and small with respect to the mean difference $\delta = 5$. Consequently, the p-value is very low.
- The second example is characterized by a huge difference in sample size, and a considerable difference in sample variance, but the p-value is lower than the common .05-threshold. Such an imbalance in sample size is not uncommon when the large sample stems from a closed simulation system, and the small sample is the scarce empirical data available. It is therefore frequent in all kinds of simulation-based analysis for which we use a small given real data set as reference for model validation, and further investigate new, slightly changed options by simulation.
- The third case might stem from a trivial queuing simulation. It shows that significant results ($p = .0253$) are possible even for tiny effects (negligible mean difference (1) in light of the standard deviation 100) if sample size is huge.
- The fourth case is contrary to the third: Here we have strong effect (almost identical to the first sample, s_2 is slightly greater) but due to the small sample size a non-significant p-value.

A statistical interpretation solely grounded on the p-values of null hypothesis significance testing (NHST) (assuming $\alpha = .05$) accepts the first three data sets as evidence of significant results and rejects the last case. Is this the only possible reasoning? What are the alternatives if we are looking for a single value or, at the most, a single interval? Is the interpretation of these numbers "fool-prove" for the cursory first inspection of Big Data (imagine 1000 rows in table 1)?

In answering these questions, the paper first tries to convince the simulation practitioner that p-values are suboptimal for *condensing statistical data into a single decisive measure*. Second, for this specific purpose, the advantages and drawbacks of confidence intervals (CI) and effect sizes (ES) are discussed. The article then demonstrates why, with respect to getting a first idea of effects in Big Data, confidence intervals for effect sizes (CI^{ES}) are superior to ordinary CI and ES. It is further argued that mathematically precise CI^{ES} are impractical for scanning under heteroscedasticity ($\sigma_1 \neq \sigma_2$) due to ambiguity. Finally, an approximation is presented that secures effect sizes based on elementary statistics. It can serve as a pragmatic remedy against the often difficult selection of the right variance under heteroscedasticity. Additionally, it is completely perspicuous even for the layman (here defined as someone who had only an introductory course in statistics).

2 DEFICITS OF SIGNIFICANCE TESTS

For decades NHST has been the standard procedure of deductive statistics. However, the use of significance testing in the analysis of research data has been discredited, both logically and conceptually, from numerous top statisticians – continuously for almost 100 years (Boring 1919, Berkson 1938, Fisher 1955, Bakan 1966, Greenwald 1975, Carver 1978, Rosnow and Rosenthal 1989, Tukey 1991, Cohen 1994, Sedlmeier 1996, Schmidt and Hunter 1997, Falk 1998). Due to methodological inertia for decades with little effect besides a recommendation of the American Psychological Associations (APA) Task Force on Statistical Inference (Wilkinson 1999) to include confidence intervals and effects sizes in any statistical summary. In recent years, however, things have started to change. Landmark publications from Gigerenzer (2004), Ioannidis (2005), Hubbard and Armstrong (2006), Armstrong (2007), Hubbard and Lindsay (2008) and Lambdin (2012) had sufficient impact to question the paramount importance of NHST in social science, medicine and psychology. It is beyond the scope of this paper to discuss all these criticism, and what has been said in support of significance tests (Mulaik, Raju, and Harshman 1997, Hagen 1997, Senn 2001), but the simulation practitioner should know at least the following major problems:

1. The central assumption of NHST, that there is absolutely no difference between the means of two groups ($\mu_1 = \mu_2$) is basically a straw man. In almost all research questions groups will have different

means if we measure precisely (Tukey (1991), 100): "They are always different for some decimal place"). The crucial question is whether this difference is practically important or not.

2. NHST applies the *modus tollens* on probabilistic statements. Cohen (1994) demonstrated that this is logically wrong.
3. NHST is based on a lot of critical assumptions. p-values are therefore extremely easy to misinterpret for non-experts (see especially Schmidt and Hunter (1997) and (Lambdin 2012). A p-value, to specify two of the most popular errors, is neither the probability that the results obtained occurred due to chance nor the probability of the null hypothesis given the data. It is only the probability to get the observed data (x) or more extreme results, assuming the null hypothesis H_0 is true in the population (precisely: $p := Pr(T(X) > T(x)|H_0)$; where $T(X)$ is the specific test statistic (Lambdin 2012)).
4. Since even the smallest effect is significant above a certain sample size (see data set 3 from table 1), and increasing sample size with simulations is trivial, NHST is next to meaningless for the comparison of two simulated data sets (White et al. 2014, Troitzsch 2014).
5. On the other hand significance tests are almost useless when small samples are involved, since only huge effects in small samples (visible at first sight) generate p-values which cause the null hypothesis to be rejected. All other effects lead to non-rejection (data set 4), which does not provide evidence for the null hypothesis. Logically, one should perform equivalence tests ($H_0 \sim |\mu_2 - \mu_1| \geq \Delta$) in such situations (Parkhurst 2001). Unfortunately, only experts seem to know about them, although equivalence tests are conceptually of equal importance to science than significance testing and often easy to perform without further effort (Neuhaeuser 2010).
6. Due to the rather arbitrary 5% rule of rejection, and the tendency of journals to publish significant results only, NHST have caused a serious "file drawer" problem (Rosenthal 1979, Fanelli 2012): Non significant studies do not get published, although each of them is an important contribution to science. This disregard is more critical than it might seem at first sight, since a series of non-published, non-significant studies could even uncover a strong effect in meta-analysis (Thompson 2007, 429).
7. Stand alone, significance test put too much emphasis on the probability of a random effect, without telling anything about the magnitude of the effect. The magnitude of an effect is, however, at least in general, the very first thing one would like to know about.
8. NHST mislead to rigid mechanical reasoning (p-value $< \alpha$). Although refutation or corroboration are indeed the central goals of science, they are not apodictic verdicts based on binary values. Magnitude and likelihood of effects have to be assessed in context of domain, newness and impact. (By the way, the famous $p < \alpha$-rule combines Fisher's evidential statistic (p-value) and Neyman-Pearson's error estimate (α) in a way not justified by either theory (Hubbard 2004).)

In a nutshell, no simulation result should ever be summarized by p-values only, or by p-values as the essential outcome. The standard replacement or at least complement proposed for NHST by most of the critics mentioned above are confidence intervals (CI) (for specific arguments see Brandstaetter (1999) and Poole (2001)) and statistical effects sizes (ES) (Harris 1991, Thompson 2008). Are they the ideal solution for condensing as much instructive information about data sets as possible into single measures?

3 CONFIDENCE INTERVALS AND EFFECT SIZES

Table 2 is a result summary for the exemplary data sets from table 1 based on confidence intervals and effect sizes. For technical explanations, symbols, and issues of interpretation see the following subsections.

Is this the ideal aggregation for interpretation without NHST? Even from a first glance, it is clear, that there is much more information assembled in this table than in a single p-value. However, the question here is not how to give maximum information (presumably by raw data put into graphs (Sedlmeier 1996)) but whether this information is absolutely necessary for a first interpretation. One reason for the reluctance

Table 2: Summarizing data without using NHST: CI and ES

No.	F-test	δ	$CI_{0.05}^{Student}$	$CI_{0.05}^{Welch}$	g	Δ_{s_1}	Δ_{s_2}
1	0.062	5	[3.203; 6.797]	[3.202; 6.798]	1.1	1.25	1
2	< 0.0001	5	[3.32; 6.68]	[0.86; 9.14]	1.18	1.25	0.5
3	0.5	1	[0.1235; 1.8765]	[0.1235; 1.8765]	0.01	0.01	0.01
4	0.082	5	$[-0.07; 10.07]$	$[-0.18; 10.18]$	0.89	1.25	0.77

to give up *p*-values is their succinctness. Hence, we should first clarify which additional information CI and ES provide, and second, search for the most concise representation.

3.1 Confidence intervals

Mathematically, ordinary confidence intervals (CI) are based on the same mathematics as significance tests. Their major advantage is that they comprehend more information. A confidence interval

1. cannot only answer the question (like NHST) how likely the results are erroneously caused by *randomness* (the α-based interval includes zero),
2. it also indicates the *absolute magnitude* of the effect, since the CI mean is exactly the difference $\delta := |\bar{x}_1 - \bar{x}_2|$, and
3. indicates the *precision* of the measurement via the quotient (=: ϑ) between the confidence range and δ.

The CI explained in most textbooks are based on the *S*tudent's t-test which presupposes equal variances (in the logic of NHST: F-test $> .05$). A corresponding CI for $\delta^* := |\mu_1 - \mu_2|$ is calculated by:

$$CI_{\alpha}^{\delta_S^*} := [(\bar{x}_1 - \bar{x}_2) \pm t \cdot s_0 \cdot \varphi] \ , \qquad t := t_{n_1+n_2-2;1-\frac{\alpha}{2}} \ , \qquad \varphi := \sqrt{\frac{n_1+n_2}{n_1 n_2}} \tag{1}$$

In case of different variances ($\sigma_1^2 \neq \sigma_2^2$), also called heteroscedasticity (Greek "hetero" ('different') and "skedasis" ('dispersion')), which is obvious for data set 2, the CI has to be based on the *W*elch test (which has also been used for data set 2 in table 1):

$$CI_{\alpha}^{\delta_W^*} := \left[(\bar{x}_1 - \bar{x}_2) \pm t' \cdot s'\right] \ , \quad s' = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \ , \quad t' = t_{\nu;1-\frac{\alpha}{2}} \ , \quad \nu = \frac{\left(\frac{1}{n_1} + \frac{u}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{u^2}{n_2^2(n_2-1)}} \ , \quad u = \frac{s_2^2}{s_1^2} \tag{2}$$

The standard interpretation of the CI in table 2 is:

- In data set 1 the confidence interval $CI_{\alpha}^{\delta_S^*}$ (Student's since F-test's *p-value* $> .05$) is far from including zero (effect very unlikely caused by *randomness*), and with respect to the *absolute magnitude* $\delta = 5$ the interval is relatively *precise* (confidence range $= 3.594$; $\Rightarrow \vartheta = \frac{3.594}{5} = 0.719 < 1$).
- In data set 2 $CI_{\alpha}^{\delta_W^*}$ (F-test's *p-value* $< .05$) does not include zero (:= "uncritical"), but *precision* is low with respect to the *absolute magnitude* $\delta = 5$ (confidence range $= 8.28$; $\Rightarrow \vartheta = \frac{8.28}{5} = 1.656 > 1$).
- In data set 3 $CI_{\alpha}^{\delta_S^*}$ (F-test's *p-value* $> .05$) is uncritical, the *precision* (confidence range $= 1.753$; $\Rightarrow \vartheta = 1.753 > 1$) is low with respect to the *effect's magnitude* $\delta = 1$.
- In data set 4 $CI_{\alpha}^{\delta_S^*}$ (F-test's *p-value* $> .05$) is critical, the *precision* (confidence range $= 10.14$; $\Rightarrow \vartheta = 2.028 > 2$) is very low with respect to the effect's *absolute magnitude* $\delta = 5$.

The interpretation of confidence intervals seems easy, their content is substantially stronger than the message of *p*-values. They have been proposed as complete replacements for significance tests (Brandstaetter

1999). However, many users misinterpret CI by concluding that they can be 95% confident that the calculated interval captures the estimated population parameter. This is nonsense since any concrete CI either captures the "true" parameter or not (Thompson 2007). The correct interpretation is: If we drew infinitely many random samples, 95% of them would capture the parameter, and 5% would not. Unfortunately, we do not know which case we face using a single CI (Note that this reasoning is based on a frequentist interpretation of probability.). A certain disadvantage of CI is that they are equally dependent from basic assumptions (normality, homoscedasticity ($\sigma_1^2 = \sigma_2^2$), similar sample size) as significance tests. They can be misused for mechanical reasoning, too, since a critical CI (inducing zero) is logically equivalent to (p-value $< \alpha$). Based on CI only, the same decisions are taken for our four exemplary data sets as with NHST: Only the last data set has a critical CI. (Using CI as simple replacements for NHST is NOT recommended: CI should be used to compare the results of different studies in meta-analytic research (Osborne 2008, 258), and as measures of precision (Poole 2001)). A minor theoretical but practically important disadvantage is that CIs are measures of the *absolute* magnitude of an effect. Since they are not normalized, they cannot be compared directly across different research questions. In addition, a tiny effect like in data set 3 ($\delta = 1$ while $\sigma^2 \approx 10^4$) is not directly visible from the CI – a drawback that leads us to statistical effect sizes, which are recommended in addition to CI by most researches.

3.2 Effect sizes

Effect sizes (ES) are the measures of first choice for the estimation of the *relative magnitude* of a difference in samples (Harris (1991), Thompson (2008), Ellis (2010)). In general, effect size statistics characterize the extent to which sample results diverge from the expectations specified in the null hypothesis. There are 40+ effects sizes (Kirk 1996) with various degrees of mathematical complexity, but effect sizes for mean comparison are almost trivial: one calculates the difference between two sample means ($\delta = |\bar{x}_1 - \bar{x}_2|$) and divides the result by a sample standard deviation (s) – in order to relate δ to a measure of variance. The critical question is which standard deviation should be used. When he originally developed the concept of effect sizes, Cohen (1962) was unsure about how to solve this problem uniquely for all cases. Ideally, we would like to compute $D = \frac{\mu_1 - \mu_2}{\sigma}$, assuming $\sigma_1 = \sigma_2 = \sigma$. Replacing the expectancy values μ_i with the samples means $\bar{x}_i$, or medians (Thompson 2007, 425) for skewed distributions, is straightforward, but the calculation of an appropriate estimator for σ necessitates some choice. If both samples are of comparable reliability, importance, and size Cohen's initial recommendation ("Cohen's d") was obvious (see equation 3). For two samples of substantially different size the pooled variance $s_0 = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ can be used to extend this idea and calculate (d_{s_0}). Hedges (1981) could even find an unbiased and consistent estimator ("Hedges g"), which is especially recommended for small samples (for $n := n_1 + n_2 < 16$ the correction factor between d an g is lower than .95). Note, however, that all three measures, to some varying degree, fail to represent the variability in either group, if s_1^2 and s_2^2 differ substantially.

$$d := \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{(s_1^2 + s_2^2)/2}}; \qquad d_{s_0} := \frac{|\bar{x}_1 - \bar{x}_2|}{s_0}; \qquad g := \left(1 - \frac{3}{4(n_1+n_2)-9}\right) \cdot d_{s_0} \tag{3}$$

If one sample (a huge "control group", for example) seems much more reliable than the other, Glass, McGaw, and Smith (1981) recommended to ignore the second variance, and compute Δ (an index is used to indicate that Δ is computed with s_1 or s_2):

$$\Delta_{s_1} := \frac{|\bar{x}_1 - \bar{x}_2|}{s_1} = \frac{\delta}{s_1}; \qquad \Delta_{s_2} := \frac{|\bar{x}_1 - \bar{x}_2|}{s_2} = \frac{\delta}{s_2} \tag{4}$$

The logic here is that the standard deviation of one of the samples s_1 or s_2 is taken to be more adequate for the calculation of the effect size. The strength of this assumption is directly proportional to the size of the group whose variance is selected (Ellis 2010). However, both variants are correct, they are "two distinct features of finding which cannot be expressed by one number (Glass, McGaw, and Smith 1981,

107)". The same is true for Hedges g. Thus, in general, there are at least three different ES one has to have a look at: g (or d), Δ_{s_1} and Δ_{s_2}.

Effect sizes are closely connected to NHST via the general qualitative relation (Sedlmeier 1996):

$$\text{significance} \approx \text{effect size} \times \text{sample size}, \quad (\text{e.g.} \quad t = d_{s_0} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \text{ t = Student's t value}) \qquad (5)$$

This relation immediately reveals why in NHST tiny effects can be compensated by huge sample sizes.

Cohen (1988) has given a general rule for the *interpretation of effect sizes* like d, g, Δ_{s_1} and Δ_{s_2}. According to them, below 0.2 any effect size is negligible, above 0.2 and below 0.5 effects are small, from 0.5 to 0.8 medium, and above 0.8 great. These thresholds can also be applied on the best known measure of effect size, Pearson's correlation coefficient r, via the transformation $d_{s_0} = \frac{2r}{\sqrt{1-r^2}}$. Using this rule we get a change in interpretation. With regard to ES it is sample 3 which is critical because all its effect sizes are far below the level of practical significance. All other values indicate strong or at least medium effects (see table 2).

Unfortunately, Cohen's rule has been criticized as being far too simplistic from various authors (e.g. Rosnow and Rosenthal 1989, Thompson 2007, Ellis 2010). The main problem is that effects have to be interpreted in context with the current body of knowledge in specific domains. Any completely new effect should initially be considered important, and some domains are notorious for exhibiting only small effects. However, for a first analysis on Big Data, and the search for strong effects these thresholds are quite useful. Less mechanical and more context sensitive interpretations should follow if one is also interested in small and negligible effects (Prentice and Miller 1992). The interpretation is rendered even more difficult by the existence of three mathematically equally justified versions for ES. The choice of a single effect size is critical under heteroscedasticity: The ratio between s_1 and s_2 is directly reflected by the ratio between the effect sizes Δ_{s_1} and Δ_{s_2} with g lying between them (see data set 2). Substantially differing variances can therefore be the most important source of effect size uncertainty for many practical applications. Respecting all these caveats what should be recommended for a first glance on Big Data? All three ES and the CI? Is there no simpler solution? Before we are going to address these questions in section 5, we have to correct an error that highlights how subtle interpretation of statistical data gets whenever basic assumptions are neglected.

4 CORRECTING A SUBTLE ERROR

The decisive attribute of data set 2 is the huge difference in sample size in connection with a significant difference in sample variance (indicating heteroscedasticity: $\sigma_1 \neq \sigma_2$). . Obviously, the difference between *sample variance* s_1 and s_2 in the example is quite large to assume $\sigma_1 = \sigma_2$ in the populations. This assumption is indeed critical. It would have wrongly led us to the ordinary Student's t-test where the unequal variance Welch-t-test is necessary. In the example the difference between the ordinary Students's t-test and the Welch-t-test is tremendous: Student-p-value: 7.64×10^{-9}; Welch-p-value 0.0199. Note that, although more powerful than the t-tests, the Mann-Whitney U test (equivalent to the Wilcoxon rank sum test) is not recommended under heteroscedasticity, see Kasuya (2001). Regrettably, heteroscedasticity is still a practical nuisance: It might seem reasonable, if not compulsory, to first test whether the variances are different, and then choose the Student or Welch t-test accordingly. This is exactly what we have done in table 2 and the following reasoning about appropriate CIs. In fact, this is even a current textbook recommendation (e.g. Pezzullo (2013)) for the corresponding tests. Unfortunately, it is a bad solution (Moser and Stevens 1992, Hayes and Cai 2007). If you use the F-test first to compare variances in order to decide which test or CI to use (Student or Welch), you will have increased your risk of a Type I error. As a remedy for this calamity it has been proposed to use the Welch t-test as a general solution (Ruxton 2006). You lose power when the standard deviations are, in fact, similar but gain power in the cases where they are not. We can simply apply this recommendation by solely using the CI based on the Welch test. This is a robust and simple solution for the problem, which, fortunately, does not change our interpretation

of table 2, since $CI_{\alpha}^{\delta_S^*}$ does not differ much from $CI_{\alpha}^{\delta_W^*}$ for the data sets 1, 3 and 4 for which the change is necessary. Is there a similar robust solution for unifying the information of CI and ES?

5 CONFIDENCE INTERVALS FOR EFFECT SIZES

Calculating effect sizes as point values can only lead to estimators indicating how important differences between means (or medians) are in light of sample variances. They ignore the likelihood of a random result, especially in small samples. That is the reason why confidence intervals (or statistical tests) have to be calculated in addition. Unfortunately, there are three different ES, and only one CI which have to be related to each other by the analyst. Consequently, with regard to a synoptic view on the *relative magnitude*, *likelihood* and *precision* of effects it seems more pertinent to directly calculate confidence intervals for effect sizes CI_{α}^{ES}. They contain all three types information in measurements adapted to the specific ES. Table 3 demonstrates the advantage of CI_{α}^{ES} for the exemplary data sets. It shows the range of a confidence interval for g in comparison to the display of $CI_{\alpha}^{\delta_W^*}$ and the ordinary ES.

Table 3: Summarizing data with CI for ES

No.	$CI_{0.05}^{Welch}$	**CI for ES** $CI_{0.05}^{g}$	**g**	Δ_{s_1}	Δ_{s_2}
1	[3.20; 6.80]	[0.68; 1.52]	1.1	1.25	1
2	[0.86; 9.14]	[0.78; 1.58]	1.18	1.25	0.5
3	[0.12; 1.88]	[0.001; 0.019]	0.01	0.01	0.01
4	[−0.18; 10.18]	[−0.03; 1.81]	0.89	1.25	0.77

With respect to information density and convenience of interpretation confidence intervals for effect sizes are obviously superior to the other measures:

- CI_{α}^{ES} contains information from ordinary CI as well as from ES.
- They are inherently attributed to a specific ES (g, Δ_{s_1} or Δ_{s_2}).
- They are measures of *likelihood*, *relative magnitude*, and *precision* of an effect in a *single interval*.
- Interpretation of CI_{α}^{ES} is technically as simple as the interpretation of ordinary CI.
- One could even transfer the (not recommended) rigid mechanical reasoning mentioned for ordinary CIs to CI_{α}^{ES}, and reject respectively not reject null hypotheses on the basis of critical CI_{α}^{ES} (see data set 4).

Calculating CI_{α}^{ES} is, however, sophisticated. Whereas formulas can be used for NHST; ordinary CI and ES, most CI_{α}^{ES} have to be calculated by computer-intensive iteration procedures (Cumming and Finch 2001). The CI_{α}^{ES} in table 3 have been calculated using an *EffectSizeCalculator* available for free (http://www.cem.org/effect-size-calculator). This calculator is a nice example of the problems with mechanical thinking based on modern computer-supported statistics. It is based on normality, homoscedasticity, and similar sample size. It is therefore incorrect to use it for data set 2. In addition, data sets 2 and 4 illustrate that the two other ES (Δ_{s_1} and Δ_{s_2}) differ substantially from g as would do the corresponding CI_{α}^{ES}. Hence, if we use CI_{α}^{ES} we still have to face the problem of three measures instead of a single one.

The inexactness of the CI_{α}^{ES} used for data set 2 is a solvable problem. Mathematically satisfying "robust" solutions for CI_{α}^{g}, and $CI_{\alpha}^{\Delta_{s_i}}$ have been first described by Algina, Keselman, and Penfield (2006) and are extended in (Keselman, Algina, Lix, Wilcox, and Deering 2008). Their solutions are based on trimmed means, winsorized variances, and bootstrapping methods. Using such advanced techniques they could create robust CI_{α}^{ES} for heteroscedasticity, non-normality in general, and skewed and heavy-tailed distributions in particular. They are here considered to be the currently favored expert's solution for the problem of condensing as much mathematically exact information into single telling measures. However, from a practitioners point of view there are some disadvantages:

1. CI_α^{ES} are advanced statistics. It takes time to fully understand the mathematics behind them. Most users will probably apply them without deep comprehension, trusting in the robustness of the method. Unfortunately, even for the trimmed and winsorized CI_α^{ES} there will be instances (data sets) with bad Type I error control. The lecturer in me would therefore prefer an elementary solution.
2. On the other hand, the preoccupation with exactness (of the confidence intervals) is in an important sense misleading. Three equally justified ES (based on means) which differ tremendously under heteroscedasticity imply that mathematical exactness is often illusory for practical interpretation.
3. In an exact CI_α^{ES} effects are always seen in light of a single selected variance (for standardization) and based on the uncertainty of both variances which are pooled. When sample size differs strongly under heteroscedasticity a different approach seems equally sensible: Uncertainty is assumed to be caused solely by the small sample variance whereas the importance of the mean difference is judged by the large sample variance. Hence, one can even construct a fourth sensible CI for ES in such a case (see next section).

Does all this mean that we have to give up the idea of a general single measure which is both instructive and perspicuous even for laymen? Under homoscedasticity an exact CI_α^{ES} for g is presumably the solution we strive for: We get a single measure containing information about likelihood, relative magnitude and precision of an effect which is easy to interpret. We can also use either $CI_\alpha^{\Delta_{s_1}}$ or $CI_\alpha^{\Delta_{s_2}}$ if we know which variance is crucial. But what can we do to construct a single interval under heteroscedasticity and uncertainty about the correct variance to scale the effect? While scanning through Big Data this last case will be the rule not the exception. A very crude approach would be to use the extrema of $CI_\alpha^{\Delta_{s_1}}$ and $CI_\alpha^{\Delta_{s_2}}$ to construct a new broader confidence interval. Regrettably, this bricolage lets the error rate α completely uncontrolled. Is there another simple solution?

6 SECURING EFFECT SIZES

In order to get an idea of such an solution, we first construct the additional CI_α^{ES} mentioned in the last enumeration (no. 3). We can use data set 2 ($\bar{x}_1 = 495$, $s_1 = 4$, $n_1 = 1000$; $\bar{x}_2 = 500$, $s_2 = 10$, $n_2 = 25$) to recalculate the effect size under the assumptions that the second sample was too favorable by chance, and that the first sample is completely unbiased, i.e. $\bar{x}_1 = \mu_1$ and $\sigma_1 = s_1$. We thereby construct a effect size which is secured with respect to the uncertainty introduced by the small sample. The rationale behind this approach is, that in an extremely large sample ($n \geq 10^3$) $\bar{x}$ and s are, in general, very good estimators for μ and σ. We will not make a big mistake equalizing them.

We simply calculate how low μ_2 might be, given only the parameters $\bar{x}_2$, n_2, and s_2 of our second sample, and a preselected α-level of .05. The corresponding value, calculated using a Student's-t-distribution with 24 degrees of freedom, is 496.58 (see figure 1). This means that there is a 5%-chance that if μ_2 is actually as low as 496.58, we get a sample mean $\bar{x}_2^{(n=25)} \geq 500$. We thereby "secure" our estimation of μ_2 against the randomness of the small sample. With our additional assumption of an unbiased first sample we get: $\delta_{0.05}(secured) = \delta_{0.05}(s) = |495 - 496.58| = 1.58$. For this scaling we prefer the standard deviation of the first sample $\sigma_1 = s_1$, since this value is (here) assumed to be an unbiased estimator of the more pertinent variance.

$$\varepsilon_{0.05}^{s_2-secured;\, s_1-scaled} = \varepsilon_{0.05}^{s_2-s_1} = \frac{\delta_{0.05}(s)}{s_1} = \frac{1.58}{4} = 0.39$$

The effect size we get by this calculation is *secured* with respect to the randomness of the small sample and *scaled* with respect to the presumably very "realistic" variance of the large sample. SES are lower "bounds" for ES. An upper "bound" is simply the extension into the other direction. It is therefore called extended effect size (EES) for matters of distinction.

$$\varepsilon_{0.05}^{s_2-extended;\, s_1-scaled} = \varepsilon_{0.05}^{e_2-s_1} = \frac{\delta_{0.05}(e)}{s_1} = \frac{8.42}{4} = 2.11$$

Figure 1: The rationale behind an secured effect size.

Instead of a mathematically exact CI_{α}^{ES} we can therefore use the simple approximation $[SES, EES]_{\alpha}^{\Delta_{s_1}} = [\varepsilon_{\alpha}^{s_2-s_1}; \varepsilon_{\alpha}^{e_2-s_1}]$.

We can summarize the steps we have taken so far by the following formula:

$$\varepsilon_{\alpha}^{s_2-secured;\, s_1-scaled} = \varepsilon_{\alpha}^{s_2-s_1} := \frac{|\bar{x}_2 - \mu_1|}{s_1} - \frac{s_2 \cdot t_{n_2-1;1-\alpha}}{s_1 \cdot \sqrt{n_2}} = \frac{|\bar{x}_2 - \bar{x}_1|}{s_1} - \frac{s_2 \cdot t_{n_2-1;1-\alpha}}{s_1 \cdot \sqrt{n_2}}. \tag{6}$$

Mathematically, we could also calculate the corresponding secured effect size with the first sample as starting point:

$$\varepsilon_{\alpha}^{s_1-secured;\, s_2-scaled} = \varepsilon_{\alpha}^{s_1-s_2} := \frac{|\bar{x}_1 - \mu_2|}{s_2} - \frac{s_1 \cdot t_{n_1-1;1-\alpha}}{s_2 \cdot \sqrt{n_1}} = \frac{|\bar{x}_1 - \bar{x}_2|}{s_2} - \frac{s_1 \cdot t_{n_1-1;1-\alpha}}{s_2 \cdot \sqrt{n_1}}. \tag{7}$$

In that case we assume that the second sample is unbiased, i.e. $\bar{x}_2 = \mu_2$. Using the corresponding EES one gets an approximate interval $[SES, EES]_{\alpha}^{\Delta_{s_2}}$ around Δ_{s_2}.

It seems as if we have made a step backwards. We still have two intervals, and now only two approximations. But we can use the basic idea to do something new: In order to generate a simple interval representing all the information we have we calculate 2α-secured bounds for both samples individually and scale them with the pooled variance s_0. This idea leads to (assuming for simplicity that $\bar{x}_2 \geq \bar{x}_1$)):

$$\varepsilon_{\alpha}^{s_1\&s_2-secured;\, s_0-scaled} = \varepsilon_{\alpha}^{s_1\&s_2-s_0} := \frac{\left(\bar{x}_2 - \frac{s_2}{\sqrt{n_2}} \cdot t_{(n_2-1;1-2\alpha)}\right) - \left(\bar{x}_1 + \frac{s_1}{\sqrt{n_1}} \cdot t_{(n_1-1;1-2\alpha)}\right)}{s_0} \tag{8}$$

Here the sample means are secured with the corresponding standard variations. The resulting difference is standardized by using the pooled variance s_0. Using the corresponding EES one gets a broad interval $[SES; EES]^{d_{s_0}}$ around Cohen's d_{s_0} (one could apply the correction factor for Hedges g, but it is almost 1 ($> .985$) for $n = n_1 + n_2 > 50$) that includes the specific uncertainty of both samples. Since SES and EES are mixtures of the variances of both samples they cannot be interpreted as exact CI_{α}^{ES}. They should be taken as first estimations of the likelihood and precision of effect sizes under the special assumptions made above. However, essentially, they communicate the same message as "exact" CI_{α}^{ES}.

6.1 Reasoning with secured and extended effect sizes

For the special purpose of getting a first idea about effects in Big Data $[SES;EES]^{d_{s_0}}$ are much more instructive than p-values (see table 4):

Table 4: Summarizing data with single measures: p-values versus secured and extended effect sizes

No.	$(\bar{x}_1;\ s_1;\ n_1)$ - $(\bar{x}_2;\ s_2;\ n_2)$	**p-value**	**[SES; EES]**$^{d_{s_0}}$	$CI_{0.05}^{ES=g}$
1	$(495;4;50)-(500;5;50)$	<0.0001	$[0.63;1.58]$	$[0.68;1.52]$
2	$(495;4;10^3)-(500;10;25)$	0.0199	$[0.33;1.84]$	$[0.78;1.58]$ (?)
3	$(495;100;10^5)-(496;5;10^5)$	0.0253	$[0.001;0.018]$	$[0.001;0.019]$
4	$(495;4;10)-(500;6.5;10)$	0.0529	$[0.08;1.77]$	$[-0.03;1.81]$

The $[SES;EES]^{d_{s_0}}$ in table 4 confirm that $CI_{0.05}^{ES=g}$ has been erroneously applied on data set 2. The approximate intervals are relatively close to the exact values for the uncritical cases 1 and 3, and only slightly too narrow for data set 4 (the shift to the right (greater effect) is here caused by the differences between d_{s_0} and g for $n=20$). For practical work I recommend to extend the intervals by calculating α-secured bounds for both samples. Tests with hundreds of data sets have corroborated the pertinence of such $[SES;EES]^{d_{s_0}}$ under normality. $[SES;EES]^{d_{s_0}}$ is a very convenient measure while searching for statistically uncritical strong effects in Big Data. Effects that would hit you right between the eyes if only you could see them in all that data. In that sense, the lower bound SES serve as triggers for what has been called the ultimate "interocular traumatic test" in a classic article from Edwards, Lindman, and Savage (1963). Triggers that cannot be as easily misunderstood as other single measures. From the four exemplary data sets only set 1 hints at such an effect, data set 3 has absolutely no potential for a strong effect, and the huge ranges of data sets 2 and 4 (around means that indicate strong effects) advise to increase the sample sizes of the real data if possible. $[SES;EES]^{d_{s_0}}$ gives information about approximate measurement precision, plausibility of a change in sign, and relative magnitude of the effects. It is based on elementary statistics only, thus completely perspicuous even for the layman without a background in advanced statistics. It is, deliberately, a simplifying approximation, but it serves well as a first step into data analysis. An obvious disadvantage of this approximation is that it based on normality. One can correct this problem by using the same methods applied from (Keselman, Algina, Lix, Wilcox, and Deering 2008) for exact CI for ES. This is an area of future research.

7 SUMMARY AND CONCLUSION

p-values, confidence intervals, and effect sizes (the "Big Three" (Hatcher 2013)) are standard measures of getting a first idea of the effects in data sets. For this purpose, however, p-values are clearly suboptimal, since confidence intervals (CI) and effect sizes (ES) do not only convey information about the likelihood of random effects, but also about the precision of the measure and the absolute and relative magnitude of the effect. This additional information, which is often crucial while searching for effects in huge data sets (Big Data), can also be condensed into confidence intervals for effect sizes, which are presumably the measures of maximum information density. They are recommended as the most instructive exact measures of first assessment. Unfortunately, they are as ambiguous under heteroscedasticity as effect sizes, and presuppose advanced statistics. A single, elementary interval that approximates exact CI for ES can be constructed using statistically secured effect sizes. It is as easy to interpret as a conventional CI, and only based on the assumption of sample normality. Increasing robustness of secured effect sizes with respect to non-normality while not decreasing simpleness is a matter of future research.

REFERENCES

Algina, J., H. Keselman, and R. Penfield. 2006. "Confidence intervals for an effect size when variances are not equal". *Journal of Modern Applied Statistical Methods* 5 (1): 2–13.

Armstrong, J. S. 2007. "Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries". *Internation Journal of Forecasting* 23:335–336.

Bakan, D. 1966. "The test of significance in psychological research". *Psychological Bulletin* 66:423–437.

Berkson, J. 1938. "Some significance of interpretation encountered in the application of the Chisquare test". *Journal of the American Statistical Association* 33:526–536.

Boring, E. 1919. "Mathematical vs. scientific significance". *Psychological Bulletin* 16:335–338.

Brandstaetter, E. 1999. "Confidence intervals as an alternative to significance testing". *Methods of Psychological Research Online* 4 (2): 33–46.

Carver, R. 1978. "The case against stastistical significance testing". *Harvard Educational Review* 48:378–399.

Cohen, J. 1962. "The statistical power of abnormal-social psychological research: A review". *Journal of Abnormal and Social Psychology* 65:145–153.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale: Lawrence Erlbaum Associates.

Cohen, J. 1994. "The earth is round (p= 0.5)". *American Psychologist* 12:997–1003.

Cumming, G., and S. Finch. 2001. "A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions". *Educational and Psychological Measurement* 61:532–574.

Edwards, W., H. Lindman, and L. J. Savage. 1963. "Bayesian statistical inference for psychological reseach". *Psychological Review* 70:193–242.

Ellis, P. D. 2010. *The essential guide to effect sizes*. Cambridge, GB: Cambridge University Press.

Falk, R. 1998. "In criticism of the null hypothesis statistical test". *American Psychologist* 53:798–799.

Fanelli, D. 2012. "Negative results are disappearing from most disciplines and countries". *Scientometrics* 90 (3): 891–904,.

Fisher, R. A. 1955. "Statistical methods and scientific induction". *Journal of the Royal Statistical Society. Series B (Methodological)* 17:69–78.

Gigerenzer, G. 2004. "Mindless statistics". *The Jounal of Socio-Economics* 33:587–606.

Glass, G. V., B. McGaw, and M. L. Smith. 1981. *Meta-analysis in social research*. Beverly Hills CA: Sage.

Greenwald, A. 1975. "Consequences of predjudice agains the null hypothesis". *Psychological Bulletin* 82:1–20.

Hagen, R. 1997. "In praise of the null hypothesis test". *American Psychologist* 52:15–24.

Harris, M. J. 1991. "Significance tests are not enough: The role of effect size estimation in theory corroboration.". *Theory and Psychology* 1:375–382.

Hatcher, L. 2013. *Advanced Statistics in Research: Reading, Understanding, and Writing Up Data Analysis Results*. Shadow Finch Media.

Hayes, A. F., and L. Cai. 2007. "Further evaluating the conditional decision rule for comparing two independent means.". *British Journal of Mathematical and Statistical Psychology* 60:217244.

Hedges, L. V. 1981. "Distribution theory for Glass's estimator of effect size and related estimators". *Journal of Educational Statistics* 6:107–128.

Hubbard, R. 2004. "Alphabet soup: Blurring the distinctions between ps and alphas in psychological research". *Theory & Psychology* 14:295327.

Hubbard, R., and J. Armstrong. 2006. "Why we don't really know what statistical significance means: Implications for educators.". *Journal of Marketing Education* 28:114–120.

Hubbard, R., and R. M. Lindsay. 2008. "Why p values are not a useful measure of evidence in stastistical significance testing". *Theory and Psychology* 18:69–88.

Ioannidis, J. 2005. "Why most puplished research findings are false". *PLoS Medicine* 2 (8): e124.

Kasuya, E. 2001. "Mann-Whitney-Test when variances are unequal". *Animal Behaviour* 61:1247–1249.

Keselman, H., J. Algina, L. Lix, R. Wilcox, and K. Deering. 2008. "A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes". *Psychological Methods* 13 (2): 110–129.

Kirk, R. E. 1996. "Practical significance: A concept whose time has come". *Educational and Psychological Measurement* 56:746–759.

Lambdin, C. 2012. "Significance tests as sorcery: Science is empirical - significance tests are not". *Theory and Psychology* 22 (1): 67–90.

Moser, B., and G. Stevens. 1992. "Homogeneity of variance in the two sample means test". *The American Statistician* 46 (1): 19–22.

Mulaik, S., N. Raju, and R. Harshman. 1997. "There is a time and a place for significance testsng". In *What if there were no significance tests?*, edited by L. Harlowand, S. Mulaik, and J. Steiger, 65–115. Erlbaum.

Neuhaeuser, M. 2010. "An equivalence test based on n and p". *Journal of Modern Applied Statistical Methods* 9 (1): 304–307.

Osborne, J. W. 2008. *Best Practices in Quantitative Methods*. Sage Publications.

Parkhurst, D. F. 2001. "Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation". *BioScience* 51 (12): 1051–1057.

Pezzullo, J. 2013. *Biostatistics for Dummies*. John Wiley and Sohns.

Poole, C. 2001. "Low p-values or narrow confidence intervals: which are more durable". *Epidemiology* 12 (3): 291–294.

Prentice, D., and D. Miller. 1992. "When small effects are impressive". *Psychological Bulletin* 112:160–164.

Rosenthal, R. 1979. "The file drawer problem and tolerance for null results.". *Psychological Bulletin* 86 (3): 638–641.

Rosnow, R., and R. Rosenthal. 1989. "Statistical procedures and the justification of knowledge in psychological science". *American Psychologist* 44:1246–1284.

Ruxton, G. D. 2006. "The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test". *Behavioral Ecology* 17 (4): 688–690.

Schmidt, F., and J. Hunter. 1997. "Eight common but false objections to the discontinuation of significance testing in the analysis of research datat". In *What if there were no significance tests?*, edited by L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 37–64. Erlbaum.

Sedlmeier, P. 1996. "Jenseits des Signifikanztest-Rituals: Ergaenzungen und Alternativen". *Methods of Psychological Research Online* 1 (4): 41–63.

Senn, S. 2001. "Two cheers for P-values?". *Journal of Epidemiology and Biostatistics* 6 (2): 193–204.

Thompson, B. 2007. "Effect sizes, confidence intervals, and confidence intervals for effect sizes". *Psychology in the Schools* 44 (5): 423–432.

Thompson, B. 2008. "Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes.". In *Best practices in quantitative methods*, edited by J. Osborne. Newbury Park CA: Sage.

Troitzsch, K. 2014. *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*, Chapter Analysing Simulation Results Statistically: Does Significance Matter?, 88–105. PA, USA: Hershey.

Tukey, J. 1991. "The philosophy of multiple comparison". *Statistical Science* 6:100–116.

White, J., A. Rassweiler, J. Samhouri, A. Stier, and C. White. 2014. "Ecologists should not use statistical significance tests to interpret simulation model results". *Oikos* 123:385–388.

Wilkinson, L. 1999. "Task force on Statistical Inference: Statistical methods in psychology journals". *American Psychologist* 54:594–604.

AUTHOR BIOGRAPHY

MARKO HOFMANN is Chief Scientist at ITIS GmbH in Neubiberg, Germany since 2000, and adjunct Professor at the University of the Federal Armed Forces in Munich, Germany since 2010. He holds a M.S., a Ph.D. and the venia legendi in Computer Science. His email address is marko.hofmann@unibw.de.