Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015

Tetsuya Sakai
Waseda University, Japan.
tetsuyasakai@acm.org

ABSTRACT

We conducted a systematic review of 840 SIGIR full papers and 215 TOIS papers published between 2006 and 2015. The original objective of the study was to identify IR effectiveness experiments that are seriously underpowered (i.e., the sample size is far too small so that the probability of missing a real difference is extremely high) or overpowered (i.e., the sample size is so large that a difference will be considered statistically significant even if the actual effect size is extremely small). However, it quickly became clear to us that many IR effectiveness papers either lack significance testing or fail to report p-values and/or test statistics, which prevents us from conducting power analysis. Hence we first report on how IR researchers (fail to) report on significance test results, what types of tests they use, and how the reporting practices may have changed over the last decade. From those papers that reported enough information for us to conduct power analysis, we identify extremely overpowered and underpowered experiments, as well as appropriate sample sizes for future experiments. The raw results of our systematic survey of 1,055 papers and our R scripts for power analysis are available online. Our hope is that this study will help improve the reporting practices and experimental designs of future IR effectiveness studies.

Keywords

effect sizes; evaluation; power analysis; sample sizes; statistical power; statistical significance; systematic review

1. INTRODUCTION

In experiments for retrieval effectiveness evaluation, the *de facto* standard is the practice of comparing the mean evaluation measure scores across topics by means of statistical significance tests such as the (paired and unpaired) t-test and Analysis of Variance (ANOVA). While alternative approaches to these standard tests exist (e.g., Carterette's Baysian inference [3], Robertson and Kanoula's view of *documents* as the source of variance [12], and Killeen's probability of replication for experiments in psychology [10]), so far they have not yet enjoyed the same popularity as the classical significance tests in the IR community.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911492

In standard signifiance testing, the *test statistic* (e.g., the t statistic for the t-test and the F statistic for ANOVA) is a function of the *effect size* (i.e., the actual magnitude of the difference between systems) and the *sample size* (i.e., the number of topics, users etc.), even though what we are usually interested in is the effect size: is the difference "substantial"? However, due to the above relationship, increasing the sample size increases the value of the test statistic, which in turn decreases the *p-value* (i.e., the probability of observing the obtained result or something more extreme under the null hypothesis). Since we conclude a result to be statistically significant if, for a predetermined *significance level* α, we observe that $p\text{-}value \leq \alpha$ holds, we can make *anything* statistically significant by using a large enough sample. Conversely, if the sample size is too small, we might be missing differences that we really should be detecting. While α is the predetermined probability of Type I error (detecting a nonexistent difference), β is the probability of Type II error (missing a real difference), and the ability to detect a real diffrence, given by $(1-\beta)$, is the *statistical power*. The combination $(\alpha, \beta) = (0.05, 0.20)$, known as *Cohen's five-eighty convention* (where "eighty" refers to 80% power), is often used as a standard setting for determining the sample size for an experiment [5, 15].

Recent statistically motivated studies have suggested that topic sets used in IR test collections should be substantially larger than they currently are in order to meet a clear set of statistical requirements. For example, Sakai's *topic set size design* results show that if researchers use the paired t-test for ad hoc IR and want to ensure Cohen's five-eighty convention for *any* system difference of 0.05 (or higher) in mean average precision, about 300 topics are required [15][1]. Urbano, Marrero and Martín conclude that "*in most cases a couple hundred*" topics are required for stable system rankings from their study based on the *Generalisability Theory* [20]. While these studies do not imply that all system comparison experiments based on 50 topics are invalid, they do suggest that some of the experiments may be *underpowered*: we may be missing a lot of real differences due to small sample sizes. On the other hand, the advent of web search engines brought with it the practice of using thousands or even millions of queries from their query log data for averaging; we suspected that some of such studies may be *overpowered*: even if the effect size is very small, statistically significant differences can be obtained due to large sample sizes. The original objective of this study was to identify IR effectiveness experiments that are extremely underpowered or overpowered.

We conducted a systematic review of 840 ACM SIGIR full papers and 215 ACM TOIS papers published between 2006 and 2015. SIGIR is considered by many to be the premier conference in IR,

[1]Over 700 topics are required if unstable measures such as expected reciprocal rank (ERR) is used [15].

while TOIS enjoys a similar status in the journal domain; this journal explicitly instructs authors as follows: "*When reporting statistics, the name of the statistic, the degrees of freedom, the value obtained, and the p-value should be reported, e.g.,* $F(3, 65) = 4.83, p < 0.01$"[2]. However, it quickly became clear to us that many IR effectiveness papers published in SIGIR and TOIS either lack significance testing or fail to report p-values and/or test statistics, which prevents us from conducting power analysis. Hence we first report on how IR researchers (fail to) report on significance test results, what types of tests they use, and how the reporting practices may have changed over the last decade. From those papers that reported enough information for us to conduct power analysis, we identify extremely overpowered and underpowered experiments, as well as appropriate sample sizes for future experiments. The raw results of our systematic survey of 1,055 papers are available online. We hope that this study will help improve the reporting practices and experimental designs of future IR effectiveness studies.

2. PRIOR ART

Two papers directly inspired us to conduct the present systematic review. The first is the study by Armstrong *et al.* [1]: they examined SIGIR papers in the period 1998-2008 and CIKM papers in the period 2004-2008 to investigate whether IR systems have improved over the years. A total of 106 papers, which contained IR effectiveness results using TREC test collections, were analysed in their survey. One of their main findings was that researchers often claim statistically significant results but use low baselines for comparison. The second is the study by Kelly and Sugimoto [9]: they analysed 127 journal and conference papers selected from 2,791 papers in the period 1967-2006 and investigated the evaluation practices in *interactive* information retrieval (IIR), including experimental designs, corpora and measures used. In the present study, we tried to follow Kelly and Sugimoto's description of *systematic reviews* to the best of our ability: "*Researchers articulate a plan for gathering and analyzing studies and attempt to be exhaustive with their coverage of the literature. Researchers also take a neutral position during analysis and attempt to create generalizations from findings. Systematic reviews adhere to strict scientific guidelines to minimize potential selection and interpretation biases to ensure replicability (and hence reliability).*" Like Armstrong *et al.* [1], we are primarily interested in *system* effectiveness studies that typically rely on averaging across a set of topics. However, as we shall see, some of the findings from the present study partially overlap with one of Kelly and Sugimoto's main finding from their systematic review of *IIR* studies: "*Because of the basic goals and design of these studies, the majority of researchers used either ANOVA or t-tests to analyze the results. [...] In some cases, the type of test conducted was not reported although statistically significant results were claimed and/or p-values were presented.*"

Sanderson and Zobel [16] report on a small-scale survey which is also relevant to our present study: they examined 26 system effectiveness papers from SIGIR 2003 and 2004 to see which significance tests and evaluation measures were used. Their findings, which are also in line with ours, are perhaps also worth quoting: "*We found that significance was not explicitly reported in 14 of the papers. In two it was implied such tests had been tried, but outcomes were not given. In three or four of these papers, the improvements were large and arguably a significance test was unnecessary. However, in at least six papers (23% of the sample) the reported improvements were small, sometimes no more than a few percent in relative MAP.*"; "*Among the 12 papers with significance tests, one used both ANOVA and the t-test, five each used either the t-test or Wilcoxon's test, and in one, the test was not identified.*"

[2]https://tois.acm.org/authors.cfm visited April 12, 2016.

Table 1: Statistics of the ACM papers examined in this study.

Year	SIGIR	TOIS (Volume.Issue)	SIGIR+TOIS
2006	74	17 (24.1-24.4)	91
2007	85	23 (25.1-26.1)	108
2008	85	24 (26.2-27.1)	109
2009	78	18 (27.2-27.4)	96
2010	87	28 (28.1-29.1)	115
2011	108	16 (29.2-29.4)	124
2012	98	25 (30.1-30.4)	123
2013	73	22 (31.1-31.4)	95
2014	82	21 (32.1-32.4)	103
2015	70	21 (33.1-33.4)	91
Total	840	215	1,055

In the present study, we manually examined 1,055 SIGIR and TOIS papers, and found that at least 862 of them appear to deserve statistical significance testing for performance differences; we then analysed these 862 papers further. This number is substantially higher compared to the above studies, although the numbers are not directly comparable due to differences in objectives and methods of analysis. More importantly, unlike prior art, we conducted power analysis for 133 papers that reported the p-value and/or the test statistic, and computed the achieved power as well as appropriate sample sizes for future experiments.

3. SURVEY METHOD

The primary purpose of this systematic survey is to examine how IR researchers, especially those working on improving IR *effectiveness*, use statistical significance tests, and to conduct power analysis wherever possible. We are interested, for example, in whether the topic set sizes are appropriate in test collection-based studies, and whether the number of participants or observations is appropriate in user-based studies. Table 1 shows the number of papers per year that we examined[3]. First, we created two lists of DOIs, one for SIGIR and the other for TOIS, on separate sheets in an Excel file, and we created columns in the Excel sheets according to a strict coding scheme as described below. Then, the author of this paper manually examined each paper at least twice. The coding started in February 2015, and was completed in October 2015. The pdf file of each paper was downloaded from the ACM digital library and was viewed on computer screen; for some papers that required careful interpretations of significance test results, hard copies were also used[4].

3.1 Coding Scheme

Each paper was coded in an Excel file as follows.

Step 1 Does the paper contain a table or figure of mean effectiveness, mean user performance, etc. that appears to deserve significance testing? If YES, select and record the name of one such table or figure, which we refer to as a *representative table* throughout this paper. No representative table was selected for papers that did not discuss IR effectiveness (e.g., those only discussing efficiency or theory). Note that we do not look at the magnitude of the difference in means here; Instead, Step 5 quantifies effect sizes wherever possible.

Step 2 Does the paper conduct a significance test? Assign exactly one category from (A)-(I) shown in Table 2. If multiple test types are used, pick a primary one, preferably from (A)-(E)

[3]TOIS published Volume 34 Issue 1 (6 papers) in October 2015. These papers are outside the scope of this study.

[4]The pdf files of three SIGIR papers were not searchable by Ctrl-F; for these cases also, hard copies were used.

Table 2: Paper categorisation scheme used in this study.

Category	
(A)	Unpaired t-test
(B)	Paired t-test
(C)	One-way ANOVA
(D)	Two-way ANOVA without replication
(E)	Two-way ANOVA
(F)	Other tests
(G)	Test type not specified
(H)	CIs used instead of significance tests
(I)	No significance tests

Table 3: Subcategories for papers that compare two systems

(Sub) category	significance test type
(A)	Unpaired t-test
(B)	Paired t-test
(F1)	Wilcoxon signed rank test
(F2)	Sign test
(F3)	Bootstrap test
(F4)	Randomisation test
(F5)	Wilcoxon rank sum test (Mann-Whitney U test)
(F6)	Other tests

to enable power analysis. Specifically, we searched for "statistical", "signi", "test" and "ANOVA" within the pdf file.

Step 3 If categorised as (I) in Step 2, check if the paper claims "significance" and record the exact expressions in an Excel popup comment. Section 4.2 provides more details.

Step 4 If categorised as (A)-(G) in Step 2, record whether the p-value and/or test statistic is reported. If only the p-value is reported, and if categorised as (A)-(E) in Step 2, compute the test statistic as described later (Eqs. 1-2).

Step 5 If the test statistic for categories (A)-(E) is obtained in Step 4 and the sample sizes actually used are indicated in the paper, conduct power analysis using tools described in Section 3.2 under Cohen's five-eighty convention ($\alpha = 0.05, \beta = 0.20$). If the achieved power is 0.99 or higher, label the paper as *overpowered*; else if the achieved power is 0.50 or below, label the paper as *underpowered*; otherwise, label it as *about right*. Also, record the effect size and the *future sample sizes* (i.e., recommended sample sizes for new experiments with similar purposes and settings) that are output by the tool. While the above power thresholds are arbitrary, note that the interested reader can easily apply different thresholds to our raw Excel file.

Step 6 In addition, copy and paste sentences from the paper that are relevant to power analysis (e.g., how exactly they report the p-values and test statistics). Save them as an Excel popup comment. Record the investigator's (i.e., the present author's) own comments in an Excel cell.

In Step 2, if the paper mentioned a t-test but did not indicate two sample sizes, we assumed that the test is paired (Category (B)) as this is the typical setting in test collection-based studies; if the paper did not indicate whether the test is two-sided or one-sided, we interpreted it as the more conservative two-sided test when conducting power analysis. Category (G) was chosen if the paper mentioned a significance level (α), a p-value, or just *statistical* significance, but did not specify the significance test used; Category (H) was chosen if the paper did not mention any significance testing *but* used CIs, boxplots, (often unexplained) error bars, or reported standard errors/deviations. Category (I) was chosen if the paper did not mention significance testing at all, even if it claimed "significant" improvements (without using the word "statistical").

In Step 4, we considered the p-value as "reported in the paper" if the *exact* p-value was reported. We did not regard "$p < 0.1, p < 0.05, p < 0.01$" as exact p-values; rather, we regarded them as the predetermined significance level α, which may or may not be substantially larger than the actual p-values. On the other hand, we count "$p < 0.001$" and smaller values as "reported in the paper" as such values should be accurate enough for power analysis. Whenever the test statistic t was not reported but the p-value and the sample sizes were, we deduced the test statistic as:

$$t = t_{inv}(\phi; p\text{-}value) \tag{1}$$

where $t_{inv}(\phi; P)$ is the critical t value for ϕ degrees of freedom and probability P, obtained with Microsoft Excel as $\texttt{T.INV.2T}(P, \phi)$ for two-sided tests, or $\texttt{T.INV}(1 - P, \phi)$ for one-sided tests[5]. For *paired* t-tests with sample size n, we let $\phi = n - 1$; for unpaired t-tests with sample sizes n_1 and n_2, we let $\phi = n_1 + n_2 - 2$. Similarly, for ANOVA results where p-values were reported but the F values were not, we computed them as follows whereover possible:

$$F = F_{inv}(\phi_1, \phi_2; p\text{-}value) \tag{2}$$

where $F_{inv}(\phi_1, \phi_2; P)$ is the critical F value for (ϕ_1, ϕ_2) degrees of freedom and probability P, or $\texttt{F.INV.RT}(P, \phi_1, \phi_2)$ with Microsoft Excel. The degrees of freedom are computed in accordance with ANOVA.

After coding each paper as described above, we focussed on papers labelled "(F) Other tests" and assigned a subcategory to each of them, for the purpose of examining the popularity of different significance tests for comparing two systems. The subcategories are shown in Table 3; (A)-(F5) are used in the analysis in Section 4.3. Category (F6) includes papers that used nonparametric tests for more than two systems.

Admittedly, even though our coding scheme is systematic, it is impossible to completely rule out the possibility that the present author has misinterpreted some of the papers. To accommodate correction, we have created a twitter account solely for this purpose[6]: if an assessment of a paper needs to be corrected, the author of that paper can point this out by using either `TOIS` or `SIGIR` followed by the last seven digits of the paper's DOI as the hashtag (e.g., `#SIGIR1148261`). Recall that our intepretation of each paper is available as a raw Excel file and therefore that it is easy to correct it. However, it is highly unlikely that these minor "bug reports" will affect the main conclusions of the present paper.

3.2 Power Analysis Tools

We conducted a power analysis for every paper in Categories (A)-(E) where the test statistic and actual sample sizes were obtained, based on modified versions of t-test and ANOVA power analysis tools provided by Toyoda [19]. The original R scripts of Toyoda, which simply rely on R's library `pwr`[7], are available at the publisher's website[8]; our own versions, which differ from Toyoda's only in input and output specifications[9], are available from our website[10], along with our raw systematic review results. We provide a function for each significant test type, for computing the

[5]For one SIGIR 2014 paper, the t statistic was computed by dividing the mean difference by the reported standard error of the mean. This paper is discussed in Section 4.6 (Table 7 Entry #12).
[6]http://twitter.com/IRsysrev
[7]https://cran.r-project.org/web/packages/pwr/pwr.pdf
[8]http://www.tokyo-tosho.co.jp/download/DL02065.zip
[9]The present author is solely responsible for the modifications and any errors introduced thereby.
[10]http://www.f.waseda.jp/tetsuya/data.html

effect size, achieved power and recommended future sample sizes. In what follows, we adhere to Toyoda's notations when we refer to effect sizes ($\hat{es}$ for t-tests; $\hat{f}$ and $\hat{f}^2$ for ANOVA).

3.2.1 *Unpaired (i.e., two-sample) t-test*

Our R function for unpaired t-tests is called `future.sample.unpairedt`, whose arguments are the t statistic (t), two sample sizes (n_1, n_2), whether the test is two-sided or one-sided (default: two-sided), α, and power ($1-\beta$) (default: Cohen's five-eighty convention). The script computes first the sample effect size $\hat{es} = |t|\sqrt{(n_1+n_2)/(n_1 n_2)}$ (which expresses the between-system difference in standard deviation units) and then the achieved power and the recommended sample size per group n' for future experiments, by calling the function `power.t.test` (if $n_1 = n_2$) or `pwr.t2n.test` (if $n_1 \neq n_2$). If a paper reports $n_1 + n_2$ but not n_1 and n_2 individually, we assume that $n_1 = n_2$.

3.2.2 *Paired t-test*

Our R function for paired t-tests is called `future.sample.pairedt`, whose arguments are the t statistic (t), the sample size (n), whether the test is two-sided or one-sided (default: two-sided), α, and power ($1-\beta$) (default: Cohen's five-eighty convention). The script computes first the sample effect size $\hat{es} = |t|/\sqrt{n}$ and then the achieved power and the future sample size n', by calling `power.t.test`.

3.2.3 *One-way ANOVA*

Our R function for one-way ANOVA tests is called `future.sample.1wayanova`, whose arguments are the F statistic (F), number of groups being compared (m), number of observations per group (n), α, and power ($1-\beta$) (default: Cohen's five-eighty convention). The script computes first the sample effect size $\hat{f} = \sqrt{\phi_A F/\phi_E}$ (where $\phi_A = m-1, \phi_E = m(n-1)$) and then the achieved power and the future sample size per group n', by calling `pwr.anova.test`. Here, $\hat{f}$ is an estimate of how large the between-group population standard deviation is compared to the within-group population standard deviation.

3.2.4 *Two-way ANOVA without Replication*

Our R function for two-way ANOVA (without replication) tests is called `future.sample.2waynorep`, whose arguments are the same as `future.sample.1wayanova`. The script first computes the sample effect size $\hat{f}^2 = \phi_A F/\phi_E$ (where $\phi_A = m-1, \phi_E = (m-1)(n-1)$) and then the achieved power and the future sample size per group n', by calling `pwr.f2.test`. Note that this tool outputs $\hat{f}^2$ rather than $\hat{f}$ as the effect size, simply because that is what `pwr.f2.test` requires as an argument[11]. An equivalent way to express $\hat{f}^2$ would be $\hat{f}^2 = \eta_p^2/(1-\eta_p^2)$, where η_p^2 expresses how much of the total variance (after removing other effects) is due to factor A; it can be computed as $\eta_p^2 = \frac{\phi_A F}{\phi_A F + \phi_E}$. Also, η_p is known as the *partial correlation ratio*.

3.2.5 *Two-way ANOVA*

Our R function for two-way ANOVA tests is called `future.sample.2wayanova`, whose arguments are the F statistics (F_A, F_B, F_{AB}), number of groups (m), number of cells per group (n), total number of observations (N), α, and power ($1-\beta$) (default: Cohen's five-eighty convention). For example, A, B, AB could represent the system and topic effects and the interaction between them, respectively. Let the degrees of freedom for A, B, AB and the residual E be $\phi_A = m-1, \phi_B = n-1, \phi_{AB} = (m-1)(n-1), \phi_E = N - mn$. Then the sample effect size for A is given by $\hat{f}_A = \sqrt{\hat{\eta}_{pA}^2/(1-\hat{\eta}_{pA}^2)}$, where $\hat{\eta}_{pA}^2 = \frac{\phi_A F_A}{\phi_A F_A + \phi_E}$; $\hat{f}_B$ and $\hat{f}_{AB}$ are computed similarly. The achieved power and the *total* sample size for future experiments N' are computed by calling `pwr.anova.test`.

[11] http://127.0.0.1:25552/library/pwr/html/pwr.f2.test.html

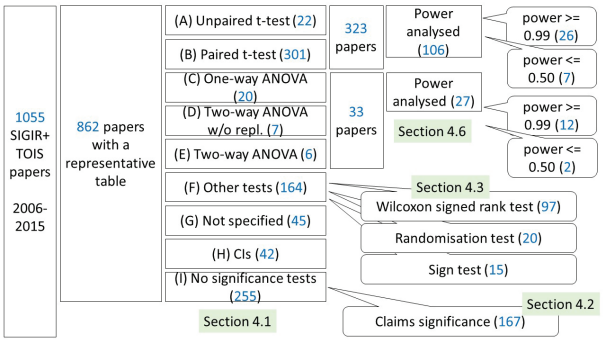

Figure 1: Number of papers concerning results reported in Sections 4.1, 4.2, 4.3, and 4.6 (SIGIR+TOIS).

4. RESULTS

In Step 1 of the coding scheme described in Section 3.1, a representative table was selected for 700 out of the 840 SIGIR papers (83%), and for 162 papers out of the 215 TOIS papers (75%). Thus, our view is that at least 862 papers out of the 1,055 papers that we examined may deserve statistical significance testing for comparing mean effectiveness scores and the like. Hereafter, we focus our attention to these 862 papers. Figure 1 provides an overview of some of the results reported in this section, in terms of paper counts.

4.1 Paper Distribution over Categories

Figure 2 shows the distributions of the aforementioned 700 SIGIR and 162 TOIS papers over the seven categories shown in Table 2. Paper counts are shown in addition to percentages. It can be observed that the distributions for SIGIR and TOIS are similar: 35-37% of the papers use the paired t-test (Category (B)); about 28-30% do not report significance test results even though these papers have a representative table (Category (I)); and about 18-24% use tests other than the t-test or ANOVA (Category (F))[12]. We shall examine Catgories (I) and (F) more closely in Sections 4.2 and 4.3, respectively.

Figure 3 shows the distribution of papers over the categories and across the timeline; the top graph shows the results for SIGIR; the bottom one shows the results obtained by summing the SIGIR and TOIS statistics. The number of TOIS papers alone per year is considered too small for our analysis and are not shown here. For each year, the number of papers in each category is divided by the total number of papers with a representative table for that year; the latter is shown below the x axis for each year. It appears that the the use of the paired t-test is now more common than a decade ago, and that we are also seeing fewer papers without statistical significance tests. In Section 4.4, we shall compare the results for 2006 and those for 2015 from this viewpoint.

[12] Recall, however, that we had to select exactly one significance test type per paper even if both t-tests and ANOVA were used in the same paper.

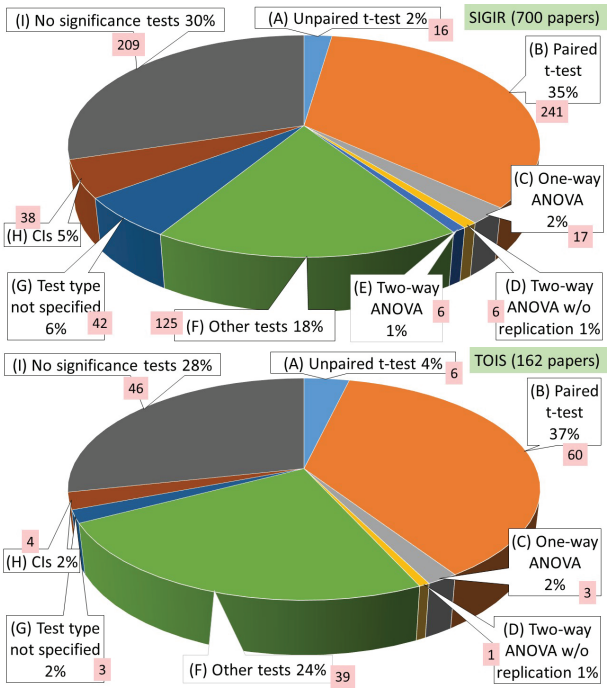

Figure 2: Distributions of papers with a representative table over categories.

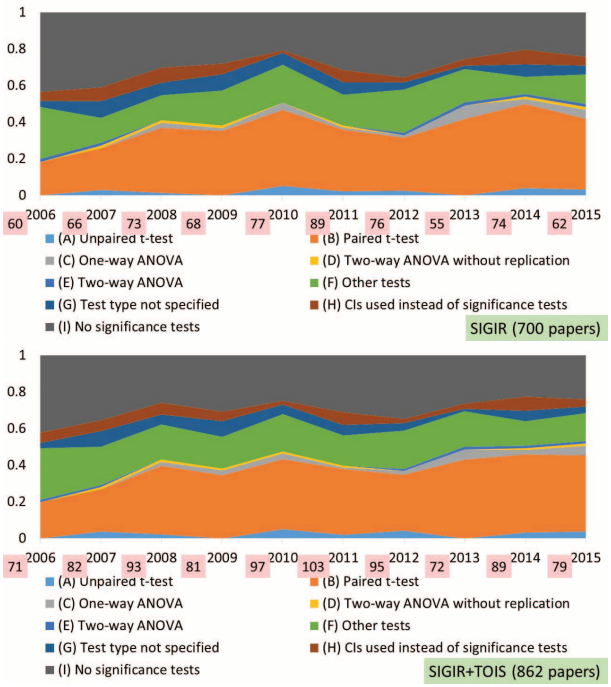

Figure 3: Distributions of papers with a representative table over categories across 10 years.

Table 4: Category (I) papers (no significance tests): Column (a) shows the number of papers that claim "significant" improvements and the like; Column (b) shows the number of papers that make no such claims. For the SIGIR and SIGIR+TOIS columns, the percentages are also shown, where the denominators are the paper counts shown in Figure 3.

	SIGIR		TOIS		SIGIR+TOIS	
	(a)	(b)	(a)	(b)	(a)	(b)
2006	**14** (23%)	12 (20%)	**1**	3	**15** (21%)	15 (21%)
2007	**21** (32%)	6 (9%)	**1**	1	**22** (27%)	7 (9%)
2008	**14** (19%)	8 (11%)	**0**	2	**14** (15%)	10 (11%)
2009	**15** (22%)	4 (6%)	**5**	1	**20** (25%)	5 (6%)
2010	**10** (13%)	6 (8%)	**5**	3	**15** (15%)	9 (9%)
2011	**13** (15%)	15 (17%)	**3**	1	**16** (16%)	16 (16%)
2012	**19** (25%)	8 (11%)	**6**	0	**25** (26%)	8 (8%)
2013	**7** (13%)	7 (13%)	**4**	1	**11** (15%)	8 (11%)
2014	**11** (15%)	4 (5%)	**4**	1	**15** (17%)	5 (6%)
2015	**10** (16%)	5 (8%)	**4**	0	**14** (18%)	5 (6%)
total	**134** (19%)	75 (11%)	**33**	13	**167** (19%)	88 (10%)

4.2 Claiming Significance without Providing Significance Test Results

Why are there so many papers without significance tests despite explicit instructions such as the ones from TOIS (Section 1)? Recall that all of these Category (I) papers have a representative table that appears to deserve significance testing. Possible good reasons include: (i) the authors consider statistical significance testing to be of limited or no value (e.g. [7, 8]); (ii) the authors judge that significance testing is unnecessary in their particular situations because either the sample size and/or the effect size is very large, or because the difference is not the central point the study. For example, one SIGIR 2012 paper (DOI: 10.1145/2348283.2348301) states: "*Since the present study does not aim to prove one retrieval method better than another, we report the findings without tests on significance of statistical differences.*" [2]. In Category (I), however, we did find some papers that use expressions such as "significant improvement" and "significantly outperform" in the context of discussing effectiveness even though *statistical* significance tests are never mentioned. Some of these claims are even made in the paper abstracts and conclusions. We argue that such practices are quite misleading and that the use of ambiguous expressions such as those mentioned above should be avoided. Table 4 breaks down the Category (I) papers from each year into papers that claim "significant" improvements and the like, and those that do not. It is worrying that 167 papers (19%) out of 862 papers with a representative table say "significant" without conducting significance tests, even if in some cases it may be clear from the context that the word is being used in the non-statistical sense.

4.3 Comparing Two Systems: Popular Tests

As was mentioned in Section 3.1, we used the subcategories shown in Table 3 to examine which significance tests are popular for comparing two systems. Historically, IR researchers in the 20th century were relatively reluctant to use parametric tests [14], but nowadays the robustness of the t-test (which is parametric) is recognised and the test is widely used. Savoy [17] and Sakai [13] advocated the use of the bootstrap test in 1997 and 2006, respectively, while Smucker, Allan and Carterette [18] advocated the randomisation test in 2007. Are these computer-based, distribution-free tests used more often now? Figure 4 breaks down 365 SIGIR papers and 99 TOIS papers that used a significance test for comparing two systems by test type. Again, the overall pictures are very similar for these two venues: 61-66% of these papers use the paired t-test; 20-23% use the Wilcoxon signed rank test[13]; 4-5% use the randomisation test; 3-4% use the sign test; only 1% use the bootstrap test. Recall, however, that we picked one primary test from each paper even if it utilised multiple test types. Figure 5 shows how the popularity of these tests have changed over the last decade; note that the bottom graph aggregates the statistics from SIGIR and TOIS as before. It appears that the paired t-test is now more popu-

[13]"Wilcoxon test" was always interpreted as the Wilcoxon signed rank test, not as the Wilcoxon rank sum test.

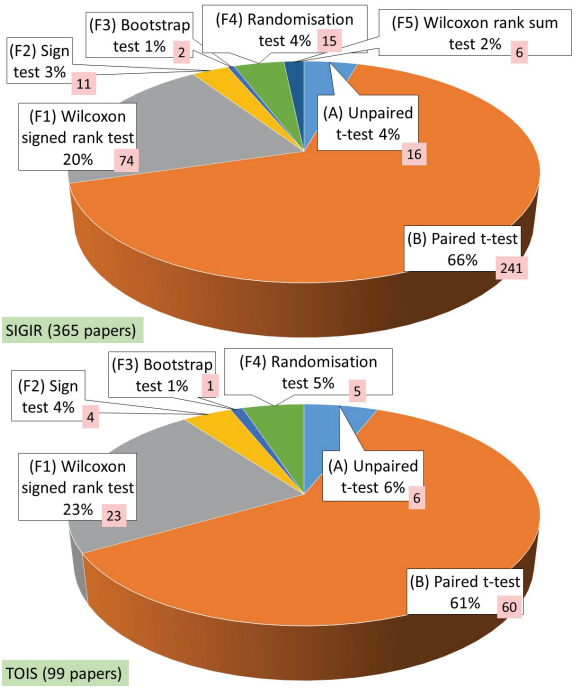

Figure 4: Distributions of papers over significance test types for comparing two systems.

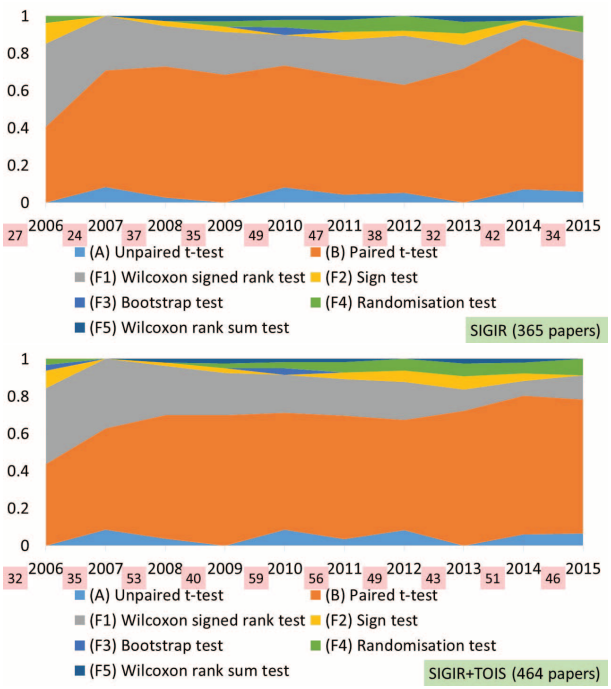

Figure 5: Distributions of papers over significance test types for comparing two systems across 10 years.

lar than it was 10 years ago, and that the Wilcoxon signed rank test is less so. In Section 4.4, we shall compare the results for 2006 and 2015 from this viewpoint.

We came across a few papers that explicitly discussed why a particular test type was chosen, although we did not conduct an exhaustive search for such comments. In a TOIS paper from 2012 (10.1145/2382438.2382445), the authors cite Sanderson and Zobel's SIGIR 2005 paper [16] and state: "*We report statistical significance test results using the nondirectional paired t-test at a confidence level of 0.01, since this test has been shown to be more reliable than the Wilcoxon and signed tests*" [6]; A SIGIR 2008 paper (10.1145/1390334.1390407) provides a similar argument. In a SIGIR 2015 paper (10.1145/2766462.2767700), the authors cite the aforementioned CIKM 2007 paper by Smucker *et al.* [18] as well as Sanderson and Zobel [16], and choose the randomisation test.

4.4 Have the Proportions Changed over the Last Decade?

Figures 3 and 5 suggest the following four clear trends: (1) In Figure 3, the proportion of Category (I) (no significance tests) papers is decreasing; (2) Similarly, the proportion of Category (B) (paired t-test) papers is increasing; (3) In Figure 5, the proportion of Category (F1) (Wilcoxon signed-rank test) papers is decreasing; (4) Similarly, the proportion of Category (B) (paired t-test) papers is increasing. Recall that Figure 3 considers all papers with a representative table, while Figure 5 only considers papers that used a significance test for comparing two systems. Ideally, these observations deserve a time series analysis, where an observation from a given year is modelled as a function of an observation from previous years along with some noise. However, this would not be very useful with our data with only ten data points. An obvious alternative is to regard the observation from each year as an independent sample to conduct standard significance testing, but in practice the independence assumption is highly unlikely to hold: for example, if many SIGIR paper authors use the t-test in 2014, those who submit a paper to SIGIR 2015 are also likely to do so, perhaps after reading some of the SIGIR 2014 papers. Due to the above considerations, we focus our attention on the comparison between 2006 and 2015, i.e., the leftmost and rightmost results in Figures 3 and 5, and conduct significance testing for two samples, each from a different population, under an independence assumption. Below, we describe our (two-sided) significance testing procedure for comparing two proportions (one from 2006 and the other from 2015) [11].

From the first population, we draw a sample of size n_1, and find that x_1 of them holds a certain property (e.g., lacks significance testing); the observed proportion is x_1/n_1. The observed proportion for the second population is defined similarly as x_2/n_2. The null hypothesis H_0 is that the true proportions are equal: $P_1 = P_2 = P$. Let $\hat{P}_1^* = \frac{x_1+0.5}{n_1+1}, \hat{P}_2^* = \frac{x_2+0.5}{n_2+1}, \hat{P}^* = \frac{x_1+x_2+0.5}{n_1+n_2+1}$, which are the estimates of the true proportions P_1, P_2, P with continuity corrections, respectively. Under H_0, it is known that the distribution of $\hat{P}_1^* - \hat{P}_2^*$ can be approximated by $N(0, P(1-P)(\frac{1}{n_1}+\frac{1}{n_2}))$. Hence a z-test can be performed using the following test statistic:

$$u_0 = \frac{\hat{P}_1^* - \hat{P}_2^*}{\sqrt{\hat{P}^*(1-\hat{P}^*)(\frac{1}{n_1}+\frac{1}{n_2})}}. \quad (3)$$

The point estimate for the true difference between the two proportions is given by $\hat{P}_1^* - \hat{P}_2^*$, whereas the margin of error (MOE) for computing the 95% CI is given by:

$$z_{\alpha/2}\sqrt{\frac{\hat{P}_1^*(1-\hat{P}_1^*)}{n_1} + \frac{\hat{P}_2^*(1-\hat{P}_2^*)}{n_2}}, \quad (4)$$

where $z_{\alpha/2}$ is the critical z value for probability $\alpha/2 = 0.025$ (since we want 95% confidence).

Our four null hypotheses, which correspond to the aforementioned trends (1)-(4), state that the population proportion computed for 2015 is equal to that for 2006. We denote them as $H_0^1, H_0^2, H_0^3, H_0^4$. Table 5 summarises the results of the significance tests; Part (a)

Table 5: Statistical significance test results: comparing the proportions from 2006 and those from 2015.

(a) SIGIR	H_0^1	H_0^2	H_0^3	H_0^4
n_1 (2006)	60	60	27	27
x_1 (2006)	26	11	12	11
n_2 (2015)	62	62	34	34
x_2 (2015)	15	24	5	24
p-value	0.028	0.015	0.013	0.023
$P_1^* - P_2^*$	0.188	−0.200	0.289	−0.289
95% CI	[0.023, 0.353]	[−0.357, −0.044]	[0.065, 0.513]	[−0.530, −0.048]
(b) SIGIR+TOIS	H_1	H_2	H_3	H_4
n_1 (2006)	71	71	32	32
x_1 (2006)	30	14	13	14
n_2 (2015)	79	79	46	46
x_2 (2015)	19	33	6	33
p-value	0.019	0.004	0.006	0.015
$P_1^* - P_2^*$	0.180	−0.217	0.271	−0.273
95% CI	[0.031, 0.329]	[−0.361, −0.074]	[0.073, 0.468]	[−0.489, −0.057]

uses data from the SIGIR papers only, while Part (b) uses both SIGIR and TOIS papers. It can be observed that all the p-values are below $\alpha = 0.05$, and that all of the above null hypotheses are rejected. More importantly, the 95% CIs for H_0^1 and H_0^3 are above zero, while those for H_0^2 and H_0^4 are below zero. Hence, we can conclude, for both Parts (a) and (b), as follows:

1. The proportion of Category (I) (no significance tests) papers among those with a representative table in 2015 is statistically significantly smaller compared to 2006;

2. The proportion of Category (B) (paired t-test) papers among those with a representative table in 2015 is statistically significantly larger compared to a 2006;

3. The proportion of Category (F1) (Wilcoxon signed-rank test) papers among those using a test for comparing two systems in 2015 is statistically significantly smaller compared to 2006;

4. The proportion of Category (B) (paired t-test) papers among those using a test for comparing two systems in 2015 is statistically significantly larger compared to 2006.

4.5 Do Researchers Report p-values and/or test statistics?

Next, we focused on papers (with a representative table) that did conduct significant tests, regardless of the test type. For 453 SIGIR papers and 112 TOIS papers in Categories (A)-(G), Step 4 described in Section 3.1 further categorised them into four classes: (i) both exact p-values and test statistics are reported; (ii) only exact p-values are reported (from which test statistics may be deduced if the sample sizes are known); (iii) only test statistics are reported; and (iv) neither is reported. Note that Class (iv) includes papers that specify a significance level α but does not report the exact p-values. Researchers (who accept and rely on classical significance tests) should report p-values and effect sizes. Saying "significant at $\alpha = 0.05$" leads to dichotomous thinking ("significant or not?") [5], and is not very informative for the reasons discussed in Section 1. Figure 6 visualises the proportion of Classes (i)-(iii) against Class (iv) papers for each year. It also contains a table of the actual number of papers. It can be observed that over one half of the papers with significance tests report neither p-values nor test statistics, and that the situation does not seem to be improving.

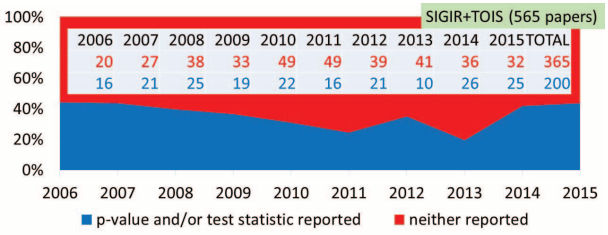

Figure 6: Proportion of papers (with a representative table) that report either p-values or test statistics (or both).

Table 6: Number of papers (with a representative table) with (a) an overpowered experiment (power ≥ 0.99); (b) an underpowered experiment (power ≤ 0.50); (c) "about right" experiment ($0.50 <$ power < 0.99).

	SIGIR			TOIS			SIGIR+TOIS		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
(A) Unpaired t-test	3	1	7	0	0	5	3	1	12
(B) Paired t-test	16	4	44	7	2	17	23	6	61
(C) One-way ANOVA	7	1	6	1	0	1	8	1	7
(D) Two-way ANOVA without replication	1	1	4	1	0	0	2	1	4
(E) Two-way ANOVA	2	0	2	0	0	0	2	0	2
total	29	7	63	9	2	23	38	9	86

4.6 Overpowered? Underpowered?

Among the 200 "good practice" papers indicated in Figure 6, we were able to conduct a power analysis for 99 SIGIR papers (75 with t-tests and 24 with ANOVA tests) and 34 TOIS papers (31 with t-tests and 3 with ANOVA test); the other 67 papers used other significance test types. Table 6 shows the paper counts for overpowered, underpowered, and "about right" papers for Categories (A)-(E). It can be observed that as many as 38 (29%) out of the 133 papers that went through power analysis are overpowered: the achieved power is 99% or higher for these papers. Whereas, only 9 (7%) papers were found to be underpowered: the achieved power is 50% or lower.

Hereafter, we focus on the power analysis results for SIGIR papers. Table 6 "SIGIR (a)" column shows that we found a total of 19 "overpowered" and 5 "underpowered" papers with t-tests (Categories (A) and (B)), as well as 10 "overpowered" and 2 "underpowered" papers with ANOVA tests (Categories (C)-(E)). Tables 7 and 8 provide complete power analysis results for these papers. Recall that n and n' denote the actual and future sample sizes *per group*, while N and N' denote the actual and future *total* sample sizes (Section 3.2). To quantify the gap between actual and future sample sizes, let us define *sample size ratio* as the actual sample size divided by the future sample size. Figure 7 plots sample size ratios against the achieved power; the baloons in the figure indicate which paper in Table 7 or 8 each dot corresponds to. For example, the top right balloon in Figure 7(a) indicates the third entry in Table 7, for which the effect size is $\hat{es} = 1.864$ (a "large" effect [4]) and the achieved power is 100%; although the actual sample size was $n = 192$, in fact $n' = 5$ is sufficient. The leftmost balloon in Figure 7(a) indicates the 24th entry in Table 7, for which $\hat{es} = 0.180$ (a "small" effect [4]) and the achieved power is only 15.2%; although the actual sample size was $n = 28$, in fact $n' = 244$ is needed to ensure 80% power. Similar relationships can be observed between Figure 7(b) and Table 8.

Table 7: SIGIR papers with overpowered (power ≥ 0.99) and underpowered (power ≤ 0.50) t-tests. Papers are sorted by achieved power, then by effect size.

#	year	doi (10.1145/*)	test type	t	actual sample size	effect size $\hat{es}$	achieved power	future sample size
1	2010	1835449.1835485	(B) Paired t-test	10.60	$n = 23$	2.209	1	$n' = 4$
2	2006	1148170.1148249	(B) Paired t-test	-25.27	$n = 182$	1.873	1	$n' = 5$
3	2009	1571941.1571991	(B) Paired t-test	25.83	$n = 192$	1.864	1	$n' = 5$
4	2007	1277741.1277784	(B) Paired t-test	5.68	$n = 45$	0.846	1	$n' = 14$
5	2010	1835449.1835518	(A) Unpaired t-test	-7.42	$n_1 = 486, n_2 = 114$	0.772	1	$n' = 28$
6	2014	2600428.2609577	(B) Paired t-test	11.41	$n = 234$	0.746	1	$n' = 17$
7	2008	1390334.1390407	(B) Paired t-test	7.46	$n = 100$	0.746	1	$n' = 17$
8	2010	1835449.1835459	(A) Unpaired t-test	7.40	$n_1 = n_2 = 605$	0.425	1	$n' = 88$
9	2008	1390334.1390352	(B) Paired t-test	7.51	$n = 600$	0.307	1	$n' = 86$
10	2012	2348283.2348313	(B) Paired t-test	8.26	$n = 7341$	0.096	1	$n' = 847$
11	2014	2600428.2609602	(A) Unpaired t-test	7.07	$n_1 = n_2 = 96762$	0.032	1	$n' = 15174$
12	2014	2600428.2609617	(B) Paired t-test	16.00	$n = 5352460$	0.007	1	$n' = 164107$
13	2008	1390334.1390412	(B) Paired t-test	5.07	$n = 8000$	0.057	0.999	$n' = 2441$
14	2007	1277741.1277821	(B) Paired t-test	5.03	$n = 25$	1.006	0.998	$n' = 10$
15	2006	1148170.1148214	(B) Paired t-test	4.90	$n = 20$	1.095	0.996	$n' = 9$
16	2011	2009916.2010052	(B) Paired t-test	4.61	$n = 11122$	0.044	0.996	$n' = 4105$
17	2015	2766462.2767712	(B) Paired t-test	4.42	$n = 3543978$	0.002	0.993	$n' = 1425634$
18	2014	2600428.2609586	(B) Paired t-test	4.51	$n = 30$	0.823	0.992	$n' = 14$
19	2008	1390334.1390370	(B) Paired t-test	4.46	$n = 60$	0.576	0.992	$n' = 26$
20	2007	1277741.1277778	(B) Paired t-test (one-sided)	1.55	$n' = 400$	0.077	0.461	$n' = 1032$
21	2014	2600428.2609632	(A) Unpaired t-test	2.60	$n_1 = n_2 = 100$	0.228	0.362	$n' = 303$
22	2010	1835449.1835536	(B) Paired t-test	1.56	$n' = 140$	0.132	0.342	$n' = 451$
23	2009	1571941.1571947	(B) Paired t-test	1.37	$n' = 48$	0.198	0.269	$n' = 203$
24	2012	2348283.2348343	(B) Paired t-test	0.95	$n' = 28$	0.180	0.152	$n' = 244$

Table 8: SIGIR papers with overpowered (power ≥ 0.99) and underpowered (power ≤ 0.50) ANOVA tests. Papers are sorted by achieved power.

#	year	doi (10.1145/*)	test type	F	actual sample size	effect size	achieved power	future sample size
1	2013	2484028.2484090	(E) Two-way ANOVA	$F_B = 68.01, m = n = 3$	$N = 82$	$\hat{f}_B = 1.365$	1	$N' = 18$
2	2015	2766462.2767746	(C) One-way ANOVA	$F = 243.42, m = 3$	$n = 2551$	$\hat{f} = 2.252$	1	$n' = 52$
3	2014	2600428.2609574	(C) One-way ANOVA	$F = 26.7m = 3$	$n = 12$	$\hat{f} = 1.272$	1	$n' = 4$
4	2014	2600428.2609596	(C) One-way ANOVA	$F = 56.52, m = 4$	$n = 1985$	$\hat{f} = 0.146$	1	$n' = 129$
5	2013	2484028.2484084	(C) One-way ANOVA	$F = 45.609, m = 2$	$n = 269$	$\hat{f} = 0.292$	1	$n' = 48$
6	2012	2348283.2348392	(C) One-way ANOVA	$F = 40, m = 7$	$n = 400$	$\hat{f} = 0.293$	1	$n' = 24$
7	2010	1835449.1835484	(C) One-way ANOVA	$F = 66.82, m = 5$	$n = 1100.2$	$\hat{f} = 0.221$	1	$n' = 51$
8	2010	1835449.1835486	(C) One-way ANOVA	$F = 31.77, m = 3$	$n = 173.3$	$\hat{f} = 0.351$	1	$n' = 28$
9	2014	2600428.2609620	(E) Two-way ANOVA	$F_B = 24.89, m = n = 2$	$N = 964$	$\hat{f}_B = 0.161$	0.999	$N' = 308$
10	2009	1571941.1572033	(D) Two-way ANOVA w/o replication	$F = 8.01, m = 4$	$n = 57$	$\hat{f}^2 = 0.143$	0.991	$n' = 35$
11	2008	1390334.1390362	(C) One-way ANOVA	$F = 1.28, m = 3$	$n = 12$	$\hat{f} = 0.279$	0.279	$n' = 43$
12	2015	2766462.2767719	(D) Two-way ANOVA w/o replication	$F = 0.63, m = 4$	$n = 17$	$\hat{f}^2 = 0.039$	0.183	$n' = 75$

4.6.1 Case Studies: Overpowered Experiments

There are a few extremely large sample sizes in the "overpowered" section of Table 7. The 12th entry in Table 7 (SIGIR 2014, 10.1145/2600428.2609617), a paper on personalisation from a search engine company, reported the difference in mean average precision (MAP) together with the standard error of the mean (SEM), which enabled us to compute the t-statistic by simply dividing the former with the latter. We chose the result with the largest difference ($MAP = 0.0224, SEM = 0.00140$) from one of their tables, which resulted in $t = 16.0$. The actual sample size (the number of impressions for averaging) was $n = 5352460$, but a recommended future sample size is $n' = 164107$. This situation is also visualised in Figure 7(a) (third balloon from the top). Note that the effect size is extremely small: $\hat{es} = 0.007$, although even such a small effect may possibly translate to a subtantial increase in profit for search engine businesses. The authors of the 17th entry in Table 7 (SIGIR 2015, 10.1145/2766462.2767712) are from academia, but they utilise commercial search engine logs, one of which contains 11,813,260 search sessions. As authors indicate that they used 30% of the above sessions for computing (mean) click entropy, we assumed that the sample size was $n = 3543978$ even though this exact number is not indicated in the paper. The authors report p-value $< 10^{-5}$, so the t statistic was computed as $t = t_{inv}(n - 1; 10^{-5}) = 4.42$, and the future sample size obtained is $n' = 1425634$. Again, the effect size is extremely small: $\hat{es} = 0.002$. As for the 11th entry in Table 7 (SIGIR 2014, 10.1145/2600428.2609602), one of its authors works for the aforementioned search engine company. The authors use the unpaired t-test with large sample sizes ($n_1 + n_2 = 193524$, from which we assume that $n_1 = n_2 = 96762$) and report a p-value of $1.5 * 10^{-12}$. However, the purpose of employing this test in their study was to quantify the association between two different user search behaviour features rather than to compare retrieval effectiveness: one feature is used to split the data into two samples, and then the other is used to compare the two samples in terms of the unpaired t-test.

There were also two SIGIR papers (not included in the aforementioned 75 papers that went through power analysis) for which the sample sizes for unpaired t-tests were too large for our power analysis tool: A SIGIR 2010 paper (10.1145/1835449.1835537) used a sample of some 95.8M query-URL pairs per group; A SIGIR 2007 (10.1145/1277741.1277771) used two samples where the total sample size was approximately 60M. These papers are from two different search engine companies: the latter paper is from the same company as the aforementioned 11th and 12th entries in Table 7. All of the above papers should be commended (not condemned!) for providing enough information in their papers for post hoc power anal-

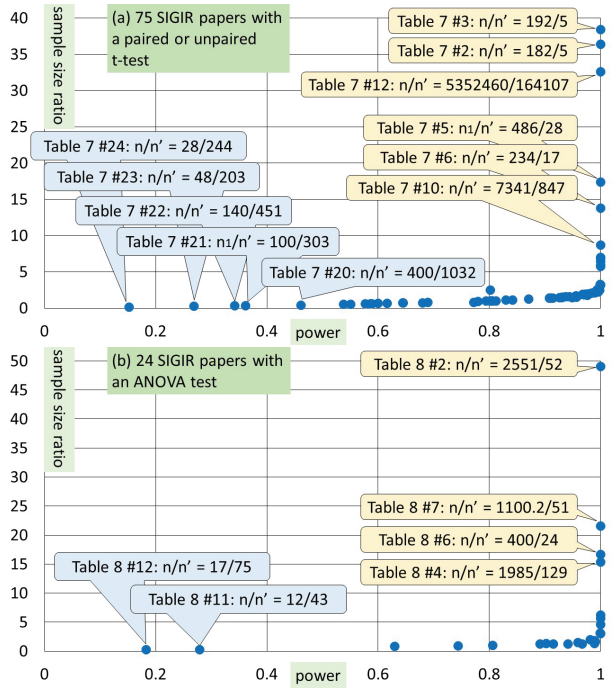

Figure 7: Summary of power analysis results: sample size ratio vs. achieved power.

ysis; however, recall that reporting p-values from an experiment with extremely large sample sizes without discussing effect sizes is generally not very informative, as was discussed in Section 1.

Next, let us examine a few overpowered *ANOVA* cases, using Table 8 and Figure 7(b). More specifically, we examine two papers indicated by the two top right balloons in Figure 7(b): The second and seventh entries in Table 8 (SIGIR 2015, 10.1145/2766462.2767746 and SIGIR 2010, 10.1145/1835449.1835484). These papers have a common first author, and the experiments in both papers utilise a commercial social media application suite for the purpose of item recommendation. In the second entry, a one-way ANOVA result is reported, where the number of groups is $m = 3$ (items rated "very interesting" vs. "interesting" vs. "not interesting"), the degree of freedom is $\phi = 7650$ and the F statistic is 243.42. Since our power analysis tool for one-way ANOVA recommends a per-group future sample size, we assume uniform group size for the analysis and let $\phi = m(n-1) = 7650$, which gives us the average group size $n = 2551$. As shown in Table 8 and Figure 7(b), this sample size is much larger than the required sample size $n' = 52$. Similarly, in the seventh entry, the authors report on a one-way ANOVA result with $m = 5$ (five recommendation systems), $\phi = 5496$. Again, if we assume equal group size for the purpose of power analysis, we obtain $n = 5496/5 + 1 = 1100.2$, whereas the future sample size is $n' = 51$. It can be observed in Table 8 that the effect size for the latter experiment is much smaller ($\hat{f} = 2.252$ vs. $\hat{f} = 0.221$). Again, let us emphasise that these are good papers which, unlike many other SIGIR papers, provide enough information for us to conduct post hoc power analysis. However, it is probably fair to say that many highly overpowered experiments come from industry, where data is abundant.

4.6.2 *Case Studies: Underpowered Experiments*

Arguably, extremely underpowered experiments may be more problematic than extremely overpowered experiments. Extremely overpowered experiments may conclude that some very small effects are statistically significant, and the small effects may or may not be *practically* significant. On the other hand, extremely underpowered experiments may hide away very important real differences forever.

First, let us have a look at two extremely underpowered paired t-test results indicated by the two leftmost balloons in Figure 7(a): the 24th and 23rd entries in Table 7. The 24th entry (SIGIR 2012, 10.1145/2348283.2348343) reports on many statistical significance results including ANOVA, but what we have selected for power analysis was a statistically insignificant result with a paired t-test, where $n = 28$ participants were involved (within-subjects design) and two systems were compared in terms of a *user experience sub-scale* called "focused attention." As Table 7 shows, the effect size is "small" ($\hat{es} = 0.180$) and the future number of participants is $n' = 244$, which is quite demanding for a user study. For this particular paper, we conducted additional power analyses for the other paired t-test results reported: "felt involvement" ($\hat{es} = 0.026$, $power = 0.052$, $n' = 12061$), "endurability" ($\hat{es} = 0.061$, $power = 0.061$, $n' = 2083$), "search effectiveness" ($\hat{es} = 0.111$, $power = 0.111$, $n' = 396$)[14]. It can be observed that the actual sample size ($n = 28$) was too small regardless of what user experience sub-scale is used, as the effect sizes are very small. The 23rd entry (SIGIR 2009, 10.1145/1571941.1571947) reports on a user study with 24 participants for comparing two algorithms (LAIR2 vs. Buckshot), but since each participants performed *two* tasks with each algorithm, we assumed that the sample size was $n = 48$ when the two algorithms were compared in terms of the F_1 measure. As Table 7 shows, $\hat{es} = 0.198$, $power = 0.269$, $n' = 203$. (Even if $n = 24$, the experiment is still underpowered: $\hat{es} = 0.280$, $power = 0.259$, $n' = 103$.) We would like to emphasise, however, that these papers are also examples of good papers, which provide enough details even for results that did not turn out to be statistically significant. This is what enables us to conduct post hoc studies, and the important question is how to design future experiments for similar studies.

Finally, we discuss the 12th and the 11th entry in Table 8, the two extremely underpowered cases indicated as the two leftmost balloons in Figure 7(b). The 12th entry (SIGIR 2015, 10.1145/2766462.2767719) reports on a statistically nonsignificant repeated-measures ANOVA result, where the relationship between four levels of search latency ($m = 4$) and Skin Conductance Responses (SCRs) were examined. This can be regarded as a two-way ANOVA without replication case, with $\phi_E = (m-1)(n-1) = 48$ and therefore the number of participants per group is $n = 17$. As can be seen at the bottom of Table 8, the effect size is $\hat{f}^2 = 0.039$, $power = 0.183$, $n' = 75$. Thus this experiment would have required 75 participants. The same paper also reports on statistically nonsignificant ANOVA results for examining the relationship between search latency and self-reported measures of engagement: if we apply the same R script to these results with $\phi_E = (m-1)(n-1) = 54, m = 4, n = 19$, we obtain $\hat{f}^2 = 0.042$, $power = 0.215$, $n' = 71$ for CSUQ (Computer System Usability Questionnaire), $\hat{f}^2 = 0.044$, $power = 0.222$, $n' = 68$ for FA (Focused Attention), $\hat{f}^2 = 0.052$, $power = 0.257$, $n' = 59$ for post-NAS (Negative Affect), and $\hat{f}^2 = 0.068$, $power = 0.332$, $n' = 46$ for post-PAS (Positive Affect). Thus, regardless of which ANOVA result we choose, relying on only $n = 17$ participants results in very low power, unfortunately. For the 11th entry (SIGIR 2008, 10.1145/1390334.1390362), we selected statistically nonsignificant one-way ANOVA results for comparing the

[14]For "perceived usability", the t statistic was zero, which suggests that there is no effect.

demographic characteristics of $m = 3$ participant groups, each with $n = 12$ subjects. As Table 8 shows, we would have wanted $n' = 43$ participants per group based on these particular results. However, these results do not constitute the main part of their study.

All of the "underpowered" papers discussed above are user study papers. This is not so surprising, as hiring good participants for experiments can always be difficult. Many such papers use statistical tests and report the results appropriately, but we argue that we should learn from past experiments as we have done in this study so that we can conduct better-designed experiments in the future.

5. CONCLUSIONS

We conducted a systematic review of 840 SIGIR full papers and 215 TOIS papers published between 2006 and 2015. Our main findings are as follows:

- Of the 862 papers that seem to deserve significance testing for comparison of means, 301 papers (35%) use the paired t-test; 255 papers (30%) lack significance testing (Figure 2).
- Of the 255 papers that lack significance testing, 167 papers (19%) make claims such as "significant improvement" and "significantly outperform" (Table 4).
- Of the 464 papers that report on a significance test for comparing two systems, 301 papers (65%) use the paired t-test; 97 papers (21%) use the Wilcoxon signed rank test (Figure 4).
- Compared to a decade ago, the proportion of papers that lack significance testing and the proportion of papers that rely on the Wilcoxon signed rank test have decreased; whereas, the proportion of papers that utilise the paired t-test has increased (Table 5). The differences are statistically significant according to two-proportions z-test.
- Of the 565 papers that report on significance test results, 365 papers (65%) report neither p-values nor test statistics (Figure 6).
- Of the 133 papers for which power analysis was possible, 38 papers (29%) were extremely overpowered ($power \geq 0.99$), while 9 papers (7%) were extremely underpowered ($power \leq 0.50$) (Table 6).
- We have observed extremely overpowered experiments in which proprietary data from industry (typically search engine companies) are utilised, as well as extremely underpowered user experiments in which the number of hired participants is limited.

To recap, p-values alone are not very informative as results of statistical significance testing, because one can obtain arbitrarily small p-values for any experiment by using a large enough sample [5, 14]. Hence, whenever researchers report on statistically significant differences based on overpowered experiments, it is vital that they report the effect sizes in addition to p-values. As for underpowered user experiments, researchers should conduct pilot studies first, or learn from similar studies in prior art about effect sizes and appropriate sample sizes.

One original question that is left unanswered in the present study is: are existing test collections with 50-100 topics good enough? Statistical requirements for comparing *any* systems suggest that test collection require many more topics [15, 20], but are current test collections actually serving the purpose for comparing *existing* systems? The question is left unanswered because, as we have seen above, many system effectiveness papers do not provide enough information for post hoc power analysis: recall that either an exact p-value or a test statistic (with a clearly stated sample size) is required for this. To make matters worse, there is the publication bias problem: researchers are often tempted not to report on statistically nonsignificant results. We already see some good reporting practices in interactive IR papers, complete with test statistics and effect sizes even for statistically nonsignificant results; we believe that similar practices are in order for the rest of the IR community as well.

Acknowledgement

We thank Professor Hideki Toyoda (Waseda University) for letting us modify his R code and distribute it.

6. REFERENCES

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of ACM CIKM 2009*, pages 601–610, 2009.

[2] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of ACM SIGIR 2012*, pages 105–114, 2012.

[3] B. Carterette. Bayesian inference for information retrieval evaluation. In *Proceedings of ACM ICTIR 2015*, pages 31–40, 2015.

[4] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (Second Edition)*. Lawrence Erlbaum Associates, 1988.

[5] P. D. Ellis. *The Essential Guide to Effect Sizes*. Cambridge University Press, 2010.

[6] S. Gerani, M. Carman, and F. Crestani. Aggregation methods for proximity-based opinion retrieval. *ACM TOIS*, 30(4), 2012.

[7] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):696–701, 2005.

[8] D. H. Johnson. The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63(3):763–772, 1999.

[9] D. Kelly and C. R. Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967-2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770, 2013.

[10] P. R. Killeen. An alternative to null hypothesis significance tests. *Psychological Science*, 16:345–353, 2005.

[11] Y. Nagata. *Introduction to Statistical Analysis (in Japanese)*. JUSE Press, 1992.

[12] S. E. Robertson and E. Kanoulas. On per-topic variance in IR evaluation. In *Proceedings of ACM SIGIR 2012*, pages 891–900, 2012.

[13] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.

[14] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

[15] T. Sakai. Topic set size design. *Information Retrieval Journal*, 2015.

[16] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of ACM SIGIR 2005*, pages 162–169, 2005.

[17] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33(4):495–512, 1997.

[18] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM CIKM 2007*, pages 623–632, 2007.

[19] H. Toyoda. *Introduction to Statistical Power Analysis: A Tutorial with R (in Japanese)*. Tokyo Tosyo, 2009.

[20] J. Urbano, M. Marrero, and D. Martín. On the measurement of test collection reliability. In *Proceedings of ACM SIGIR 2013*, pages 393–402, 2013.