# Statistical Reform in Information Retrieval?

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

**Abstract**

IR revolves around evaluation. Therefore, IR researchers should employ sound evaluation practices. Nowadays many of us know that statistical significance testing is not enough, but not all of us know exactly what to do about it. This paper provides suggestions on how to report effect sizes and confidence intervals along with $p$-values, in the context of comparing IR systems using test collections. Hopefully, these practices will make IR papers more informative, and help researchers form more reliable conclusions that "add up." Finally, I pose a specific question for the IR community: should IR journal editors and SIGIR PC chairs require (rather than encourage) reporting of effect sizes and confidence intervals?

## 1 Introduction

The objective of this paper is to initiate a discussion on better practices in reporting experimental results in IR, especially in the context of laboratory experiments using test collections. Early IR researchers were rather reluctant to use statistical significance tests (especially parametric tests); that has changed over the past decades, but many modern IR researchers who use test collections report only the $p$-values (or worse, we often just say "the difference is statistically significant at $\alpha = 0.05$"[1].

It is well-known that reporting $p$-values is not enough, because they reflect both the *sample size* (number of topics in our case) and the *effect size* (magnitude of the difference between systems) (e.g., [4, 7, 25]). For example, suppose that an IR researcher compared systems $X$ and $Y$ using a test collection with $n$ topics, and obtained per-topic performance differences $(d_1, \ldots, d_n) = (x_1 - y_1, \ldots, x_n - y_n)$ in terms of some evaluation measure. The *test statistic* for a paired $t$-test in this case would be

$$t_0 = \frac{\bar{d}}{\sqrt{V/n}} = \sqrt{n}\frac{\bar{d}}{\sqrt{V}} \quad (1)$$

where $\bar{d} = \sum_{j=1}^{n} d_j/n$ (the sample mean, which is the unbiased estimate of the population mean) and $V = \sum_{j=1}^{n}(d_j - \bar{d})^2/(n-1)$ (the unbiased estimate of the population variance). Now, as $t_0$ becomes larger (i.e., more extreme), the corresponding $p$-value becomes smaller and therefore the result is more likely to be considered statistically significant. However, it

[1] Or even worse: "the difference is *significant*."

is clear from Eq. 1 that a large $t_0$ may mean either (a) the sample size $n$ is large; or (b) the sample effect size $\bar{d}/\sqrt{V}$, the *difference between X and Y measured in standard deviation units*, is large. A $p$-value does not tell us which is the case. In short, discussing effect sizes along with $p$-values means trying to isolate the real difference from the sample size effect.

Outside the IR community, some top-tier journals now require reporting of effect sizes and *confidence intervals* along with $p$-values [4, 5, 6, 7, 13, 14]. Whereas, take the ACM TOIS *Information for Authors* page[2] as an example: as of May 14, 2014, it only mentions classifical significance testing: "*Statistical tests should be included to support empirical claims. When reporting statistics, the name of the statistic, the degrees of freedom, the value obtained, and the p-value should be reported, e.g.,* $F(3, 65) = 4.83, p < 0.01$." Thus, while effect sizes and confidence intervals may be nothing new to many IR researchers[3], it appears that the reporting practice has not really prevailed, at least in test collection-based studies. Hence, this paper is primarily intended for those who routinely use IR test collections to compare two or more systems and conduct statistical significance tests. I shall try to provide a tentative guideline for reporting effect sizes and confidence intervals in the above contexts. If the IR community standardise such reporting practices to some extent, this may help us form reliable and *practically significant*[4] conclusions through accumulation of informative results from different studies.

The remainder of this paper is organised as follows. Section 2 is a collection of quotations from past IR research, which reflects how the IR community slowly adopted (parametric) statistical significance testing, and how it has been questioned recently. Readers not interested in the history can skip this section and go directly to Section 3, which suggests exactly how to report effect sizes and confidence intervals for researchers trying to compare two or more IR systems using a test collection with $n$ topics. Section 4 discusses whether and how a "statistical reform" [5, 6] in IR should be brought about, as well as a few other possibilities.

# 2 Looking Back

In the early days, IR researchers were very cautious about the assumptions underlying (parametric) significance testing.

"*An attractive feature of the sign test is that normality of the input data is not required, and since this normality is generally hard to prove for statistics derived from a request-document correlation process, the sign-test probabilities may provide a better indicator of system performance than the t test.*" (Salton and Lesk, 1968 [18], p.15)

"*Parametric tests are inappropriate because we do not know the form of the underlying distribution.*" (Van Rijsbergen, 1979 [23], p.136)

"*Since the form of the population distributions underlying the observed performance values is not known, only weak tests can be applied; for example, the sign test.*" (Sparck Jones and Willet, 1997 [21], p.170)

If researchers are worried about the underlying assumptions such as normality, there are

[2] `http://tois.acm.org/authors.html`

[3] See, for example, Peter Bailey's SIGIR paper writing tip No.5: `http://research.microsoft.com/en-us/people/pbailey/sigir-paper-writing-tips.aspx`.

[4] "*A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important.*" [4]

distribution-free, computer-based alternatives to parametric significance testing, namely, the *bootstrap test* [16, 19] and the *randomisation test* [20, 22]. On the other hand, the modern view is that parametric tests such as the $t$-test and ANOVA are in fact applicable to many situations in IR, as these tests are known to be robust to assumption violations. The $t$-test, for example, is known to be highly consistent in practice with the distribution-free bootstrap and randomisation tests.

"*While the errors may not be normal, the t-test is relatively robust to many violations of normality. Only heavy skewness (lack of symmetry) or large outliers (observations very far from the mean) will seriously compromise its validity.*" (Hull, 1993 [8], p.334)

"*These simulation experiments suggest that if we are going to worry about assumptions, homoscedasticity and linearity rather than normality are the ones we should worry about, though even then the errors are small and may very well cancel out.*" (Carterette, 2012 [3], p.23)

But while the IR community began to accept parametric significance tests, the aforementioned limitations of significance testing began to receive more attention than before, and statistical reforms were introduced in several research disciplines outside IR [5, 6]. Carterette's recent remarks on significance testing in IR is in line with these developments:

"*With all of these questions, it is hard to escape the feeling that statistical inference is ultimately subjective, only providing a thin veneer of objectivity that makes us feel a little more comfortable about experimental rigor. That does not mean we should throw out the entire edifice — on the contrary, though we believe our analysis conclusively shows that a p-value cannot have any objective meaning, we still believe p-values from paired t-tests provide a good heuristic that is useful for many of the purposes they are currently used for. We only argue that p-values and significance test results in general should be taken with a very large grain of salt, and in particular have a limited effect on publication decisions and community-wide decisions about "interesting" research directions.*" (Carterette, 2012 [3], p.27)

As was mentioned earlier, $p$-values are considered uninformative (or even harmful) in some research disciplines [5, 6]. So what exactly can we do about it? This paper attempts to provide a tentative answer, by means of a draft guideline for reporting effect sizes and confidence intervals in IR experiments using test collections.

Before concluding this section, it is worth recalling a well-known fact that some early IR researchers did try to distinguish between statistical significance and *practical significance*. Discussing effect sizes is a more theoretically-grounded approach to measuring the latter.

"*It must nevertheless be admitted that the basis for applying significance tests to retrieval results is not well established, and it should also be noted that statistically significant performance differences may be too small to be of much operational interest.*" (Sparck Jones, 1981, p.243)

"*In a broad way, I shall characterise performance differences, assumed statistically significant, as interesting if they are at least noticeable, i.e. of the order of 5-10% different, and as rather more interesting if they are material, i.e. more than 10%.*" (Sparck Jones, 1974, p.5)

# 3 Looking Forward

This section provides a draft guideline for reporting effect sizes (ESs) and confidence intervals (CIs) along with $p$-values. We consider two typical cases: comparing two systems with a

paired $t$-test, and comparing $m(>2)$ systems with a two-way ANOVA without replication[5]. The former test should be applied when we have two systems $X$ and $Y$ and their per-topic performance differences $(d_1, \ldots, d_n) = (x_1 - y_1, \ldots, x_n - y_n)$; the latter test may be applied when we have $m(>2)$ systems with their per-topic performance scores $x_{ij}$ $(i = 1, \ldots, m,\ j = 1, \ldots, n)$. Note that when $m > 2$, conducting a $t$-test with a significance criterion $\alpha$ independently for every system pair results in a *familywise error rate* of $1 - (1 - \alpha)^{m(m-1)/2}$ [3, 4]: for example, if $m = 10$ and $\alpha = 0.05$, the probability that there is at least one Type I error (finding a nonexistent difference) amounts to $1 - 0.95^{45} = 0.90$. A better approach in this case would be to conduct ANOVA first to test the hypothesis that all of the $m$ systems are equally effective; if this null hypothesis is rejected, then a multiple comparison test such as the *randomised Tukey HSD test* [3, 17] can be applied, as discussed later[6].

## 3.1 Comparing Two Systems

### 3.1.1 Two-sided Paired $t$-test

Given $\{x_j\}$ and $\{y_j\}$, the per-topic performance scores for systems $X$ and $Y$ $(j = 1, \ldots, n)$, we assume that the scores are independent and that $x_j \sim N(\mu_X, \sigma_X^2)$ and $y_j \sim N(\mu_Y, \sigma_Y^2)$. Under these assumptions, $d_j = x_j - y_j \sim N(\mu, \sigma^2)$ where $\mu = \mu_X - \mu_Y$ and $\sigma^2 = \sigma_X^2 + \sigma_Y^2$, and $t = (\bar{d} - \mu)/\sqrt{V/n} \sim t(n-1)$ (i.e., $t$ distribution with $n-1$ degrees of freedom). Therefore, under the null hypothesis $H_0$ (namely, $\mu_X = \mu_Y$), $t_0 = \bar{d}/\sqrt{V/n} \sim t(n-1)$.

Given a significance criterion $\alpha$, we reject $H_0$ if $|t_0| \geq t(n-1;\alpha)$ where $t(\phi;\alpha)$ is the two-sided critical $t$ value with $\phi$ degrees of freedom for probability $\alpha$[7]. As was mentioned earlier, the actual $p$-value should be reported[8].

### 3.1.2 Reporting Effect Sizes

For the paired $t$-test (See Eq. 1), a useful effect size to report would be the sample effect size given by

$$ES_{pairedt} = \frac{|\bar{d}|}{\sqrt{V}} \ . \tag{2}$$

Recall that $\bar{d} = \sum_{j=1}^{n} d_j/n$ (the sample mean, which is the unbiased estimate of the population mean $\mu$) and $V = \sum_{j=1}^{n}(d_j - \bar{d})^2/(n-1)$ (the unbiased estimate of the population variance $\sigma^2$).

It should be noted that Eq. 2 is *not* an estimate of the population effect size given by $\mu/\sigma$: rather, it happens to be an estimate of a population effect size of the form $\mu/\sigma\sqrt{2(1-\rho_{XY})}$, where $\rho_{XY}$ denotes the population correlation coefficient between $X$ and $Y$ [14]. Hence, effect sizes for paired $t$-tests as defined by Eq. 2 should not be compared directly with effect sizes for *unpaired* tests such as *Cohen's d* [4]. If all IR test collection users stick to the paired test and Eq. 2 wherever appropriate, there should be no problem.

[5] "without replication" means that, for every system-topic pair, there is exactly one measurement, which is usually true in experiments based on test collections.

[6] Ellis [4] remarks that the approach of adjusting $\alpha$ (e.g., Bonferroni correction) "*may be a bit like spending $1,000 to buy insurance for a $500 watch*".

[7] With Microsoft Excel, `TINV`$(\alpha, \phi)$ or `T.INV.2T`$(\alpha, \phi)$.

[8] With Microsoft Excel, the $p$-value may be obtained by `TDIST`$(t_0, \phi, 2)$ or `T.DIST.2T`$(t_0, \phi)$.

### 3.1.3 Reporting Confidence Intervals

Since we have assumed that $t = (\bar{d} - \mu)/\sqrt{V/n} \sim t(n-1)$, it follows that

$$Pr[-t(n-1;\alpha) \leq t \leq t(n-1;\alpha)] = 1 - \alpha \ , \tag{3}$$

which can be rewritten as

$$Pr[\bar{d} - ME \leq \mu \leq \bar{d} + ME] = 1 - \alpha \ , \tag{4}$$

where the *margin of error ME* is given by:

$$ME = t(n-1;\alpha)\sqrt{V/n} \ . \tag{5}$$

Thus, using Eq. 5, a $100(1-\alpha)\%$ CI for $\mu$ is given by $[\bar{d} - ME, \bar{d} + ME]$.

## 3.2 Comparing $m(> 2)$ Systems

### 3.2.1 Two-way ANOVA without replication

Suppose we have $m$ systems evaluated with $n$ topics; we have the performance scores $x_{ij}$ ($i = 1, \ldots, m$ and $j = 1, \ldots, n$) which we assume to be independent. We furthermore assume that $x_{ij}$ can be modeled as $x_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $\sum_{i=1}^{m} a_i = 0, \sum_{j=1}^{n} b_i = 0$. Here, $\mu$ is the grand mean of the population, $a_i$ is the system effect, $b_i$ is the topic effect, and $\varepsilon_{ij}$ is the residual[9]. Note that we assume a common variance $\sigma^2$ across the $m$ systems (just like when we conduct an *unpaired* $t$-test where the variance is unknown): this is called the *homoscedasticity* assumption.

From the observed values $\{x_{ij}\}$, we first compute the grand mean of the sample $\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$, the system means $\bar{x}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$, and the topic means $\bar{x}_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$. Then we compute the following *sum of squares*:

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 \ , \tag{6}$$

$$S_A = n\sum_{i=1}^{m}(\bar{x}_{i\bullet} - \bar{x})^2 \ , \quad S_B = m\sum_{j=1}^{n}(\bar{x}_{\bullet j} - \bar{x})^2 \ , \tag{7}$$

$$S_E = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2 = S_T - S_A - S_B \ . \tag{8}$$

Note that $S_T$ measures how the observed values differ from the grand mean, while $S_A$ measures how each system mean differs from the grand mean; $S_E$ represents what is left after removing the between-system variation $S_A$ and the between-topic variation $S_B$ from the total variation.

Our null hypothesis $H_0$ for the system effect is that $a_1 = \ldots = a_m = 0$, i.e., that all systems are equivalent. Let the *mean squares* be $V_A = S_A/\phi_A, V_B = S_B/\phi_B, V_E = S_E/\phi_E$ where $\phi_A = m - 1, \phi_B = n - 1, \phi_E = (m-1)(n-1)$. Then under $H_0$, it is known that

$$F_0 = V_A/V_E = (n-1)S_A/S_E \sim F(\phi_A, \phi_E) \tag{9}$$

[9] In Two-way ANOVA without replication, the system-topic interaction cannot be separated from the residual.

where $F(\phi_A, \phi_E)$ is the $F$ distribution with $(\phi_A, \phi_E)$ degrees of freedom. Hence, we reject $H_0$ if $F_0 \geq F(\phi_A, \phi_E; \alpha)$, where $F(\phi_A, \phi_E; \alpha)$ is the critical $F$ value (with $\phi_A$ and $\phi_E$) for probability $\alpha$[10]. It is clear from Eq. 9 that a large $F_0$ may imply either (a) a large sample size $n$; or (b) that the between-system variation $S_A$ is substantially higher than $S_E$, which suggests that not all systems are equal[11].

If $H_0(a_1 = \ldots = a_m = 0)$ is rejected, then a Tukey HSD (Honestly Significant Differences) test or its computer-based randomised version [3, 17] can be applied to find exactly which pairs are statistically significantly different, while keeping down the familywise error rate to $\alpha$[12]. For systems $X$ and $Y$ whose difference in means is given by $\bar{d}_{XY}$, the classical Tukey HSD test rejects the null hypothesis that $X$ and $Y$ are equivalent if $|\bar{d}_{XY}|/\sqrt{V_E/n} \geq q(m, (m-1)(n-1); \alpha)$, where $q(m, (m-1)(n-1); \alpha)$ denotes the critical value of the *studentised range distribution* with $(m, (m-1)(n-1))$ degrees of freedom at $\alpha$, which needs to be obtained by a table lookup. Note that $V_E$ can be obtained from the ANOVA. Whereas, the *randomised* Tukey HSD is distribution-free and can be automatically computed for any degrees of freedom.

### 3.2.2 Reporting Effect Sizes

With ANOVA, the population effect sizes often discussed are:

$$\eta^2 = \frac{\sigma_A^2}{\sigma_T^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_E^2} \ , \tag{10}$$

$$\eta_p^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \tag{11}$$

where $\sigma_T^2, \sigma_A^2, \sigma_B^2, \sigma_E^2$, are the unknown population variances that correspond to the sample *variations* $S_T, S_A, S_B, S_E$, respectively. Thus, $\eta^2$ represents "how much of the total variance can be explained by the between-system variance," while the *partial* effect size $\eta_p^2$ represents the same quantify *after* removing the between-topic variance. The above effect sizes are often estimated from samples as $\hat{\eta}^2 = S_A/S_T$ and $\hat{\eta}_p^2 = S_A/(S_A + S_E)$, but it is known that $\hat{\eta}^2$ and $\hat{\eta}_p^2$ are positively biased: they heavily overestimate $\eta^2$ and $\eta_p^2$ especially when the sample size is small [14]. (Weren't we trying to isolate the effect size from the sample size?) For two-way ANOVA without replication, more accurate estimates of $\eta^2$ and $\eta_p^2$ can be obtained as [7, 14, 15]:

$$\omega^2 = \frac{\phi_A(V_A - V_E)}{S_T + V_B} \ , \tag{12}$$

$$\omega_p^2 = \frac{\phi_A(V_A - V_E)}{S_A + (n - \phi_A)V_E} \ . \tag{13}$$

The above effect sizes measure the system effect as a whole in the context of ANOVA. However, what IR researchers are often more interested in is the Tukey HSD result which shows exactly which of the systems are statistically significantly different. When researchers

[10] With Microsoft Excel, `FINV`$(\alpha, \phi_A, \phi_E)$ or `F.INV.RT`$(\alpha, \phi_A, \phi_E)$. The $p$-value for $F_0$ can be obtained as `FDIST`$(F_0, \phi_A, \phi_E)$ or `F.DIST.RT`$(F_0, \phi_A, \phi_E)$.

[11] The topic effect may also be tested in a similar way: the null hypothesis in this case is that $b_1 = \ldots = b_n = 0$, i.e., that all systems are equivalent, and $F_0 = (m-1)S_B/S_E \sim F(\phi_B, \phi_E)$.

[12] An implementation of Carterette's randomised Tukey HSD test is available from `http://www.f.waseda.jp/tetsuya/tools.html`.

report on the $p$-values of the Tukey HSD test (or its randomised version), one is tempted to report effect sizes of the form similar to Eq. 2, that is not directly affected by the sample size and measures the difference in standard deviation units. Since the test statistic for the classical Tukey HSD test is given by $\bar{d}_{XY}/\sqrt{V_E/n}$, a useful sample effect size might be:

$$ES_{HSD} = |\bar{d}_{XY}|/\sqrt{V_E} \ . \quad (14)$$

Recall that $V_E$ is the *remainder* mean square, obtained after removing the between-system and between-topic variations.

As there are several effect size estimates available for ANOVA [7, 15], it is important for us to clarify exactly *which* effect sizes we are reporting in our papers. Moreover, we should provide the basic statistics such as $S_A$ and $S_E$ in our papers so that other researchers can compute their favourite effect sizes from our data. Of course, making the raw data publicly and permanently available to the community would solve a lot of problems.

### 3.2.3 Reporting Confidence Intervals

In the context of ANOVA where we accept the homoscedasticiy assumption, the following margin of error can be used to obtain a CI for every system [9]:

$$ME = t(\phi_E; \alpha)\sqrt{V_E/n} \ . \quad (15)$$

Thus, for the $i$-th system, its $100(1-\alpha)\%$ CI is given by $[\bar{x}_{i\bullet} - ME, \bar{x}_{i\bullet} + ME]$. Note that the above gives a tigher margin compared to Eq. 5.

Finally, it is important to visualise CIs by means of *error bars* on graphs. Every graph that shows mean statistics should contain error bars (that represent CIs)[13]. With Microsoft Excel, it is very easy to add error bars to graphs, with user-specified ME values[14].

## 3.3 Examples

### 3.3.1 Example: Comparing Two Systems

Table 1: Example: per-topic performances and differences for systems $X$ and $Y$ ($n = 10$).

| Topic ID | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| System X | 0.39 | 0.28 | 0.31 | 0.21 | 0.19 | 0.64 | 0.75 | 0.36 | 0.66 | 0.54 | 0.43 |
| System Y | 0.27 | 0.04 | 0.18 | 0.08 | 0.19 | 0.54 | 0.57 | 0.28 | 0.20 | 0.40 | 0.28 |

Suppose we have obtained the results shown in Table 1. Then $\bar{d} = 0.16, V = 0.015, t_0 = \bar{d}/\sqrt{V/n} = 4.1$ so the $p$-value is $\texttt{TDIST}(4.1, 10-1, 2) < 0.0029$. The sample effect size is $ES_{pairedt} = \bar{d}/\sqrt{V} = 1.3$. Also, if $\alpha = 0.05$, then $ME = \texttt{TINV}(0.05, 10-1) * \sqrt{V/n} = 0.088$, so the 95% CI for the difference is $0.16 \pm 0.088$. These results may be reported as follows:

"According to a two-sided paired $t$-test for the difference in means $\bar{d} = 0.16$ (with the unbiased estimate of the population variance $V = 0.015$), System $X$ statistically significantly outperforms $Y$ ($t(9) = 4.1, p < 0.0029, ES_{pairedt} = 1.3$, 95% CI $[0.07, 0.25]$)."

[13] Error bars are sometimes used to represent standard deviations or standard errors, so it is important to clearly indicate what the bars represent.

[14] See, for example, `http://office.microsoft.com/en-us/excel-help/add-error-bars-to-a-chart-HA102840044.aspx` .

### 3.3.2 Example: Comparing $m(>2)$ Systems

Table 2: Example: per-topic performances for systems $X, Y, Z$ ($m = 3, n = 10$).

| Topic ID | 01 | 02 | 03 | 04 | 05 | $\bar{x}_{i\bullet}$ |
|---|---|---|---|---|---|---|
| System X | 0.40 | 0.44 | 0.42 | 0.40 | 0.39 | 0.41 |
| System Y | 0.35 | 0.40 | 0.40 | 0.39 | 0.40 | 0.39 |
| System Z | 0.35 | 0.40 | 0.37 | 0.38 | 0.39 | 0.38 |

Table 3: Two-way ANOVA (without replication).

| | Sum of squares | Degrees of freedom | Mean squares | $F_0$ |
|---|---|---|---|---|
| Between-system | $S_A = 0.0027$ | $\phi_A = 2$ | $V_A = 0.0013$ | 6.8 |
| Between-topic | $S_B = 0.0034$ | $\phi_B = 4$ | $V_B = 0.00084$ | 4.3 |
| Within | $S_E = 0.0016$ | $\phi_E = 8$ | $V_E = 0.00020$ | – |

Suppose we have obtained the results shown in Table 2. The grand mean is then $\bar{x} = 0.39, S_A = n\sum_{i=1}^{m}(\bar{x}_{i\bullet} - \bar{x})^2 = 0.0027, S_B = m\sum_{j=1}^{n}(\bar{x}_{\bullet j} - \bar{x})^2 = 0.0034, S_T = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{ij} - \bar{x})^2 = 0.0076, S_E = S_T - S_A - S_B = 0.0016, V_A = S_A/\phi_A = 0.0027/(3-1) = 0.0013, V_B = S_B/\phi_B = 0.0034/(10-1) = 0.00084, V_E = S_E/\phi_E = 0.0016/(3-1)(10-1) = 0.00020$. Therefore, $F_0 = (n-1)S_A/S_E = 6.76$, and the $p$-value is $\texttt{FDIST}(6.76, \phi_A, \phi_E) < 0.020$. (The topic effect can be tested in a similar way.) The population effect sizes for the ANOVA can be estimated as $\omega^2 = 0.27, \omega_p^2 = 0.70$ (Eqs. 10 and 11). According to a randomised Tukey HSD test, only the difference between $X$ and $Z$ is found to be statistically significant ($p = 0.029$), and $ES_{HSD} = \bar{d}_{XZ}/\sqrt{V_E} = 0.022/\sqrt{0.00020} = 1.6$. As for the common margin of error, if $\alpha = 0.05$, $ME = \texttt{TINV}(0.05, \phi_E) * \sqrt{V_E/n} = 0.015$. This is then applied to the system means $0.41, 0.39, 0.38$ to obtain the CIs. These results may be reported as follows:

"Table 3 shows the result of a two-way ANOVA (without replication) applied to the comparison of $m = 3$ systems with $n = 5$ topics. The system effect is statistically significant ($F(2, 8) = 6.8, p < 0.020$)[15]. The population effect size and the *partial* population effect size for the ANOVA can be estimated from Table 3 as $\omega^2 = \phi_A(V_A - V_E)/(S_T + V_B) = 0.27$ and $\omega_p^2 = \phi_A(V_A - V_E)/(S_A + (n - \phi_A)V_E) = 0.70$. A randomised Tukey HSD test shows that only the difference between $X$ and $Z$ is statistically significant ($p = 0.029, ES_{HSD} = 0.022/\sqrt{V_E} = 1.6$). Figure 1 shows the mean performances of the three systems with 95% CIs using the same $V_E$ from Table 3."

# 4 Summary

Following the practices in some research disciplines outside IR, this paper provided a draft guideline for reporting experimental results in IR with $p$-values, effect sizes and confidence intervals. I do not claim that this is *the* right way to report results, but I argue that it is important to (a) make papers as informative as possible, for example, by separating the effect size from the sample size and by reporting basic statistics such as $V$ (Section 3.1.2), $S_A, S_E$ (Section 3.2.1); and (b) share some basic reporting practices across IR researchers to facilitate *meta-analysis* [4] and to form conclusions that "add up" [1].

[15]The topic effect is also statistically significant ($F(4, 8) = 4.3, p < 0.039$).

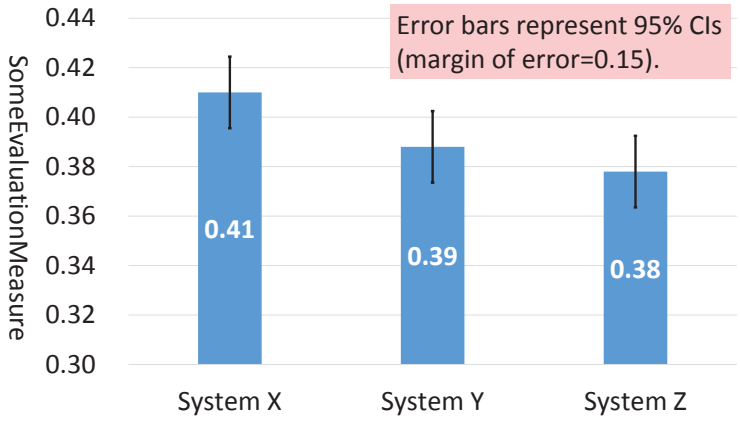


Figure 1: Comparison of systems $X$, $Y$ and $Z$.

According to Fidler *et al.* [5, 6] who have observed the (slow progress of) statistical reforms in medicine, psychology and ecology over the past decades, it is extremely hard to bring about statistical reforms. They conclude that "*requirements may be more effective than recommendations.*" To IR researchers, this translates to: "*Maybe IR journal editors and SIGIR PC chairs should* require *(rather than encourage) reporting of effect sizes and confidence intervals?*" I think they should.

Another important point to remember is that reporting effect sizes and confidence intervals is not an end in itself. Based on the effect size estimates, researchers should compare the impact of their research with others, and wherever possible discuss whether their results may be *practically* significant.

Finally, there are alternatives to classical hypothesis testing. In psychology, Killeen [12] recently proposed $p_{rep}$ to replace $p$-values: it represents the probability that a replication of a study would give a result in the same direction as the original study, which to some of us may sound reminiscent of the topic-splitting approach of Zobel [26] and Voorhees and Buckley [24]. Another alternative is the Bayesian approach to hypothesis testing [2, 10, 11], which was first conceptualised in 1935. Will "new" approaches such as these eventually replace classical significance testing in IR?

# References

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 19998. In *Proceedings of ACM CIKM 2009*, pages 601–610, 2009.

[2] B. Carterette. Model-based inference about IR systems. In *ICTIR 2011 (LNCS 6931)*, pages 101–112, 2011.

[3] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.

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

[5] F. Fidler and G. Cumming. Lessons learned from statistical reform efforts in other disciplines. *Psychology in the Schools*, 44(5):441–449, 2007.

[6] F. Fidler, C. Geoff, B. Mark, and T. Neil. Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics*, 33:615–630, 2004.

[7] R. J. Grissom and J. K. John. *Effect Sizes for Research: Univariate and Multivariate Applications (Second Edition)*. Routledge, 2012.

[8] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of ACM SIGIR 1993*, pages 329–338, 1993.

[9] S. Ishimura. *Analysis of Variance (Ninth Edition) (in Japanese)*. Tokyo Tosho, 1999.

[10] V. E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19313–19317, 2013.

[11] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

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

[13] R. Nuzzo. Statistical errors. *Nature*, 506(13):150–152, 2014.

[14] M. Okubo and K. Okada. *Psychological Statistics to Tell Your Story: Effect Size, Confidence Interval (in Japanese)*. Keiso Shobo, 2012.

[15] S. Olejnik and J. Algina. Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Comtemporary Educational Psychology*, 25:241–286, 2000.

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

[17] T. Sakai. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, 2014.

[18] G. Salton and M. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15:8–36, 1968.

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

[20] 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.

[21] K. Sparck Jones and P. Willet, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[22] J. Urbano, M. Marrero, and D. Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM SIGIR 2013*, pages 925–928, 2013.

[23] C. J. van Rijsbergen. *Information Retrieval (Second Edition)*. Butterworths, 1979.

[24] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR 2002*, pages 316–323, 2002.

[25] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proceedings of ACM CIKM 2008*, pages 571–580, 2008.

[26] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR 1998*, pages 307–314, 1998.