How do Metric Score Distributions affect the Type i Error Rate of Statistical Significance Tests in Information Retrieval?

Urbano, Julián; Corsi, Matteo; Hanjalic, Alan

DOI
10.1145/3471158.3472242

Publication date
2021

Document Version
Final published version

Published in
ICTIR 2021

How do Metric Score Distributions affect the Type I Error Rate of Statistical Significance Tests in Information Retrieval?

Julián Urbano
Delft University of Technology
Delft, The Netherlands
j.urbano@tudelft.nl

Matteo Corsi
Delft University of Technology
Delft, The Netherlands
m.corsi@tudelft.nl

Alan Hanjalic
Delft University of Technology
Delft, The Netherlands
a.hanjalic@tudelft.nl

ABSTRACT

Statistical significance tests are the main tool that IR practitioners use to determine the reliability of their experimental evaluation results. The question of which test behaves best with IR evaluation data has been around for decades, and has seen all kinds of results and recommendations. Definitive answer to this question has recently been attempted via stochastic simulation of IR evaluation data, allowing researchers to compute actual Type I error rates because they can control the null hypothesis. One such research line simulates metric scores for a fixed set of systems on random topics, and concluded that the t-test behaves the best. Another such line simulates retrieval runs by random systems on a fixed set of topics, and concluded that the Wilcoxon test behaves the best. Interestingly, two recent surveys of the IR literature have shown that the community has a clear preference precisely for these two tests, so further investigation is critical to understand why the above simulation studies reach opposite conclusions. It has been recently postulated that a reason for the disagreement is the distributions of metric scores used by one of these simulation methods. In this paper we investigate this issue and extend the argument to another key aspect of the simulation, namely the dependence between systems. Following a principled approach, we analyze the robustness of statistical tests to different factors, thus identifying under what conditions they behave well or not with respect to the Type I error rate. Our results suggest that differences between the Wilcoxon and t-test may be explained by the skewness of score differences.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Statistical significance, Simulation, Type I error, Skewness

ACM Reference Format:
Julián Urbano, Matteo Corsi, and Alan Hanjalic. 2021. How do Metric Score Distributions affect the Type I Error Rate of Statistical Significance Tests in Information Retrieval?. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3471158.3472242

ICTIR '21, July 11, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8611-1/21/07.
https://doi.org/10.1145/3471158.3472242

1 INTRODUCTION

The Information Retrieval (IR) practice heavily relies nowadays on statistical significance tests to report the reliability of test collection based experimental results [2, 15]. However, IR evaluation data do not comply with typical assumptions made by these tests, such as the assumption of normally distributed scores behind the t-test. The natural question of which tests should be used is one that triggered early discussion [9, 17, 20, 27]. The follow-up question of which tests actually perform best with IR data is one for which several experimental works have been published mostly in the past two decades [6, 16, 17, 20, 21, 25, 28–30], with highly conflicting recommendations. Two recent and parallel lines of work have pointed to limitations in how these experimental works approached the problem [13, 14, 24, 26]. The gist is mainly in that the data they use are limited by the dozens of systems and topics in the TREC archive, with no control over the null hypothesis. As an alternative, they proposed stochastic simulation frameworks that allow us to *generate* IR-like data and study how well statistical tests behave by computing actual Type I and Type II errors.

On the one hand, Urbano [23] and Urbano and Nagler [26] developed a simulation framework that builds a model for the joint distribution of effectiveness scores of a set of systems. The model contains two parts: the marginal distribution of each system (ie. their distribution regardless of other systems), and a copula [12] to model the dependence among systems (ie. how they tend to behave for the same topic). Given this model, they may simulate effectiveness scores on new random[1] topics for the same systems. In a later work, Urbano et al. [24] used this simulation framework to compute Type I error rates for a range of tests and under different conditions. Besides other results, they find that i) the t-test and permutation test maintain error rates at the α level remarkably well; ii) the bootstrap test is biased towards small p-values, but large sample sizes tend to correct the bias; and iii) both the Sign and Wilcoxon tests have high error rates, but large sample sizes actually tend to increase the bias.

On the other hand, Parapar et al. [14] developed a simulation framework that builds a model for the retrieval score distribution [10] of a system and a topic. The model consists in a mixture of distributions for relevant and non-relevant documents. Given this model, they simulate new random runs[2] for the same topic. In a follow-up work, Parapar et al. [13] simulate runs from a logistic model that captures the relationship between document ranks and relevance. Besides other results, they find that i) the Wilcoxon and permutation tests maintain error rates at the α level remarkably

[1] Note that the term "random" here means "stochastic", that is, generated from a probabilistic model. It does *not* mean that it is uninformative or uniformly random.
[2] Similarly, the term "random" refers here to a run generated from a stochastic model.

well; ii) the t-test and Sign test have a medium bias towards high p-values, and iii) the bootstrap test has a very high bias towards high p-values.

It is surprising that these two works, based on the same principle of simulating data with control of the null hypothesis, reach shockingly opposite conclusions, specially with regard to the t-test and Wilcoxon test. These tests turn out to be the most popular in the IR literature [2, 15], so further investigation is critical to understand *why* these simulation studies reach opposite conclusions. Parapar et al. [13] postulate that the results by Urbano et al. [24] are caused by "a fundamental limitation of [their] approach", namely "if simulated models are fitted from pre-selected classes of distributions, the comparison is biased towards significance tests that follow certain parametric assumptions [...] such pre-selection of certain parametric distributions is an artifact of the simulation". Paraphrasing, if data are simulated from a model that aligns with one of the tests, of course that test will behave better than others. Unfortunately, this point was not analyzed or verified, so in this paper we explore the issue and extend the argument also to the copula families, that is, perhaps some of them benefit some tests more than others. Shedding light on this matter is important not only to find out if one simulation method is more appropriate than the other, but also to gain more understanding regarding how different factors such as score distributions and system dependencies may affect the behavior of significance tests when used with *real* IR evaluation data.

In this paper we therefore present an *exploratory* but principled investigation of the data generated by Urbano et al. with regard to Type I errors, and in particular study the behavior of tests across distribution families, copula families, and sample set sizes. Our results show that differences across tests do not appear to be caused by the metric score distribution families as suggested by Parapar et al., but rather by the different degrees of skewness induced by the dependence between system.

2 DATA

Figure 1 shows how Urbano et al. [24] simulated effectiveness data under the null hypothesis to compute Type I error rates. To fit the model (red flow), a baseline system B and an experimental system E are randomly chosen and their marginal distributions are estimated (ie. the distribution of metric scores, regardless of other systems). From these distributions, the so-called pseudo-observations U and V are computed and used to estimate a copula that models their dependence (ie. how they tend to behave for the same topic). To simulate scores on a random topic (blue flow), new pseudo-observations U^* and V^* are generated from the copula, and they are transformed into effectiveness scores B^* and E^*. Because the same distribution F_B is used in the last step, both systems have the same expected value and the null hypothesis holds; a test yielding $p \leq \alpha$ is thus making a Type I error.

For the margins, they considered 3 parametric distributions (Truncated Normal, Beta and Beta-Binomial) and 6 non-parametric through Kernel Smoothing and various kernels (Truncated Normal, Beta, and Discrete with 4 degrees of smoothness) [24, 26], and chose in each case the distribution that best described the data according to the Akaike Information Criterion (AIC). For the dependence, they

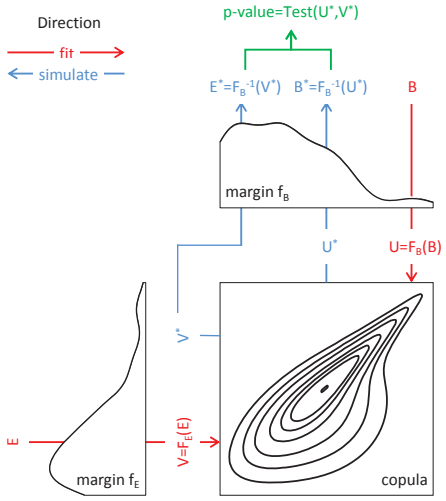

Figure 1: Stochastic simulation model used by Urbano et al. [24] (figure adapted from theirs). Note that the model is fitted (red) with two different margins, but data are simulated (blue) with only one, so that the null hypothesis holds.

considered 11 parametric copulas (Gaussian, Student t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, BB8 and Tawn) plus their 3 rotations [24, 26], and similarly chose the best fit based on the AIC.

Urbano et al. created just over $50K$ such stochastic models from the 363 TREC 5–8 Ad hoc runs and the 228 TREC 2010–2013 Web runs, across five effectiveness metrics (*AP*, *nDCG*@20, *ERR*@20, *P*@10 and *RR*), making $25M$ simulations across three topic set sizes (25, 50 and 100). They studied the paired t-test, Wilcoxon Signed Rank, Sign, bootstrap-shift and Permutation tests, leading to a total of $125M$ 2-tailed p-values. These are the data we analyze in this paper, which are publicly available from the Github repository linked from their paper[3].

3 ANALYSIS

The data described in Section 2 allow us to study how different factors affect the Type I error rate of the tests, namely the effectiveness metric and the score distributions it produces (ie. the margins), the dependence between systems (ie. the copula), and the sample size. Figure 2 shows that some distribution families and copula families are chosen more often than others[4]. This indicates that some models describe actual IR data better than others, so in principle it seems like a good idea to consider many different families and choose the best one, as Urbano et al. [24] did. In addition, the diversity of choices indicates that we should actually consider as many families as possible. Still, as argued by Parapar et al. [13], maybe some families favor some significance tests more than others, so when using simulated data to study their behavior the comparison might not be fair.

[3] https://github.com/julian-urbano/sigir2019-statistical/

[4] We do not report *nDCG*@20 and *ERR*@20 because results are very similar to those of *AP*. Likewise, we will not report on the permutation and Sign tests because they are very similar to the t-test and Wilcoxon test, respectively. Full results, along with all data and code, are available online at https://github.com/julian-urbano/ictir2021-metric

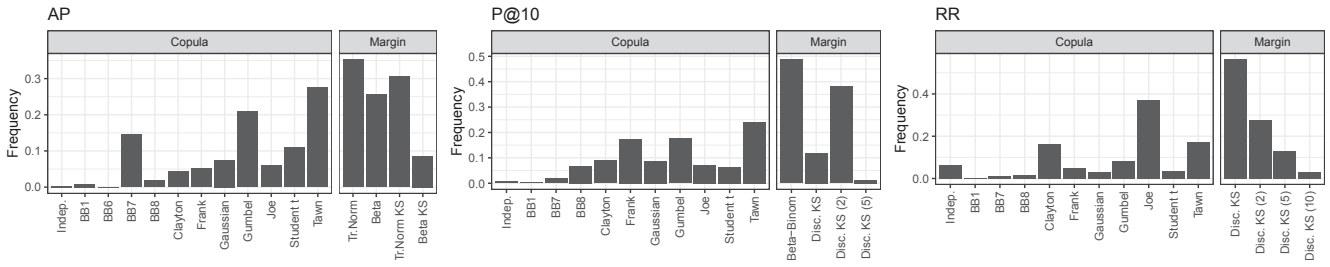

Figure 2: Distribution of copula and marginal distribution families used in the simulation models.

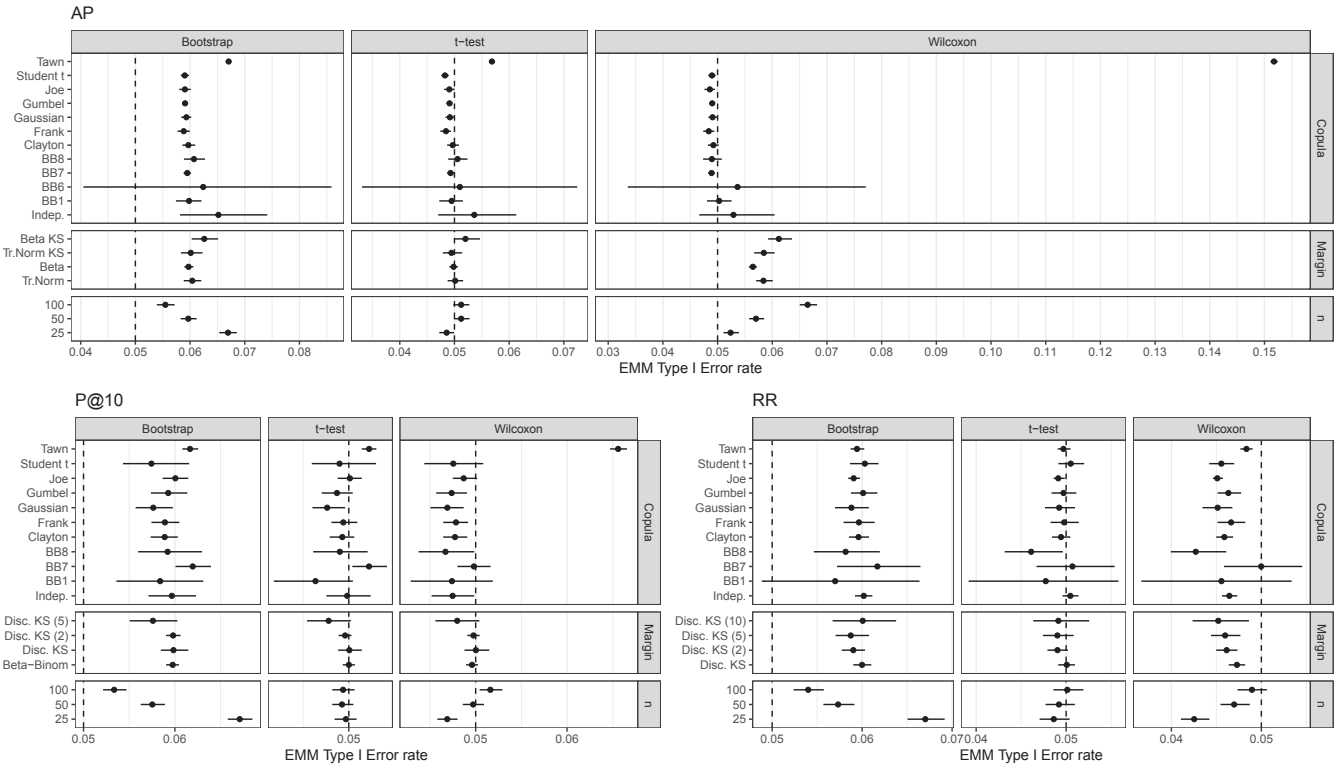

Figure 3: Estimated marginal Type I error rates for the copula, margin and sample size factors, along with 95% BC bootstrap confidence intervals. The vertical dashed lines mark the ideal rate $\alpha = 0.05$. Note that, for the same metric, the x-axes have the same scale; panels have different widths because they cover different ranges.

To answer this question, and contrary to Urbano et al. [24], we do not report on the observed error rates across each of the factors. Figure 2 showed that the data are highly imbalanced, so simply reporting the observed rates would lead to confounded effects [11]. For example, the Wilcoxon test had an observed error rate of 0.074 for *AP* and the Truncated Normal margin. However, as many as 30% of those cases came from a Tawn copula, which means that the 0.074 error rate may largely be due to the *combination* of Truncated Normal margin and Tawn copula.

Descriptive results would therefore be biased towards popular models, so instead we look at Estimated Marginal Means (EMM) [19]. In particular, for each effectiveness metric we calculate the observed Type I error rate for every combination of sample size, margin family and copula family. We then fit a linear model for the error rate including these three factors and their two-factor interactions. From this model, we compute the EMMs for each factor, which estimate the error rate for each case *while controlling* for the effect of the other factors, that is, it removes confounding. For instance, the EMM error rate for the Truncated Normal margin is 0.058, while it is 0.152 for the Tawn copula. Such an analysis would therefore suggests that the cause of the high error rate in the above example (ie. the observed 0.074) does not seem to be the margin, but the copula. To further reduce noise in our analysis, we compute 95% bias-corrected bootstrap intervals [7] around the EMMs, based on 1, 000 bootstrap samples.

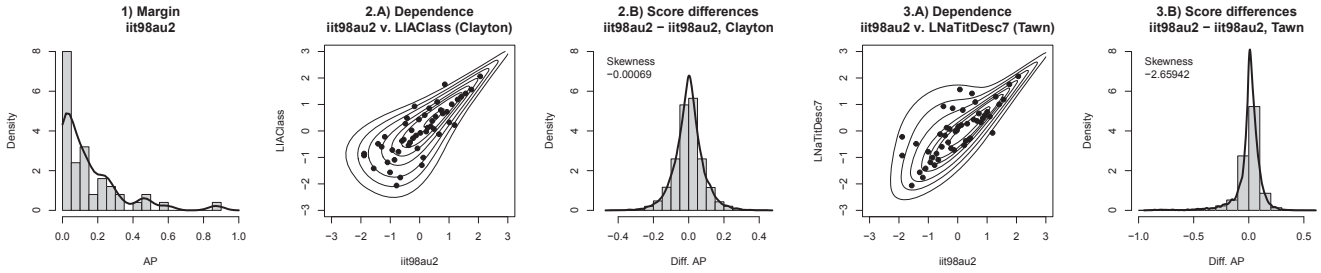

Figure 4: Illustration of how copula asymmetricity affects the skewness of metric score differences. 1) Distribution of *AP* scores of the TREC 7 system `iit98au2`. 2.A) Copula fitted for the dependence with system `LIAClass`; the best fit is a Clayton copula (symmetric). 2.B) Distribution of per-topic *AP* differences using the copula in 2.A with the margin in 1; the distribution is *not* skewed. 3.A) Same as 2.A but with system `LNaTitDesc7`; the best fit is a Tawn copula (asymmetric). 3.B) Same as 2.A but with the copula in 3.A; the distribution *is* skewed. By construction, both 2.B and 3.B comply with the null hypothesis because both simulated systems have the same marginal distribution in 1.

Figure 3 shows the Estimated Marginal Mean Type I error rates for each of the factors, where the dashed vertical lines mark the ideal error rate at the $\alpha = 0.05$ level. If we first turn our attention to the effect of the topic set size n, we see that the t-test is indeed robust, the bootstrap test tends to correct the bias with higher sample sizes, but the Wilcoxon test tends to make the bias even worse. As suggested by Urbano et al. [24], this makes sense because the bootstrap test is able to better estimate the sampling distribution of the mean if it has more data, and the Wilcoxon test may be too liberal with high sample sizes if its assumptions are not met. This observation agrees with the studies by Smucker et al. [20, 21] and Urbano et al. [25], who compared the tests with *real IR data* from the Ad hoc and Robust TREC tracks, and similarly found that the bootstrap test was the one producing smallest p-values. Regarding marginal distributions, we see that both the bootstrap and Wilcoxon tests have a consistent bias for *all* distribution families, while the t-test is robust and only displays a minor and non-significant bias for the Beta-Kernel Smoothing case. Interestingly, the effect of the margins on the Wilcoxon test is not consistent across metrics. Looking at the error rates across copula families, we see that the bootstrap test is still consistently biased. The Wilcoxon test shows mild bias in *P*@10 and *RR*, but it is robust for *AP* except for a very clear bias with the Tawn copula. The t-test is very robust to changes in the copula, again with a minor bias for the Tawn copula.

Generally speaking, Figure 3 points in the direction of the Tawn copula family, as it seems to yield higher error rates for all metrics and tests, specially the Wilcoxon test with *AP* scores. A close look at its definition [22], reveals that this copula is asymmetric, while *all* the other 10 copula families used are symmetric (an example of both can be seen in Figure 4 2.A and 2.B). In their experiment, Urbano et al. [24] fitted all 11 copulas and their rotations to every pair of systems, and simply selected the one with lowest AIC score. The Tawn copula was selected 22% of the times, followed by the Joe and Gumbel copulas 16% of the times each (see Figure 2). As Figure 4 illustrates, symmetric copulas are likely to yield symmetric distributions of per-topic score differences, while asymmetric copulas may yield distributions with different degrees of skewness. This is critical, because *those* are the distributions fed to the statistical

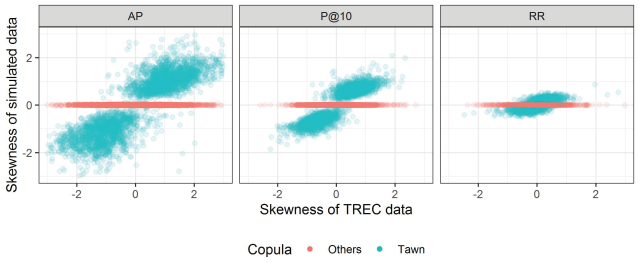

Figure 5: Skewness of the score difference distributions in TREC and simulated. Only the Tawn copula allows for skewed distributions.

tests (recall that these are *paired* tests). The side-effect of choosing the copula model based on AIC alone, is that whatever skewness there was in the TREC data is removed by the model if the selected copula is other than the Tawn copula. Conversely, a Tawn copula may also enforce the skewness observed in the TREC data even when the underlying process was not skewed (ie. it was just an artifact of the topic selection). Figure 5 shows that such skewness is indeed present in actual IR data, but to different degrees across metrics.

To further analyze the effect of skewness alone, Figure 6 shows the EMM error rates for different levels of skewness. As the plot evidences, all tests are affected to some extent, but the Wilcoxon test is clearly the least robust to skewed data. As a matter of fact, that the score distributions are symmetric (ie. zero skewness) is one of the assumptions of the test [5].[5] One may immediately wonder why the t-test is not affected that much, given that it assumes Normal distributions. We may link this to the Central Limit Theorem, stating that the sampling distribution of the mean approaches a Normal distribution as the sample size increases, regardless of the distribution of the data. This means that, regardless of how skewed

[5]Technically, it is not. The Wilcoxon test is originally for differences in the median, but it is often used as a non-parametric alternative to the t-test for differences in means, thus adding the assumption of symmetricity (ie. mean and median are the same).

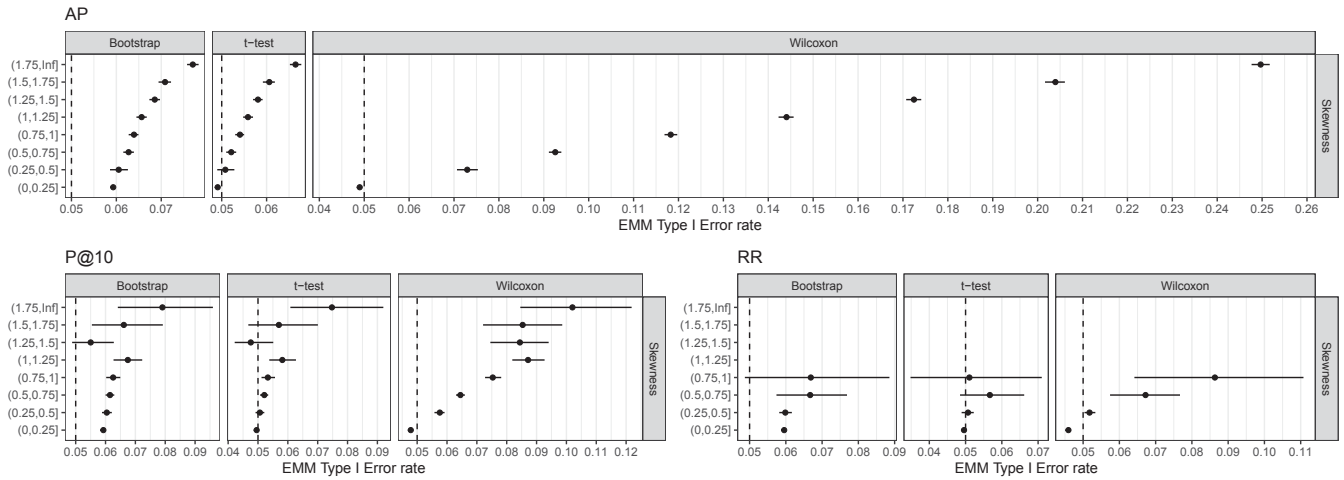

Figure 6: Estimated marginal Type I error rates for the skewness factor, along with 95% BC bootstrap confidence intervals. The vertical dashed lines mark the ideal rate $\alpha = 0.05$. Note that, for the same metric, the x-axes have the same scale; panels have different widths because they cover different ranges.

the data are, the sample mean converges to a distribution with zero skewness. Specifically, the Berry–Esseen theorem [1, 8] determines the rate of convergence via an upper bound on the Kolmogorov-Smirnov distance, which is in fact proportional to the third absolute moment $\mathrm{E}[|X|^3]$, which is itself proportional to the skewness. In short, the sampling distribution of the mean, which the t-test assumes to be Student's t (ie. symmetric), does converge to symmetric as the sample size increases, at a rate inversely proportional to the skewness of the score distribution.

4 DISCUSSION

Two parallel lines of research have recently used stochastic simulation to compute actual Type I error rates of statistical significance tests for Information Retrieval. Shockingly, they conclusively reach opposite conclusions regarding the popular Wilcoxon and t-tests. Parapar et al. [13] recently claimed that the experiments by Urbano et al. [24] favored the t-test because they used parametric marginal distributions to simulate data. In reality though, they also used non-parametric distributions based on Kernel Smoothing, which are arguably as free of assumptions as one can be, providing models that adjust to the data better than parametric models should. Following a similar logic, one could be tempted to argue that the t-test is at disadvantage with such distributions. However, our results show that it behaves equally well across all score distributions, while the Wilcoxon test showed systematically high error rates across the board. Thus, the claim by Parapar et al. seems unjustified because the t-test does *not* appear to benefit from parametric distributions.

We extended their argument to the copula families, and our analysis points in the direction of the skewness of the bivariate joint distribution (ie. the dependence between systems), and how it affects the skewness of the per-topic score differences. The Wilcoxon test assumes that score differences have zero skewness, which may explain why it performed that poorly in [24]. In contrast, the t-test is more robust to such skewness, which was similarly observed in statistical studies given a moderately large sample size [3, 4, 18]. The difference is that the Wilcoxon test has the *direct* assumption that the distribution of scores is symmetric, while the t-test assumes that the (standardized) sample mean follows a Student's t distribution, which happens to be symmetric. The first assumption is thus harder than the second one, where symmetricity is actually a *consequence* of the assumption.

Unfortunately though, the experiment by Urbano et al. [24] simulated data that, in the majority of cases, came from zero-skew models. One the one hand, it could be argued that they simply chose the models that best described the TREC data. After all, the real distribution could have zero skewness, and whatever skewness is observed in the TREC data is just an artifact of the sampling of topics. In any case, the purpose of the models is *not* to describe the retrieval systems underlying the original TREC data, but to build a realistic model of how *a* retrieval system behaves. Choosing models based on AIC offers a trade-off between simple and parsimonious models (usually chosen when optimizing the Bayesian Information Criterion), and complex and overfitted models (usually chosen optimizing Log-Likelihood [26]. On the other hand though, it could be argued that they should have fitted only asymmetric copulas which, in most cases, would have had little skewness anyway.

In any case, our analysis shows that removing skewness actually benefits the Wilcoxon test. Because most of the models fitted by Urbano et al. were expected to have zero skewness, the test that had the advantage was actually the Wilcoxon test! So all in all, while our analysis points to a potential flaw in their experiment, it also makes their case against the Wilcoxon test even stronger.

This paper is just a first attempt at clarifying the source of the discrepancy. Further work should corroborate our findings in a controlled setting, and explore all the factors at play in both simulation approaches. Such work is important not only for the ultimate goal of making sound recommendations as to what tests should be used and when, but to further our understanding of the properties of IR evaluation data and how they may affect our research conclusions.

ACKNOWLEDGMENTS

Work funded by European Union's H2020 programme (770376-2–TROMPA). Mal Ramos y mal Florentino. Peor Luis Enrique.

REFERENCES

[1] Andrew C. Berry. 1941. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Trans. Amer. Math. Soc.* 49, 1 (1941), 122–136.
[2] Ben Carterette. 2017. But Is It Statistically Significant?: Statistical Significance in IR Research, 1995-2014. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1125–1128.
[3] Wilkie W Chaffin and Steven G Rhiel. 1993. The effect of skewness and kurtosis on the one-sample T test and the impact of knowledge of the population standard deviation. *Journal of statistical computation and simulation* 46, 1-2 (1993), 79–90.
[4] G Cicchitelli. 1989. On the robusness of the one one sample t test. *Journal of Statistical Computation and Simulation* 32, 4 (1989), 249–258.
[5] William J. Conover. 1999. *Practical Nonparametric Statistics*. Wiley.
[6] Gordon V. Cormack and Thomas R. Lynam. 2007. Validity and Power of t-test for Comparing MAP and GMAP. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 753–754.
[7] Bradley Efron. 1981. Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics* 9, 2 (1981), 139–158.
[8] Carl-Gustav Esseen. 1942. On the Liapunoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik* A28 (1942), 1–19.
[9] David Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 329–338.
[10] R. Manmatha, Toni M. Rath, and Fangfang Feng. 2001. Modeling Score Distributions for Combining the Outputs of Search Engines. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–275.
[11] Douglas C. Montgomery. 2020. *Design and Analysis of Experiments* (10th ed.). Wiley.
[12] Roger B. Nelsen. 2006. *An Introduction to Copulas* (2nd ed.). Springer.
[13] Javier Parapar, David E. Losada, and Alvaro Barreiro. 2021. Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation. In *ACM Simposyum on Applied Computing*.
[14] Javier Parapar, David E. Losada, Manuel A. Presedo Quindimil, and Alvaro Barreiro. 2020. Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the American Society for Information Science and Technology* 71, 1 (2020), 98–113.
[15] Tetsuya Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 5–14.
[16] Mark Sanderson and Justin Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 162–169.
[17] Jacques Savoy. 1997. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing and Management* 33, 4 (1997), 495–512.
[18] Shlomo S Sawilowsky and R Clifford Blair. 1992. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological bulletin* 111, 2 (1992), 352.
[19] S. R. Searle, F. M. Speed, and G. A. Milliken. 1980. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician* 34, 4 (1980), 216–221.
[20] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *ACM International Conference on Information and Knowledge Management*. 623–632.
[21] Mark D. Smucker, James Allan, and Ben Carterette. 2009. Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 630–631.
[22] Jonathan A. Tawn. 1988. Bivariate Extreme Value Theory: Models and Estimation. *Biometrika* 75, 3 (1988), 397–415.
[23] Julián Urbano. 2016. Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation. *Information Retrieval Journal* 19, 3 (2016), 313–350.
[24] Julián Urbano, Harlley Lima, and Alan Hanjalic. 2019. Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 505–514.
[25] Julián Urbano, Mónica Marrero, and Diego Martín. 2013. A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 925–928.
[26] Julián Urbano and Thomas Nagler. 2018. Stochastic Simulation of Test Collections: Evaluation Scores. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 695–704.
[27] Cornelis J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths.
[28] Ellen M. Voorhees. 2009. Topic Set Size Redux. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 806–807.
[29] W. John Wilbur. 1994. Non-parametric Significance Tests of Retrieval Performance Comparisons. *Journal of Information Science* 20, 4 (1994), 270–284.
[30] Justin Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments?. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307–314.