Stop Using the Wilcoxon Test: Myth, Misconception and Misuse in IR Research

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

Abstract

In benchmarking of Information Retrieval systems, the Wilcoxon signed-rank test is often treated as a safer alternative to the t-test. This belief is fueled by textbooks and recommendations that portray Wilcoxon as the proper non-parametric alternative because metric scores are not normally distributed. We argue that this narrative is misleading and harmful. A careful review of Statistics textbooks reveals inconsistencies and omissions in how the assumptions underlying these tests are presented, fostering confusion that has propagated into IR research. As a result, Wilcoxon has been routinely misapplied for decades, creating a false sense of safety against a threat that was never there to begin with, while introducing another one so severe that it virtually guarantees the test will break down and mislead researchers. Through a combination of systematic literature review, analysis and empirical demonstrations with TREC data, we show how and why the Wilcoxon test easily loses control of its Type I error rate in IR settings. We conclude that the continued use of Wilcoxon in IR evaluation is unjustified and that abandoning it would improve the methodological soundness of our field.

CCS Concepts

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

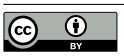
Keywords

Statistical significance, Student's t-test, Wilcoxon test

ACM Reference Format:
Julián Urbano. 2026. Stop Using the Wilcoxon Test: Myth, Misconception and Misuse in IR Research. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3805712.3808540

1 Introduction

In Information Retrieval (IR) evaluation, system effectiveness is usually measured over the fixed set of topics in a test collection or dataset. Because topic difficulty varies, observed differences between two systems, say X and Y, inevitably reflect both genuine performance gaps and random variation due to topic sampling. Statistical testing is typically used to separate signal from noise: to ask whether the apparent improvement of X over Y is more than a coincidence due to the selection of topics in the test collection.



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808540

Formally, for each topic $i = 1, \ldots, n$ we record scores X_i and Y_i according to some metric like NDCG, and the null hypothesis of interest is that the systems have equal mean effectiveness in the population of topics, $H_0 : \mu_X = \mu_Y$. Since both systems are tested on the same topics, this is called a *paired two-sample problem*, where dependence across topics (difficulty) is accounted for. This leaves us with two classical procedures that dominate the Statistics literature and IR practice, usually described as follows:

(1) Paired Student t-test: a parametric test assuming that scores follow a normal distribution.
(2) Wilcoxon signed-rank test: a non-parametric alternative when the normality assumption is violated.

It is worth noting that once we define the per-topic differences $D_i = X_i - Y_i$, this problem is equivalent to testing the one-sample hypothesis $H_0 : \mu_D = 0$. As we will show in Section 2.3, this alternative view simplifies definitions and discussion.

Over the decades, IR research has repeatedly asked which test is "best" for system benchmarking. A first wave of papers discussed mostly theoretical arguments [33, 75, 80]. A second wave took an empirical angle, using resampling and random topic set splits [9, 20, 53, 57, 58, 63, 64, 72, 76, 82]. A third wave is represented by simulation studies that recently calculated actual Type I and II errors for IR-like data,[1] but reached opposite conclusions: [70, 71, 73] recommend the t-test and discourage the Wilcoxon test, while [47, 48] actually recommend the latter. In practice, surveys of venues like SIGIR and ECIR show that about 65% of papers use the t-test and 25% Wilcoxon, with other methods being a minority [11, 54].

In this paper, we argue that the Wilcoxon test should no longer be used in IR evaluation, not because it is sub-optimal, but because it is actively *harmful*. Unlike earlier empirical or simulation-based work, we take a different route and present a systematic review of 25 Statistics textbooks. This reveals the root of a problem: the very sources we rely upon to shape our collective understanding perpetuate a confusing and misleading dichotomy between parametric and non-parametric methods. We frame the issue in terms of *myth*, *misconception*, and *misuse*. The myth is that non-parametric methods are inherently safer when parametric assumptions fail. The misconception is that the t-test requires normally distributed scores, when in fact it only requires their mean to be approximately normal. The misuse is the reliance on the Wilcoxon test as a quick remedy when normality is questioned, because its own assumption of *symmetry* is actually far stronger, yet almost always ignored. Through analysis and simulation, we show that this oversight makes the Wilcoxon

[1] As a reminder, a Type I error occurs when the null hypothesis is true but it is rejected (claiming a difference when there is none, i.e., a false positive), and a Type II error occurs when the null is not true and yet it is not rejected (missing a difference when there is one, i.e., a false negative).

test fail catastrophically for common IR data, whereas the t-test remains robust. In short, the supposed "safe" alternative actually turns out to be the riskier choice, and continuing its use only deepens the confusion and undermines the credibility of our findings.[2]

2 Systematic Review of the Statistics Literature

In order to understand the confusion around statistical testing, we review a selection of textbooks to investigate how the Statistics literature describes these methods. In particular, we selected a total of 25 textbooks, categorized as follows:

- 11 books on Statistics in general [8, 21, 34, 35, 39, 46, 52, 60, 74, 79, 84].
- 6 books specifically on Non-parametric Statistics [18, 31, 32, 45, 50, 61, 78].
- 8 books on Statistics for particular fields such as Behavioral Sciences [51], Biology [65], Business and Economics [1], Computer Science [3], Education [15], Engineering [43], Environmental Science [38], and Psychology [2].

We examined and annotated all these sources with respect to how they discuss recommendations, non-parametric methods, the t-test, and the Wilcoxon test.

2.1 Recommendations

When researchers seek guidance on which statistical test to use, many sources offer quick roadmaps with questions like *"how many groups do you have?"* or *"are your data normally distributed?"* to guide them. Of the 25 books we reviewed, 10 provided such shortcuts in the form of tables or decision trees (see Table 1).

The first noteworthy observation is the diversity of descriptions about the object of the test leading to the paired-sample Wilcoxon test, ranging from vague statements like "differences between groups," to precise ones such as "location, median," or confusing ones such as "the ordering of data in two dependent populations." As the table shows, most sources contrast parametric and non-parametric methods, and/or link test choice to the level of measurement of the data. In the latter cases, Wilcoxon is recommended for ordinal data, with the exception of [18, pp. 252], who argues that D being ordinal implies that both X and Y must be interval. On the other hand, [61, pp. 76] argues for X and Y on an "ordered metric" scale, which lies between ordinal and interval in strength [19].

The diversity and vagueness in Table 1 suggest that a proper understanding of the underlying principles is essential. In the worst case, a researcher may simply pick "whatever seems to fit my problem". In the best case, they find quick guidance but are expected to grasp the subtleties, their implications and risks. This raises the question: how does the Statistics literature actually explain concepts such as *non-parametric methods*, and does it enable sound choices?

2.2 The Myth: Non-parametric Safety

Defining *non-parametric methods* precisely proves so difficult that some authors avoid the question altogether. For example, L. Wasserman, in his popular *All of Nonparametric Statistics* [78], "did not venture to give one, no doubt I would be barraged with dissenting opinions." He simply says that they require "as few assumptions as possible" and proposes the alternative term "infinite-dimensional." Similar hesitation appears across our 25 books [8, 18, 31, 50, 79], with some of them simply choosing to not give a definition at all, however vague [46, 84].

Table 1: Textbooks with decision trees for selecting the appropriate statistical test. The main decision is made with respect to parametric vs. non-parametric methods, or level of measurement (for these, the table lists the level specified for the Wilcoxon test). The reported object of the test refers to the description used to lead into the paired-sample Wilcoxon test. Implicit mentions are indicated with square brackets.

	Object of the test	Param vs. Non-param	By level, Wilcoxon as
General			
[34]	Central tendency	•[a]	
[60]	Hypothesis about the ordering of data in two dependent populations		ordinal / rank-order
[74]	Difference between two samples or groups	•[b]	ordinal / non-param[b]
Non-parametric			
[18]	Means (medians)	[•]	interval
[32]	Location, median	[•]	
[61]	–	[•]	ordinal[c]
Field-specific			
[2]	–	•[d]	
[15]	Difference	•	ordinal
[43]	Difference in means from two normal distributions in a paired analysis	•[e]	
[51]	Differences between groups	•	ordinal

[a] The dichotomy is actually parametric vs. distribution-free.
[b] The dichotomy is actually measurement (parametric) vs. ordinal (non-parametric).
[c] D must also be ordinal, not just X and Y.
[d] The dichotomy is actually parametric vs. rank-order.
[e] Only lists non-parametric tests for the unpaired case.

Table 2 summarizes how the books describe non-parametric statistics. Diversity is again the norm, with passages like "few or no assumptions", "no assumptions about how normal" or "no assumption about distribution parameters". We can generally classify books in two groups: 18 say that these methods do not make *any* assumptions, and 7 say that they make *few or fewer* assumptions. In 4 books, these assumptions are specified in general or vague terms, sometimes not even clarifying what the object of the assumption is, while in another 7 they clearly talk about assumptions regarding the distributions. In 12 books the authors explicitly talk about *specific* distributions, their shape or functional forms, while 4 books narrow it down to the specific parameters of these distributions. This differentiation might seem nitpicky, but it is an important one. Imagine a test that only assumes that distributions are symmetric:[3] this is certainly not distribution-free, it does not deal with a specific distribution family or its parameters, but it definitely constrains its shape. It is therefore understandable how the statements in Table 2 might mislead a reader unaware of the details. In fact, because of the difficulty in properly defining the term, 5 books end up giving inconsistent statements.

Acknowledging this confusion, some authors propose alternative terms such as *exchangeability theory* methods [8] or *assumption freer* [60], while others prefer to emphasize their use of ranks [2, 31, 39]. However, as many as 9 books explicitly call them

[2] Data and code are available at https://github.com/julian-urbano/sigir2026-wilcoxon.

[3] This is in fact an assumption of the Wilcoxon signed-rank test (see Section 2.4).

Table 2: Description of *non-parametric methods* in Statistics textbooks. They may mention few(er) or no assumptions in general (G), about the distribution (D), about its functional form (F) or about its parameters (P). Some say they are also called *distribution-free*, and some give conflicting definitions. Some key excerpts are slightly rewritten due to space constraints.

	Key excerpt	Assumptions			
		None	Few(er)	Dist-free	Conflict
	General				
[8]	"replace the normal-iid assumption for the assumption of exchangeability"			•	
[21]	"does not assume a particular parametric model"	F			
[34]	"no assumption has to be made regarding the frequency distribution [...] parametric only if normal"	D			
[35]	"without assuming a given functional form for the distribution, such as the normal"	F			
[46]	"alternative when the population distribution is non-normal"				
[39]	"few or no distributional assumptions are required [...] hypotheses in terms of population distributions, not parameters "	D	D		
[52]	"do not assume that the data follow any particular distributional form"	F			
[60]	"no assumptions with regard to the population parameters that characterize the distributions [...] really not assumption free"	P			
[74]	"make no assumptions about distributions"	D			
[79]	"do not require the specification of the underlying distribution [...] no common agreement among statisticians"	F		•	
[84]	–				
	Non-parametric				
[18]	"do not assume a particular population probability distribution [...] valid for any population with any distribution"	D, F			•[a]
[31]	"based on ranks"				
[32]	"require few assumptions about the underlying populations [...] in particular, the traditional assumption that they are normal"		D		
[50]	"when important test properties hold even if only very general assumptions are made about the probability distributions"		G	•	
[61]	"no conditions about the parameters of the population [...] assumptions are fewer and much weaker"	F, P	G	•	•[b]
	Field-specific				
[1]	"without making an assumption about the specific form of the population's probability distribution"	F		•	
[2]	"no assumptions about the shape of populations [...] or about population parameters"	F, P		•	•[c]
[3]	"does not assume any particular distribution"	F			
[15]	"few or no assumptions about the distribution [...] no assumptions about how normal, even and regular the distribution"	F	D		•[d]
[38]	"often a question of whether the population has a normal distribution or not [...] parametric tests require more assumptions"		G		
[43]	"does not depend on the form of the underlying distribution of the observations"	F		•	
[45]	"neither require a specific distributional assumption nor a high level of measurement"	F			
[51]	"do not require that the data in the population be normality distributed [...] data can have any type of distribution"	D		•	•[e]
[65]	"their null hypothesis is not concerned with specific parameters but only with the distribution of the variables"	P	G	•	

[a] "No particular distribution" does not mean "*any* distribution".
[b] Scores not distributed in a certain way [...] but *still* certain assumptions.
[c] "No assumption about *parameters*" of the distribution does not mean "no assumptions about *shape*".
[d] First says "*few or no* assumptions", and later says "*no* assumptions".
[e] "Non-normal distribution" does not mean "*any* distribution".

distribution-free, reinforcing a false sense of generality that they do not actually possess because they, in fact, *do make assumptions* about distributions. As Box et al. [8] noted, practitioners should be forgiven for being misled by such terminology.

In practice, guidelines very often reduce the decision to whether normality holds or not, but rejecting normality is trivial in IR evaluation with simple arguments like "metric scores are discrete" or "metric scores are bounded in [0,1]" (see e.g. [9] for a larger discussion). Such arguments, while undeniably proving non-normality, may have catastrophic consequences as researchers would automatically lean towards non-parametric methods, not realizing that they make assumptions of their own, so much that they may actually pose more severe risks than their parametric counterparts. We illustrate this next, showing that the parametric Student t-test behaves just fine despite assuming normality, while the non-parametric Wilcoxon signed-rank test breaks dramatically when its own assumptions, often unheard of, are slightly violated.

2.3 The Misconception: Student t-test

Recall that $D = X - Y$ represents the per-topic difference between systems X and Y, and our hypothesis of interest is $H_0 : \mu_D = 0$. In other words, we test whether the expected difference between systems, in the topic population, is zero. Under the assumption that D is normally distributed with mean μ_D and variance σ_D^2, it follows from closure of the normal distribution that the sample mean $\overline{D}$ is also normally distributed with mean μ_D and variance σ_D^2/n, where n is again the number of topics. The standardized mean, also known as the z-score, follows a standard normal:

$$z = \frac{\overline{D}}{\sigma_D/\sqrt{n}} \sim \mathcal{N}(0,1) \ . \tag{1}$$

The p-value could be computed by placing this z-score in the *cdf* of the standard normal. Unfortunately, under typical experimental settings σ_D is unknown. To address this, Student [66] introduced the t-score, replacing σ_D with the *observed* standard deviation s_D:

$$t = \frac{\overline{D}}{s_D/\sqrt{n}} \sim \mathcal{T}(n-1) \ , \tag{2}$$

where $\mathcal{T}(n-1)$ is the t distribution with $n-1$ degrees of freedom. Essentially, this accounts for the random noise in estimating σ_D via s_D, making the t-distribution similar to a normal but with heavier tails. As n increases though, it converges to a standard normal.

The main objection to the t-test arises when D is non-normal—admittedly the norm in experimental research [40]—motivating the choice of the non-parametric Wilcoxon test. However, although the classical derivation of the test assumes the normality of D, we must note that this assumption has *no practical relevance*. What really matters is whether the sample mean $\overline{D}$ remains normal, because the p-value is ultimately determined only by the t-score, regardless

Table 3: Description of the *t*-test and Wilcoxon test in Statistics textbooks. The first column indicates whether they describe the equivalence between the paired two-sample and one-sample problems. For *t*-test: they may place the normality assumption on the scores (X), their difference (D), or the means (M); some of them discuss behavior for large *n*, and some give specific warnings depending on the data. For Wilcoxon: they may mention the continuity assumption and the consequences of zeros and ties, as well as the assumption of symmetry and the consequence of actually testing the median.

	paired = 1-sample	*t*-test Normality	*t*-test Large *n*	*t*-test Warnings	Wilcoxon test Continuity	Wilcoxon test Zeros	Wilcoxon test Ties	Wilcoxon test Symmetry	Wilcoxon test Median
General									
[8]	[•][a]	?, [M]							
[21]		?[b]	•	Small *n*					
[34]	[•][a]	X			•			[•][b]	
[35]	•	D	•	Heavy tails				•	
[39]	•	X, [D][c]	•			•[d]	•[d]		
[46]	•	D		Skewness		•[d]	•[d]	•	•
[52]	[•][a]	D	•	Small *n*, non-normal			•	•	
[60]		X						•	•
[74]		[?][e]							
[79]	•	X, [D]				•[d]	•[d]	•	•
[84]	•	M, [D]	•						
Non-parametric									
[18]	•	D				•[d]	•[d]	•	•
[31]		?[b]						•	•
[32]					•	•	•	•	•
[45]								•	•
[50]					•	•	•	•	•
[61]		D		Assumptions unrealistic					
Field-specific									
[1]	•	D	•			•[d]	•[d]	•	•
[2]	•	D	•	Skewness					
[3]		?[b]			•	•	•	•	•
[15]	[•]	?	•						
[38]	•	D							
[43]	•	D	•	Skewness, not unimodal	•		•	•	•
[51]	•	X	•						
[65]	[•][e]	[D][e]							

[a] By using the same equations, but without explicit mention.
[b] Only in the context of the one-sample problem.
[c] Only in the context of confidence intervals.
[d] Discusses how to handle zeros and ties, but does not mention them as consequences of continuity.
[e] Only in the context of regression or ANOVA.

of the shape of D itself. In this sense, the assumption of normality really concerns the mean, not the observed scores. Strictly speaking, Cramér's theorem tells us that $\overline{D}$ is exactly normal only if D is normal as well, but exact normality is not the relevant criterion in practice. What matters is whether $\overline{D}$ is normal *enough*. Indeed, by the Central Limit Theorem (CLT), $\overline{D}$ converges to a normal distribution as the sample size n increases, provided that D has finite non-zero variance. This means that the t-test is asymptotically correct under very general conditions, regardless of what distribution D actually has [35, §13.2.1]. Therefore, the exact normality of D is just a derivation condition for small sample problems; practical validity depends on the approximate normality of $\overline{D}$.

The left-hand side of Table 3 shows how the 25 Statistics textbooks describe the paired t-test. Only 11–16 books show the equivalence between the paired two-sample test of $\mu_X = \mu_Y$ and the one-sample test of $\mu_D = 0$. This is important because, when the equivalence is omitted, readers may misinterpret the normality assumption as concerning X and Y when in reality it concerns D. This happens in 5 books, but it is perfectly possible for X and Y to diverge significantly from normality and yet result in a D that is very close to normal. Another 9–13 books correctly placed it on D, and 6 books did not make it clear. Somewhat surprisingly, only 2 books identified the key assumption on the mean: [8] did this as a consequence of score normality, while only [84] explicitly described $\overline{D} \sim \mathcal{N}$ as *the* assumption. Another assumption easily misinterpreted is homoskedasticity (i.e., $\sigma_X^2 = \sigma_Y^2$), which does not apply to our paired samples case.[4] This inconsistency in communication likely stems from the primarily didactic focus of textbooks: they usually explain the t-test through its formal derivation from normal variables, just as we did through Eqs. (1) and (2), while far less attention is given to the practical conditions under which the test remains valid when applied to real data.

As noted earlier, the literature generally recommends a non-parametric alternative when normality is not present. As Table 3 shows, some books explicitly warn the reader about small n, skewness and heavy-tails.[5] To illustrate the effect of departures from normality directly in D, let us consider the four most common ways to break the assumption: a skewed distribution that introduces asymmetry, a high-kurtosis distribution that introduces heavy tails, a discrete and bounded distribution that modifies the support, even in

[4] Homoskedasticity would apply to the *unpaired* two-sample case, for example when evaluating with two different collections [55].
[5] We follow standard definitions, so skewness γ refers to the third standardized moment and kurtosis κ refers to the fourth. We always compute excess kurtosis.

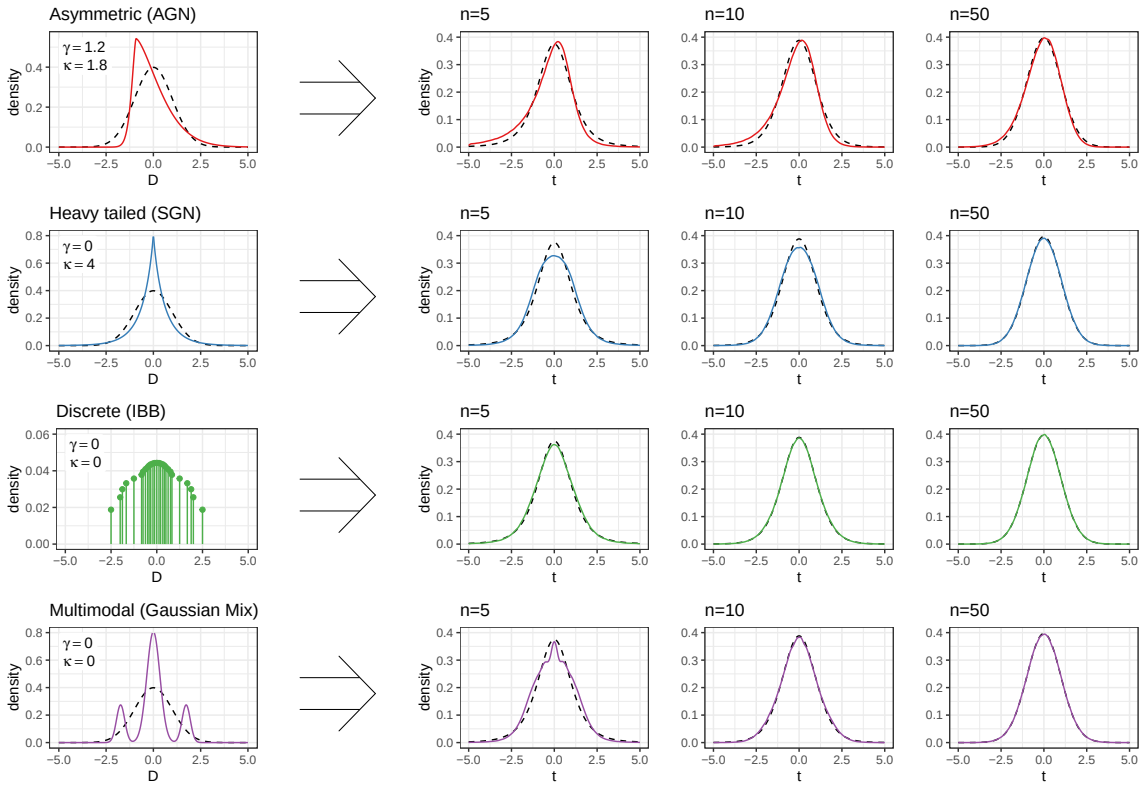


Figure 1: Effect of asymmetry, tail heaviness, discrete support and multimodality of D (left), on the distribution of the t statistic for sample sizes $n = 5, 10, 50$ (right), illustrating the effect of the Central Limit Theorem. All distributions (see Appendix A for details) are standardized to zero mean and unit variance; γ and κ denote skewness and kurtosis. Sampling distributions are obtained empirically (1 million replicas). Dashed lines represent the distributions when D is normally distributed.

irregular ways, and a multimodal distribution (see Figure 1). These distributions are *far* from normal, yet for moderate n the empirical distributions of the resulting t-scores are nearly indistinguishable from the $\mathcal{T}(n-1)$ expected under normality. Therefore, in practice the t-test may behave well provided n is not trivially small, even under gross violations of the normality assumption in D. Unfortunately, a similar discussion about normality and large samples is present only in 10 of the 25 books, needlessly pushing the reader towards the Wilcoxon test under the pretext of non-parametric safety.

2.4 The Misuse: Wilcoxon Signed Rank Test

Wilcoxon's goal was not to create a test robust to non-normality, but to simplify calculations in the pre-computer era [81]. His solution was rank-order statistics: instead of the raw differences D_i, the test works with their ranks R_i after sorting by absolute value, thus discarding magnitudes. The test assumes *exchangeability*, meaning that X and Y have the same distribution and the X/Y labels are arbitrary. Under H_0, a difference is equally likely to be positive or negative, so the statistic simply sums the ranks of positive differences:[6]

$$W^+ = \sum_{i:\, D_i>0} R_i \,, \tag{3}$$

[6]For example, $D = \{-0.4, -0.1, 0.4, 0.8\}$ results in $R = \{2.5, 1, 2.5, 4\}$ and $W^+ = 6.5$.

yielding a p-value via the Wilcoxon signed-rank distribution with sample size n (equivalently, one may sum the negative ranks W^-).

In his original formulation, Wilcoxon assumed D to be continuous, as otherwise it would be impossible to tabulate the distribution of W^+ because it depends on the number of zeros and ties. Dropping zeros, as he proposed, remains the default in software packages, though alternatives exist [16, 49]. When ties or zeros occur, a normal approximation is used for the distribution of W^+. Conover [17] later showed that strict continuity is unnecessary; it suffices that $P(D = d) < 1$ for all d, making the method applicable to ordinal data. Still, as Table 3 shows, 5 books continue to mention continuity as a requirement, while most of them simply focus on the practicalities of handling ties and zeros.

But there is an additional subtlety with the Wilcoxon test: because it relies on rank-order statistics, it is formally a test for the *median* of D, not the mean. This is explicitly mentioned in 11 books, mostly the ones on non-parametric methods. More importantly, a direct consequence of the exchangeability assumption is that the distribution of D is *symmetric* around 0, which makes the median equal to the mean. For this reason, the Wilcoxon test is legitimately used for the null hypothesis $H_0 : \mu_D = 0$, but at the price of the *extra* assumption of symmetry. Unlike the t-test, whose validity improves with larger samples via the CLT, violations of symmetry are not

diminished as n increases. Quite the opposite! In the presence of asymmetry, a larger sample size only makes the systematic drift of W^+ even more pronounced. For this reason, the null hypothesis underlying the Wilcoxon test is effectively *twofold*: it asserts both $\mu_D = 0$ and symmetry of the distribution of D. If there is asymmetry, the null distribution of W^+ is no longer appropriate, and rejection may easily occur even when $\mu_D = 0$. In such cases the test is no longer testing for a difference between groups, but for asymmetry in the distribution of differences.[7]

The contrast can be understood in terms of pivotality.[8] The t-statistic is an exact pivot under normality and remains asymptotically pivotal under very mild conditions, as its limiting distribution does not depend on the shape of D. In contrast, the Wilcoxon W^+ statistic is distribution-free only under symmetry. When symmetry fails, its null distribution depends on the underlying distribution of D, and this dependence does not disappear as n increases.

2.5 Empirical Demonstration

To demonstrate the robustness of the t-test and Wilcoxon test to violations of their assumptions, as well as the effect of sample size, we carried out a small Monte Carlo simulation study of Type I error rates under the null hypothesis $H_0 : \mu_D = 0$ and for each of the four distributions in Figure 1. In particular, for each distribution we simulated 100K samples of sizes $n = 5, 50, 500, 5000$, and recorded the observed error rates at $\alpha = .05$. Recall that, under these conditions, the tests are expected to have a Type I error rate of 5%. Table 4 shows that both tests are quite robust under the multimodal, heavy-tailed and discrete cases, although for Wilcoxon this holds only when the sample size is not extremely small. However, while the t-test is able to maintain the nominal 5% error rate under asymmetry, the Wilcoxon test fails catastrophically, and increasingly so as the sample size increases.

This simple demonstration shows that choosing the Wilcoxon test simply because of observing non-normal data can actually be a terrible idea: the symmetry assumption poses a much higher risk than non-normality does. This is ironic, to say the least, because asymmetry is probably the easiest feature to identify non-normality and turn to the Wilcoxon test, rendering it unreliable in the very situation for which it is most often recommended!

3 Impact on IR Research

The previous section demonstrated that asymmetry *can* severely distort the Wilcoxon test, but whether such distortions are likely in IR experimentation depends on how far IR data departs from normality. In this section, we therefore present simulations that progressively diverge from normality in controlled and interpretable ways, and evaluate both the t-test and Wilcoxon test with respect to their Type I error rates. Our primary interest lies in the *trends* as sample size increases and departures from normality become more severe, allowing us to draw general conclusions about robustness without tying them to specific configurations.

[7] Ironically, this is one of the uses of the Wilcoxon test statistic [36].

[8] A statistic is called *pivotal* if its sampling distribution does not depend on unknown parameters. A classical example is the z-statistic computed from a normal $\mathcal{N}(\mu, \sigma^2)$, as it follows a standard normal regardless of μ and σ^2.

Table 4: Type I error rates at $\alpha = .05$ with the distributions from Figure 1 and various sample sizes. Green for good error rates (within .001 of α), yellow for reasonable rates (within .005 of α), and red for poor rates.

	Asymmetric		Heavy tailed		Discrete		Multimodal	
n	*t*-test	Wlcxn	*t*-test	Wlcxn	*t*-test	Wlcxn	*t*-test	Wlcxn
5	.084	0	.032	0	.040	0	.024	0
50	.056	.117	.048	.048	.050	.049	.051	.050
500	.052	.629	.051	.051	.051	.051	.050	.050
5000	.050	1	.050	.050	.050	.050	.050	.050

3.1 Departures from Normality

Let us recall the four most common ways in which a distribution may depart from normality: asymmetry, tail heaviness, discrete support and multimodality. We must note that our goal is not to identify *the* true distributional family underlying IR data,[9] but rather to assess how progressively increasing departures along each of these dimensions affects test validity. We thus define 6 graded levels of departure from normality, labeled as low, medium, high, very high, extremely high, and pathologically high.

First, we deliberately do not assess the effect of multimodality because, unlike the other dimensions, even quantifying the amount of multimodality is itself a hard and ill-posed problem [29, 62]: it does not naturally admit a scalar measure to be progressively increased, and both the number and prominence of modes depend on smoothing parameters, model assumptions and diagnostics. We therefore restrict attention to the more interpretable dimensions of asymmetry, tail heaviness and discreteness.

Regarding asymmetry and tail heaviness, we anchor the levels of departure in realistic regimes by first examining real IR metric scores from TREC. We collected data from somewhat recent and continued tracks with a large number of runs: Web (ad hoc) 2011–13, Microblog (ad hoc) 2012–14, Deep Learning (passage) 2019–21, Clinical Decision Support 2014–16 and Precision Medicine (abstracts) 2017–19. This resulted in 1,209 runs, and pairing all systems within the same track edition yielded 53,703 samples of paired differences D. For each pair, we computed AP, NDCG[10], P@10 and RR. Next, we computed skewness and excess kurtosis as measures of asymmetry and tail heaviness at $n = 50$ topics (over/under-sampling where necessary). Figure 2 shows the distributions observed in TREC data and, for reference, the expected distributions under normality (dashed lines). It is clear that IR data depart substantially from the symmetry and tail heaviness expected from normally distributed scores. We note though that high sample skewness may still be observed with symmetric distributions if they have heavy tails, as we appear to have. Therefore, we also compare with the expected skewness under symmetric distributions with tails as heavy as observed in the data (dotted lines); it is clear that the observed skewness is still much higher than expected under symmetric populations. To set departure levels for asymmetry, we use the values of (absolute) skewness that roughly correspond to the empirical percentiles .25, .5, .75, .9, .99 and .999: $\gamma = 0.25, 0.5, 1, 1.5, 3, 5$, respectively (no departure would be $\gamma = 0$). For tail heaviness we use excess kurtosis,

[9] One should question whether such an enterprise is even attainable in practice.

[10] For the Clinical Decision Support and Precision Medicine runs we followed the track guidelines and actually used infAP and infNDCG.

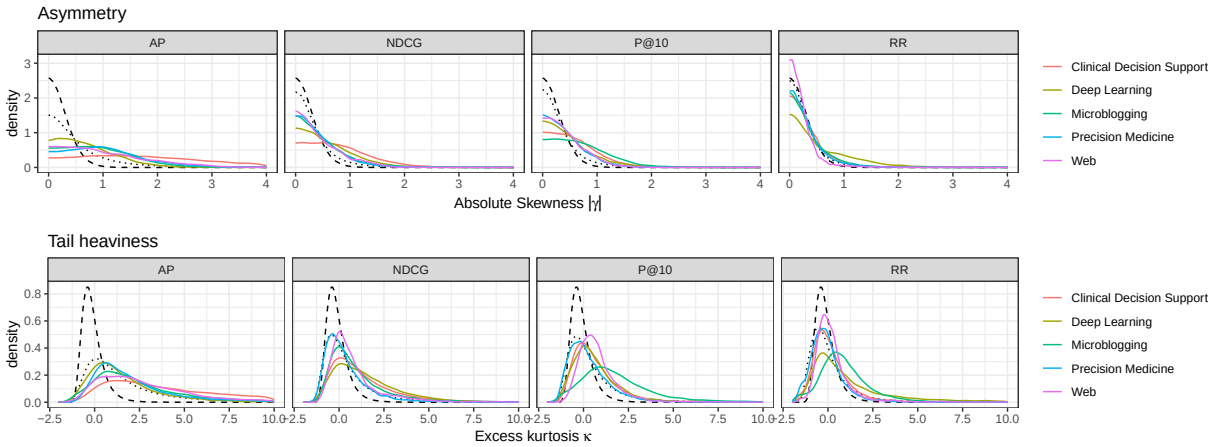


Figure 2: Symmetry and tail heaviness observed in D distributions from TREC data at $n = 50$. For reference, dashed lines represent the sampling distributions expected if D were normally distributed, and likewise dotted lines if they had tails as heavy as observed but still symmetric. The largest observed values are $|\gamma| \approx 7$ and $\kappa \approx 45$, but the axes are trimmed for clarity.

treating heavier and lighter tails separately because their interpretation is not symmetric as in skewness (i.e. an excess kurtosis of $+0.5$ is not conceptually comparable to -0.5). For heavy tails, we again use values roughly corresponding to the same percentiles: $\kappa = 0.5, 1.5, 3, 5, 15, 30$ (no departure would be $\kappa = 0$). For light tails we cap calibration from below at -1.2, as more negative values correspond to U-shaped distributions rather than merely lighter-tailed shapes; modeling them would require explicit multimodal mechanisms, which we excluded from this study. The departure levels for light tails are $\kappa = -0.2, -0.4, -0.7, -0.9, -1.1, -1.2$. Together, these values define progressively increasing deviations from normality along each dimension.

Regarding discreteness, departures from normality arise naturally in IR metrics because their supports are finite and often irregular. This is particularly evident in metrics such as P@k and RR@k (e.g., a RR@k score can only be one of $0, 1/k, 1/(k-1), ..., 1$). We model discreteness via the support structure. For a metric M with cut-off k, let Ω denote the set of all possible values when computing the difference between two M@k scores (rounded to 3 decimal digits for simplicity). The support of D is Ω, and the level of discreteness is governed by k (e.g., $|\Omega| = 21$ for P@10, whereas $|\Omega| = 95$ for RR@10). To set departure levels for discreteness, we therefore use cut-off values $k = 1000, 500, 100, 50, 10, 5$. Larger k produce fine-grained supports that approximate continuity, while smaller k induce coarser supports. In the limit, $k \to \infty$ roughly corresponds to a continuous setting with no departure due to discreteness.

3.2 Non-Normal Distributional Mechanisms

To generate distributions exhibiting the desired shapes, we employ the families described in Appendix A:

- Asymmetry. We use the Asymmetric Generalized Normal (AGN) and Tukey G-and-H (TGH, with $h=0$) distributions, varying their asymmetry parameters ξ and g to match the target skewness while keeping tail heaviness controlled.
- Tail heaviness. We use the Symmetric Generalized Normal (SGN) and TGH (with $g=0$), varying their shape parameters β and h to match the target kurtosis while ensuring symmetry.
- Discreteness. We use the Irregular Beta-Binomial (IBB), varying the support parameter Ω to match the target level of discreteness. Specifically, we simulate separately under P@k and RR@k regimes. Symmetry is guaranteed by construction, and tail heaviness is partially determined by the parameter p to the underlying Beta-Binomial.

These distributions are chosen for flexibility, interpretability and control. We do not assess goodness-of-fit, and we of course do not claim that they represent the "true" data-generating mechanism underlying IR evaluation. Instead, we manipulate general distributional *properties* —symmetry, tail heaviness and support structure— in a controlled manner. These families serve as useful parametric mechanisms to achieve target levels of these properties while holding others fixed, so we can focus on the impact of their related assumptions on Type I error rates, not the adequacy of specific distributional models for IR data. Finally, we hold σ_D constant in all distributions, because Type I error rates depend on distributional shape and sample size, not on scale. Rescaling does not affect skewness or kurtosis, cancels out in the t-statistic, and leaves Wilcoxon ranks unchanged. Since we already vary sample size n, allowing σ_D to vary too would not change the trends of interest. Therefore, we fix $\sigma_D = 0.22$, corresponding to the median observed in TREC data.

3.3 Type I Error Rates under Non-Normality

For each distributional mechanism, departure level and sample size $n = 5, 50, 500, 5000$, we simulate 100K samples under $H_0 : \mu_D = 0$. We apply both the t-test and Wilcoxon test to every sample and record whether the null is rejected at $\alpha = .05$. These empirical rejection proportions provide estimates of the Type I error rate in each condition. As emphasized earlier, our interest lies in the *trend* of these rates as departures become more severe, rather than in the value observed at any single configuration.

Table 5 reports the Type I error rates. Starting at the bottom, for both discreteness and tail heaviness the t-test and the Wilcoxon test maintain error rates at the nominal $\alpha = .05$ across all practically relevant sample sizes. The only notable deviations occur at $n = 5$. In that setting, the Wilcoxon test produces virtually no rejections across all configurations due to the limited number of attainable signed-rank scores, resulting in a conservative test. The t-test exhibits mild to large deviations under extreme tail behavior, yielding inflated error rates for light tails and deflated rates for heavy tails. However, these effects vanish rapidly as n increases, and for $n \geq 50$ both tests remain well calibrated even under pathological tail conditions. Overall, departures from normality in the form of discreteness or tail heaviness do not meaningfully compromise Type I error control in sample sizes typical of IR experimentation.

As expected, the situation changes drastically under asymmetry. For low and medium departures, the t-test maintains error rates at the nominal level or close. However, from very high departures onward, noticeable distortions emerge at smaller sample sizes. In particular, for the typical $n = 50$ the t-test behaves just fine in the vast majority of cases, but it may inflate error rates up to approximately 10% under pathological asymmetry. As sample size increases, the distortion diminishes due to the CLT, and for $n \geq 500$ the error rate returns to nominal levels even under extreme asymmetry.

In contrast, the Wilcoxon test fails *systematically* under asymmetric distributions. At $n = 5$ it again produces zero rejections due to discreteness constraints. For moderate sample sizes ($n = 50$), it is close to nominal error rates only under low departures from symmetry. As asymmetry increases, the rejection rate rises very sharply. At $n = 500$, even moderate asymmetry leads to substantial inflation, and under sufficiently strong asymmetry the test rejects the null hypothesis nearly always. This trend becomes more pronounced as sample size increases: rather than converging to nominal behavior like the t-test, the Wilcoxon test increasingly rejects under the null hypothesis of $\mu_D = 0$ in the presence of asymmetry. As discussed earlier, this occurs because it eventually becomes a test of symmetry rather than a test of mean differences.

4 Perspectives

4.1 On Test Optimality

The two most recent lines of work on statistical testing in IR studied test optimality via simulation, but framed the question in different ways. On the one hand, Urbano et al. [70, 71, 73] follow the orthodox perspective: one should choose the test that maximizes power *subject to maintaining the nominal Type I error rate.* A test unable to control its false positive rate is, under this view, simply invalid because it is testing a different hypothesis. On the other hand, Parapar et al. [47, 48] follow a trade-off perspective that emphasizes the compromise between Type I errors and power: the most powerful test should be used to accelerate scientific innovation, *even if it deviates mildly from the nominal Type I error rate.*

Regardless of which of these perspectives one adopts, our results indicate that the Wilcoxon signed-rank test should not even be part of the discussion for IR experimentation. The reason is structural: Wilcoxon assumes symmetry of the distribution of paired differences. As Figure 2 shows, asymmetry is not an exotic edge case in IR data but rather the norm, and under such asymmetry it rejects the null hypothesis almost surely as sample size increases. In effect, it becomes a test of symmetry rather than a test of mean differences. Under the orthodox view, this alone is disqualifying. Under the trade-off view, the argument is even stronger: the Wilcoxon test cannot meaningfully trade false positives for power in the name of optimality, because its error rate is not even remotely close to the nominal level.[11]

We emphasize that we do *not* advocate the t-test as a universal solution. It appears here primarily as the natural counterpart to Wilcoxon, given that the latter is routinely proposed as its non-parametric alternative. Our central object of study is the Wilcoxon test and its behavior under empirically realistic departures from normality. Resampling-based procedures, such as permutation and bootstrap tests, have also been examined in the IR literature using both limited but real TREC data [63, 64, 72], as well as stochastic simulation of TREC-like data [47, 48, 70, 71, 73]. These studies generally suggest that resampling tests maintain the nominal Type I error rates in IR-like scenarios. However, they have not been tested as we did here, by analyzing how they are affected by departures from their key assumptions as seen in IR data. Extending the present analysis to these procedures is therefore a natural next step.

This connects to a broader methodological point. Attempting to determine *the* "true" distribution of IR metric scores leads to an underdetermined problem: parametric, non-parametric, or relevance-based generative models all embed structural assumptions that cannot be validated beyond the observed samples. Our aim was precisely to avoid that goodness-of-fit rabbit hole by focusing instead on general distributional properties and their departure from test assumptions. The same principle applies when comparing inferential procedures: the question should not be which method aligns best with a specific configuration or generative model, but which procedures remain valid across realistic distributional regimes.

For similar reasons, we deliberately excluded multiple comparison procedures, because these corrections operate downstream of elementary tests and presuppose well-calibrated p-values. If a base test fails to control its Type I error rate under realistic asymmetry, no multiplicity adjustment can repair that defect. What is clear is that the structural role of asymmetry should inform the assessment of any inferential framework, also under a Bayesian approach [10].

4.2 On Importing Results from Other Fields

A broader perspective concerns the way methodological results travel across disciplines. The review exercise in Section 2 illustrates a twofold risk. First, isolated results from another field may not reflect the full scope of debate within that field. In our case, the Statistics literature contains extensive discussions on robustness, asymptotics, and assumption violations (e.g., [4, 6, 7, 13, 14, 37, 59]), which are rarely captured by general summaries. Second, even canonical sources like textbooks may present views that are historically contingent or internally contested. Therefore, we should avoid methodological imports from other fields based on single sources, as it risks the authoritative introduction of knowledge that may

[11] Works such as [63, 71] also concluded that the Wilcoxon test fails to maintain the nominal Type I error rate, even under simulation regimes with little asymmetry [70]. In contrast, [47, 48] reported that it does maintain the nominal level, which can perhaps be explained by how they simulate data under H_0: runs are independently generated from the same model and labeled X and Y, perfectly satisfying exchangeability.

Table 5: Type I error rates at $\alpha = .05$ under various levels of departure from normality with respect to asymmetry (top), tail heaviness (middle) and discreteness (bottom), as exhibited in IR-like data, and for various sample sizes. Color codes as in Table 4.

	t-test						Wilcoxon					
n	Low	Med.	High	Very H.	Extr. H.	Patho. H.	Low	Med.	High	Very H.	Extr. H.	Patho. H.
	Asymmetric											
	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$
5	.052	.056	.070	.087	.136	.189	0	0	0	0	0	0
50	.051	.052	.054	.058	.075	.098	.052	.061	.092	.141	.316	.493
500	.051	.051	.051	.051	.054	.058	.077	.156	.452	.757	.992	1
5000	.049	.050	.050	.050	.051	.051	.321	.841	1	1	1	1
	Heavy tailed											
	$\kappa = .5$	$\kappa = 1.5$	$\kappa = 3$	$\kappa = 5$	$\kappa = 15$	$\kappa = 30$	$\kappa = .5$	$\kappa = 1.5$	$\kappa = 3$	$\kappa = 5$	$\kappa = 15$	$\kappa = 30$
5	.047	.043	.039	.036	.030	.028	0	0	0	0	0	0
50	.050	.050	.049	.049	.046	.045	.049	.049	.049	.049	.049	.049
500	.051	.051	.051	.050	.050	.049	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
	Light tailed											
	$\kappa = -.2$	$\kappa = -.4$	$\kappa = -.7$	$\kappa = -.9$	$\kappa = -1.1$	$\kappa = -1.2$	$\kappa = -.2$	$\kappa = -.4$	$\kappa = -.7$	$\kappa = -.9$	$\kappa = -1.1$	$\kappa = -1.2$
5	.052	.054	.057	.060	.064	.066	0	0	0	0	0	0
50	.051	.051	.051	.051	.051	.051	.049	.049	.049	.049	.049	.049
500	.051	.051	.050	.051	.050	.050	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
	Discrete											
	$k = 1000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$	$k = 5$	$k = 1000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$	$k = 5$
5	.043	.043	.042	.039	.037	.044	0	0	0	0	.001	.002
50	.050	.050	.050	.050	.049	.049	.049	.049	.049	.049	.048	.047
500	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.051	.050	.050	.050	.050	.050

itself be contested or even outdated. We need broader scrutiny, as otherwise we may inadvertently hard-code contested views into IR practice and even into calls for papers or reviewing guidelines [56].

In our case, and setting aside the larger discussion around definitions, the Wilcoxon test is frequently introduced in textbooks simply as the non-parametric alternative to the *t*-test, sometimes described as appropriate only for ordinal data. Such statements, when taken in isolation, obscure the nuance required to interpret Wilcoxon as a test of central tendency. For instance, if one relied on only one of the books in Table 3, there would be roughly a 50% chance of missing the symmetry assumption; indeed, only about half of the IR papers cited in Section 1 mention it. The issue, however, is not merely that the assumption exists, but that its origin, implications and practical consequences for IR evaluation have remained insufficiently clarified. We fill that gap in this paper.

A parallel example in IR is the recent debate on levels of measurement: arguments based on Stevens' taxonomy have been used to question averaging certain metrics [24, 27, 42, 56], in spite of the broad discussion and criticism around this taxonomy, and the existence of others [41]. As pointed out in Section 2.1, the Wilcoxon test has itself been subject to similar debates over the years.

4.3 On Moving from Testing to Modeling

A broader methodological issue underlies the discussion around statistical tests: the reliance on contrast-based analyses that isolate a single experimental factor and treat all others as fixed or nonexistent. This paradigm appears in classical pairwise comparisons between systems and, increasingly, in ablation studies that toggle individual system components and compare specific variations. Such approaches allow us to assess whether a change produces a statistically detectable difference, but they account for only one source of variability (i.e., topics). They do not allow us to quantify how much variability is attributable to other factors, control for them when comparing systems, or reveal how components interact across datasets or conditions.

Several lines of work already point toward a more explicit modeling approach. For example, Bootstrap ANOVA and related techniques have been proposed to account for document collection variation and system–collection interactions, allowing uncertainty to reflect not only topic sampling but also corpus variability [25, 77, 83]. Earlier, Generalizability Theory was applied to IR test collections to decompose variance across facets such as topics and assessors [5]. While limited in flexibility, this line of work made it clear that topic sampling is only one of many components of experiment uncertainty [12, 68]. More recently, linear and generalized linear modeling approaches have been explored to test system differences and to study the contribution of system components within unified regression frameworks [9, 22, 26]. These approaches allow us to control for multiple factors other than topics, handle interaction effects, increase statistical power and use structured error components, thereby moving beyond repeated pairwise contrasts [28].

The perspective we advocate for is therefore not merely to replace one test with another, but to move from isolated hypothesis tests toward explicit analysis models that make assumptions transparent, allow diagnostics, and properly handle multiple sources of variance in a single framework. In that setting, significance tests become downstream summaries of a structured model rather than standalone decisions applied to aggregated topic differences.

5 Conclusions

We have revisited the role of statistical testing in IR evaluation, not by adding yet another empirical study but by systematically examining the foundations that shape our practice: Statistics textbooks.

Our review shows that they routinely present a simplistic parametric vs. non-parametric dichotomy, reinforcing the idea that the Wilcoxon test is the "safe" alternative to the t-test whenever normality is in doubt. We find this narrative misleading and dangerous.

We identified three core problems. The *myth* is that non-parametric methods are assumption-free, when in fact they have assumptions of their own; Wilcoxon silently demands symmetry. The *misconception* is that the t-test requires normal data, when it really relies on the approximate normality of the sample mean, a condition satisfied in practice through the Central Limit Theorem. The *misuse* is the routine application of Wilcoxon instead of the t-test while ignoring symmetry, which leads to distorted null distributions and inflated error rates even under the null hypothesis. As our test collections continue to grow, these issues do not fade away but are amplified: Wilcoxon becomes even more sensitive to slight departures from symmetry, virtually guaranteeing a rejection of the null, and probably for the wrong reason.

The implications for IR are severe. Simulations confirm that Wilcoxon fails precisely under the conditions that dominate IR data, while the t-test remains robust. What is widely seen as the conservative and safe choice is, in fact, the one carrying higher risk. The Wilcoxon test should be retired from IR evaluation.

A Non-Normal Distributions

This appendix describes the distribution families we used to generate non-normal data; some were illustrated already in Figure 1:

- Symmetric Generalized Normal (SGN) [44, 67]. This is a generalization of the normal distribution where parameter β controls tail heaviness. It is symmetric by design.
- Asymmetric Generalized Normal (AGN) [23]. This is another generalization of the normal distribution where parameter ξ controls asymmetry and parameter ν controls tail heaviness.
- Tukey G-and-H (TGH) [30, 69]. This family is defined via transformations of a normal variable, where parameter g controls asymmetry and parameter h controls tail heaviness.

These distribution families allow us to vary skewness and kurtosis while offering different tail decay behaviors.

For the generation of discrete data as if produced by an IR metric M, we should consider the set of all possible values when calculating the difference between two scores; let us refer to this set as Ω. Our goal is therefore to simulate discrete data with an arbitrary support Ω. To do so, we use a Beta-Binomial distribution with parameters $n = |\Omega| - 1$ and $\alpha = \beta = p$ to generate an integer random variable J, which can then be used to index Ω. We call this the Irregular Beta-Binomial (IBB). Formally, $D \sim IBB(\Omega, p)$ is calculated as $D = \Omega_{J-1}$, where $J \sim BetaBin(|\Omega| - 1, p, p)$. By construction, IBB is therefore symmetric and tail heaviness depends on Ω and p. Note that the amount of "discreteness" (i.e., the size of Ω and the spacing between its values) is entirely determined by M.

Acknowledgments

This is dedicated to those who, somehow, always thought there was something off with FC Barcelona; to those who finally think so now that the Negreira case is public; to those who think, deep down, that nothing will happen or change; and to Pepe Kollins for ensuring we all keep thinking about it. Keep thinking.

References

[1] David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Michael J. Fry, and Jeffrey W. Ohlmann. 2019. *Statistics for Business and Economics* (14 ed.). Cengage Learning.
[2] Arthur Aron, Elliot J. Coups, and Elaine N. Aron. 2013. *Statistics for Psychology* (6 ed.). Pearson. 744 pages.
[3] Michael Baron. 2014. *Probability and Statistics for Computer Scientists* (2 ed.). CRC Press. 473 pages.
[4] R. Clifford Blair and James J. Higgins. 1985. Comparison of the Power of the Paired Samples t test to that of Wilcoxon's Signed-ranks Test Under Various Population Ahapes. *Psychological Bulletin* 97, 1 (1985), 119–128. doi:10.1037/0033-2909.97.1.119
[5] David Bodoff. 2008. Test Theory for Evaluating Reliability of IR Test Collections. *Information Processing and Management* 44, 3 (2008), 1117–1145. doi:10.1016/j.ipm.2007.11.006
[6] C. Alan Boneau. 1960. The Effects of Violations of Assumptions Underlying the t test. *Psychological Bulletin* 57, 1 (1960), 49–64. doi:10.1037/h0041412
[7] George E. P. Box. 1953. Non-Normality and Tests on Variances. *Biometrika* 40, 3/4 (1953), 318. doi:10.2307/2333350
[8] George E. P. Box, J. Stuart Hunter, and William G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation and Discovery* (2 ed.). Wiley.
[9] Ben Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems* 30, 1 (2012). doi:10.1145/2094072.2094076
[10] Ben Carterette. 2015. Bayesian Inference for Information Retrieval Evaluation. In *International Conference on the Theory of Information Retrieval*. 31–40. doi:10.1145/2808194.2809469
[11] Ben Carterette. 2017. But Is It Statistically Significant?. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1125–1128. doi:10.1145/3077136.3080738
[12] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline Evaluation Options for Recommender Systems. *Information Retrieval Journal* 23, 4 (2020), 387–410. doi:10.1007/s10791-020-09371-3
[13] 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. doi:10.1080/00949659308811494
[14] G. Cicchitelli. 1989. On the Robustness of the One-Sample t Test. *Journal of Statistical Computation and Simulation* 32, 4 (1989), 249–258. doi:10.1080/00949658908811181
[15] Louis Cohen, Lawrence Manion, and Keith Morrison. 2018. *Research Methods in Education* (8 ed.). Routledge.
[16] William Jay Conover. 1973. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Amer. Statist. Assoc.* 68, 344 (1973), 985–988. doi:10.1080/01621459.1973.10481460
[17] William Jay Conover. 1973. Rank Tests for One Sample, Two Samples, and k Samples Without the Assumption of a Continuous Distribution Function. *The Annals of Statistics* 1, 6 (1973), 1105–1125.
[18] W. J. Conover. 1999. *Practical Nonparametric Statistics* (3 ed.). Wiley. doi:10.2307/1271101
[19] Clyde H. Coombs. 1950. Psychological Scaling Without a Unit of Measurement. *Psychological Review* 57, 3 (1950), 145–158. doi:10.1037/h0060984
[20] 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. doi:10.1145/1277741.1277892
[21] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester. 2005. *A Modern Introduction to Probability and Statistics: Understanding Why and How* (1 ed.). Springer.
[22] Guglielmo Faggioli, Nicola Ferro, and Norbert Fuhr. 2022. Detecting Significant Differences Between Information Retrieval Systems via Generalized Linear Models. In *ACM International Conference on Information and Knowledge Management*. 446–456. doi:10.1145/3511808.3557286
[23] Carmen Fernández and Mark F. J. Steel. 1998. On Bayesian Modeling of Fat Tails and Skewness. *J. Amer. Statist. Assoc.* 93, 441 (1998), 359–371. doi:10.2307/2669632
[24] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales. *IEEE Access* 9 (2021), 136182–136216. doi:10.1109/access.2021.3116857
[25] Nicola Ferro, Yubin Kim, and Mark Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM Transactions on Information Systems* 37, 3 (2019). doi:10.1145/3310364
[26] Nicola Ferro and Gianmaria Silvello. 2016. A General Linear Mixed Models Approach to Study System Component Effects. In *International ACM SIGIR conference on Research and Development in Information Retrieval*. 25–34. doi:10.1145/2911451.2911530
[27] Norbert Fuhr. 2018. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *ACM SIGIR Forum* 51, 3 (2018), 32–41. doi:10.1145/3190580.3190586

[28] Frank E. Harrell. 2015. *Regression Modeling Strategies, with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2 ed.). Springer.

[29] J. A. Hartigan and P. M. Hartigan. 1985. The Dip Test of Unimodality. *The Annals of Statistics* 13, 1 (1985), 70–84. doi:10.1214/aos/1176346577

[30] Todd C. Headrick, Rhonda K. Kowalchuk, and Yanyan Sheng. 2008. Parametric Probability Densities and Distribution Functions for Tukey g-and-h Transformations and Their Use for Fitting Data. *Applied Mathematical Sciences* 2, 9 (2008), 449–462.

[31] Thomas P. Hettmansperger. 1984. *Statistical Inference Based on Ranks.* Wiley.

[32] Myles Hollander and Douglas A. Wolfe. 1973. *Nonparametric Statistical Methods* (1 ed.). Wiley.

[33] 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. doi:10.1145/160688.160758

[34] Gopal K. Kanji. 2006. *100 Statistical Tests* (3 ed.). Sage.

[35] E. L. Lehmann and Joseph P. Romano. 2022. *Testing Statistical Hypotheses* (4 ed.). Springer.

[36] Heng Li and Terri Johnson. 2014. Wilcoxon's Signed-rank Statistic: What Null Hypothesis and Why it Matters. *Pharmaceutical Statistics* 13, 5 (2014), 281–285. doi:10.1002/pst.1628

[37] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. 2002. The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health* 23, 1 (2002), 151–169. doi:10.1146/annurev.publhealth.23.100901.140546

[38] Bryan F. J. Manly. 2008. *Statistics for Environmental Science and Management* (2 ed.). CRC Press.

[39] William Mendenhall, Barbara M. Beaver, and Robert J. Beaver. 2018. *Introduction to Probability and Statistics* (15 ed.). Cengage.

[40] Theodore Micceri. 1989. The Unicorn, the Normal Curve, and Other Improbable Creatures. *Psychological Bulletin* 105, 1 (1989), 156–166. doi:10.1037/0033-2909.105.1.156

[41] Joel Michell. 1986. Measurement Scales and Statistics: A Clash of Paradigms. *Psychological Bulletin* 100, 3 (1986), 398–407. doi:10.1037/0033-2909.100.3.398

[42] Alistair Moffat. 2022. Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning. *IEEE Access* 10 (2022), 105564–105577. doi:10.1109/access.2022.3211668

[43] Douglas C. Montgomery and George C. Runger. 2014. *Applied Statistics and Probability for Engineers* (6 ed.). Wiley.

[44] Saralees Nadarajah. 2005. A Generalized Normal Distribution. *Journal of Applied Statistics* 32, 7 (2005), 685–694. doi:10.1080/02664760500079464

[45] Markus Neuhäuser. 2012. *Nonparametric Statistical Tests: A Computational Approach.* CRC Press.

[46] R. Lyman Ott and Michael Longnecker. 2015. *An Introduction to Statistical Methods and Data Analysis* (7 ed.). Cengage.

[47] Javier Parapar, David E. Losada, and Álvaro Barreiro. 2021. Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation. In *ACM/SIGAPP Symposium on Applied Computing.* 655–664. doi:10.1145/3412841.3441945

[48] Javier Parapar, David E. Losada, Manuel A. Presedo Quindimil, and Álvaro Barreiro. 2020. Using Score Distributions to Compare Statistical Significance Tests for Information Retrieval Evaluation. *Journal of the Association for Information Science and Technology* 71, 1 (2020), 98–113. doi:10.1002/asi.24203

[49] John W. Pratt. 1959. Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *J. Amer. Statist. Assoc.* 54, 287 (1959), 655–667. doi:10.1080/01621459.1959.10501526

[50] John W. Pratt and Jean D. Gibbons. 1981. *Concepts of Nonparametric Theory* (1 ed.). Springer.

[51] Gregory J. Privitera. 2014. *Statistics for the Behavioral Sciences* (3 ed.). Sage.

[52] John A. Rice. 2007. *Mathematical Statistics and Data Analysis* (3 ed.). Duxbury. 650 pages.

[53] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *International ACM SIGIR Conference on Research and Development in Information Retrieval.* 525–532. doi:10.1145/1148170.1148261

[54] 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. doi:10.1145/2911451.2911492

[55] Tetsuya Sakai. 2016. Two Sample T-tests for IR Evaluation: Student or Welch?. In *International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1045–1048. doi:10.1145/2911451.2914684

[56] Tetsuya Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *ACM SIGIR Forum* 54, 1 (2020), 1–8. doi:10.1145/3451964.3451976

[57] 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. doi:10.1145/1076034.1076064

[58] Jacques Savoy. 1997. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing and Management* 33, 4 (1997), 495–512. doi:10.1016/s0306-4573(97)00027-7

[59] 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. *Quantitative Methods in Psychology* 111, 2 (1992), 352–360. doi:10.1037//0033-2909.111.2.352

[60] David J. Sheskin. 2000. *Handbook of Parametric and Nonparametric Statistical Procedures* (2 ed.). Chapman & Hall.

[61] Sidney Siegel. 1956. *Nonparametric Statistics for the Behavioral Sciences* (1 ed.). McGraw Hill.

[62] B. W. Silverman. 1981. Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 43, 1 (1981), 97–99. doi:10.1111/j.2517-6161.1981.tb01155.x

[63] 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. doi:10.1145/1321440.1321528

[64] 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. doi:10.1145/1571941.1572050

[65] Robert R. Sokal and F. James Rohlf. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research* (3 ed.). W.H. Freeman.

[66] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (1908), 1–25. doi:10.2307/2331554

[67] M. Th. Subbotin. 1923. On the Law of Frequency of Error. *Matematicheskii Sbornik* 31, 2 (1923), 296–301.

[68] Jean Tague-Sutcliffe. 1992. The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing and Management* 28, 4 (jul 1992), 467–490. doi:10.1016/0306-4573(92)90005-k

[69] John W. Tukey. 1977. Modern Techniques in Data Analysis. In *NSF-sponsored regionalresearch conference at Southern Massachusetts University.*

[70] 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 *ACM SIGIR International Conference on the Theory of Information Retrieval.* 245–250. doi:10.1145/3471158.3472242

[71] 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. doi:10.1145/3331184.3331259

[72] 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. doi:10.1145/2484028.2484163

[73] 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. doi:10.1145/3209978.3210043

[74] Ivan Valiela. 2001. *Doing Science: Design, Analysis and Communication of Scientific Research* (1 ed.). Oxford University Press. doi:10.1093/oso/9780195079623.001.0001

[75] Cornelis J. van Rijsbergen. 1979. *Information Retrieval.* Butterworths. doi:10.1145/511829.511831

[76] Ellen M. Voorhees and Chris Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *International ACM SIGIR Conference on Research and Development in Information Retrieval.* 316–323. doi:10.1145/564376.564432

[77] Ellen M. Voorhees, Daniel Samarov, and Ian Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems* 36, 2 (2017). doi:10.1145/3086701

[78] Larry Wasserman. 2006. *All of Nonparametric Statistics* (1 ed.). Springer.

[79] Samaradasa Weerahandi. 2003. *Exact Statistical Methods for Data Analysis* (1 ed.). Springer.

[80] W. John Wilbur. 1994. Non-parametric Ssignificance Tests of Retrieval Performance Comparisons. *Journal of Information Science* 20, 4 (1994), 270–284. doi:10.1177/016555159402000405

[81] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. doi:10.2307/3001968

[82] 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. doi:10.1145/290941.291014

[83] Justin Zobel and Lida Rashidi. 2020. Corpus Bootstrapping for Assessment of the Properties of Effectiveness Measures. In *ACM International Conference on Information and Knowledge Management.* ACM, 1933–1952. doi:10.1145/3340531.3411998

[84] Mine Çetinkaya and Johanna Hardin. 2024. *Introduction to Modern Statistics* (2 ed.). OpenIntro.