stats

MDPI

Article

Assessing the Reproducibility Probability of Tukey's Honest Significant Difference Test via Nonparametric Predictive Inference Bootstrap

Norah D. Alshahrani

Department of Mathematics, College of Science, University of Bisha, Bisha 61922, Saudi Arabia; ndmuflih@ub.edu.sa

Abstract

Reproducibility is a fundamental component of reliability in scientific research, reflecting the extent to which statistical conclusions remain stable under repeated experimentation. This paper investigates the reproducibility probability (RP) of Tukey's Honest Significant Difference (HSD) test, a widely used post hoc procedure following one-way ANOVA, within the framework of nonparametric predictive inference (NPI), extending our earlier work on the reproducibility of the omnibus ANOVA decision to the pairwise post hoc decisions that follow it. RP is estimated via the NPI-bootstrap (NPI-B) across seven distributional scenarios (equal-variance normal, heteroscedastic normal, heavy-tailed, right-skewed (exponential and log-normal), bimodal, and discrete (Poisson) data) and sample sizes ranging from 10 to 100 per group, separately for the rejection and non-rejection regions defined by the significance threshold. Two patterns appear consistently across all scenarios. First, within the rejection region, RP for a fixed mean difference depends not only on the magnitude of the mean difference but on how far original p-values lie from the significance threshold: scenarios in which Tukey's HSD attains high power even at small sample sizes (log-normal, Poisson, and bimodal data) achieve substantially higher RP than scenarios with more modest power (equal-variance normal, heteroscedastic, and heavy-tailed data), even for identical mean differences; RP is lowest and least sensitive to sample size under heavy-tailed data. Second, within the non-rejection region, RP decreases systematically as sample size increases, without exception, across every scenario and comparison examined, indicating that non-significant decisions obtained from larger samples are, in this sense, less reproducible than those obtained from smaller ones when a true effect is present. These findings show that RP captures aspects of decision stability not reflected in the p-value or statistical power alone, and its practical use is illustrated through an application to the PlantGrowth dataset.

Keywords: reliability of statistical inference; Post Hoc Test for ANOVA; Tukey's Honest Significant Difference (HSD) test; reproducibility probability (RP); nonparametric predictive inference (NPI)

Check for updates

Academic Editor: Ayman Alzaatreh

Received: 1 August 2026
Revised: 28 August 2026
Accepted: 7 September 2026
Published: 10 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Tukey's Honest Significant Difference (HSD) test is a widely used multiple comparison procedure for assessing pairwise differences between group means. It provides simultaneous pairwise comparisons while controlling the family-wise error rate, making it an

important post hoc procedure for identifying which group means differ. For each pair of groups, the null hypothesis considered by Tukey's HSD is

$$H_0 : \mu_i = \mu_j,$$

against the alternative

$$H_1 : \mu_i \neq \mu_j, \qquad i \neq j.$$

Although a statistically significant result provides evidence against the corresponding null hypothesis, the statistical decision itself does not indicate how stable that decision would be if the experiment were repeated under comparable conditions.

In recent years, increasing attention has been directed toward the reliability and reproducibility of statistical results, which are crucial to maintaining the credibility of scientific findings. Within the framework of statistical inference, the reproducibility probability (RP) quantifies the likelihood that a test decision, such as rejecting a null hypothesis, would be obtained again under identical experimental conditions. From this perspective, RP provides information about the stability of a statistical decision rather than simply whether the corresponding null hypothesis was rejected. Although Tukey's HSD is widely used for pairwise comparisons, its reproducibility from a predictive perspective has not been systematically investigated. Assessing the reproducibility probability of Tukey's HSD therefore provides a means of evaluating the stability of its pairwise decisions under repeated sampling.

The absence of a universally accepted definition of reproducibility complicates its evaluation. According to the National Academies of Sciences, Engineering and Medicine [1], reproducibility involves consistent results using the same data and methods, while replicability refers to obtaining consistent results with new data. Several authors have emphasized the limitations of p-values for assessing reproducibility [2,3], and recent work by Simkus et al. [4] identified multiple definitions of reproducibility and promoted predictive interpretation, introducing nonparametric predictive inference (NPI) as a suitable framework.

Nonparametric predictive inference (NPI) is a frequentist methodology that relies on minimal assumptions and is designed to make predictions about future observations. NPI has been successfully applied in various domains, including diagnostic accuracy [5], finance [6], and operations research [7]. Its predictive framework is well-suited for evaluating statistical reproducibility, a perspective pioneered by Coolen and BinHimd [8,9] in the context of nonparametric tests. Subsequent research has extended NPI-based reproducibility assessments to areas such as population quantiles [10,11] and t-tests in pharmaceutical applications [12].

It is useful to clarify how RP relates to, and differs from, several neighboring concepts. First, RP should be distinguished from the broader concepts of reproducibility and replicability. Following the terminology of the National Academies of Sciences, Engineering, and Medicine [1], reproducibility concerns consistency of results obtained using the same data and methods, whereas replicability concerns consistency of results obtained using new data. In contrast, RP is a predictive quantity computed from the original sample that estimates the probability that the same statistical decision will be obtained under comparable future sampling. Thus, RP provides a quantitative predictive measure of the stability of a particular observed statistical decision.

Second, RP is distinct from statistical power. Power is a long-run, design-based property of a statistical procedure: under a specified data-generating model and true effect, it is the probability of rejecting H_0 over repeated samples. It is therefore defined with respect to the specified population and does not depend on any particular observed dataset. RP, in contrast, is conditional on the particular dataset observed and estimates the

probability that its statistical decision will be reproduced under future predictive sampling. In the present simulation study, the empirical rejection rate is used as a descriptive measure related to power. It is not the same quantity as RP. The two may be empirically related, for example through sample size and effect magnitude, but they answer different questions: power concerns the probability of rejection under a specified data-generating mechanism, whereas RP concerns the probability of reproducing the decision obtained from a particular observed dataset.

Third, RP should not be interpreted as a Bayesian posterior probability of H_0 or H_1. RP is estimated using nonparametric predictive inference (NPI), a frequentist predictive methodology that makes probability statements about future observations based on minimal assumptions (the exchangeability of future and past observations), rather than by placing a prior distribution over parameters and updating it via Bayes' theorem. Consequently, RP does not require the specification of a prior and is not a posterior probability that either hypothesis is true; rather, it is a predictive probability that the same statistical decision will be obtained again under NPI-B predictive sampling.

In the conventional one-way analysis, Tukey's HSD is applied following the omnibus ANOVA and relies on conditions such as approximate normality of the within-group errors and homogeneity of variances. Building on our previous work, which examined the reproducibility of the ANOVA decision [13], the present study focuses on the reproducibility of the pairwise decisions produced by Tukey's HSD. Thus, ANOVA is retained as the preceding omnibus step of the analysis procedure, while the reproducibility analysis focuses specifically on the subsequent pairwise Tukey decisions.

The purpose of considering different distributional settings in the present study is not to recommend the use of Tukey's HSD when the assumptions of the underlying analysis are clearly inappropriate. Rather, the different distributions and variance structures are treated as controlled data-generating mechanisms for investigating how the reproducibility of Tukey's HSD decisions changes under different sampling conditions. The simulation study therefore includes equal-variance normal, heteroscedastic normal, heavy-tailed, right-skewed, bimodal, and discrete distributions. This design allows us to examine whether distributional shape and variance structure influence the stability of Tukey HSD decisions under repeated sampling.

A central objective of this study is to investigate the role of sample size in the reproducibility of Tukey HSD decisions. We consider sample sizes ranging from $n = 10$ to $n = 100$ per group and estimate RP separately for decisions in the rejection and non-rejection regions. In addition to sample size, we examine whether reproducibility is related to the magnitude of the underlying mean difference, distributional characteristics, statistical power, and the continuous value of the original Tukey HSD p-value.

Reproducibility probability is estimated using the nonparametric predictive inference bootstrap (NPI-B). For each original dataset, one-way ANOVA is first applied, followed by Tukey's HSD for the post hoc pairwise comparisons. The same sequence of analyses is then applied to each NPI-B predictive sample. For each pairwise comparison, RP is estimated as the proportion of predictive samples that reproduce the same Tukey HSD decision as that obtained from the original dataset. Thus, although ANOVA forms part of the analysis procedure, the reproducibility measure considered in this study is defined for the individual pairwise Tukey HSD decisions.

Because the pairwise comparisons are based on the same underlying groups, the corresponding predictive samples generate dependent pairwise decisions. Accordingly, the present study evaluates the reproducibility of individual pairwise decisions rather than a joint reproducibility probability for the complete set of Tukey comparisons. The possibility

of extending the framework to a joint, dependence-aware measure of reproducibility is considered as a direction for future research.

The main contribution of this study is an NPI-B assessment of the reproducibility of Tukey's HSD pairwise decisions across a range of sample sizes and data-generating scenarios. By examining rejection and non-rejection decisions separately, we investigate the stability of significant and non-significant Tukey HSD conclusions and identify how reproducibility varies with sample size and the characteristics of the underlying data.

The remainder of this paper is organized as follows. Section 2 presents an overview of the predictive approach to statistical reproducibility within the NPI framework. Section 3 examines the reproducibility of post hoc tests, employing NPI-B methodology. Section 4 reports the results of the simulation study, Section 5 discusses these results and their implications, and Section 6 illustrates the method on a real dataset. Finally, Section 7 provides a summary and conclusions.

2. Nonparametric Predictive Inference Bootstrap (NPI-B)

Nonparametric predictive inference (NPI) [7,14] is a frequentist approach grounded on Hill's assumption $A_{(n)}$ [15], which enables direct probabilistic predictions for future observations based solely on observed data. Suppose we have n ordered real-valued observations,

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)},$$

originating from continuous and exchangeable random variables $X_1, X_2, \ldots, X_n, X_{n+1}$. For convenience, define $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$, or set $x_{(0)} = 0$ when variables are non-negative. These observations partition the real line into $n+1$ intervals:

$$I_j = (x_{(j-1)}, x_{(j)}), \quad j = 1, \ldots, n+1.$$

Hill's assumption $A_{(n)}$ states that the probability of the next observation X_{n+1} falling into any of these intervals is uniform:

$$P(X_{n+1} \in I_j) = \frac{1}{n+1}, \quad j = 1, \ldots, n+1. \tag{1}$$

This assumption can be generalized to multiple future observations through a sequential extension $A_{(.)}$ [16]. Let O_i denote one of the possible orderings of m future observations among the existing n data points, with $i = 1, \ldots, \binom{n+m}{n}$, all assumed equally likely. For each ordering O_i, let S_j^i be the number of future observations falling in interval I_j, where $j = 1, \ldots, n+1$, satisfying $\sum_{j=1}^{n+1} s_j^i = m$. The probability of ordering O_i is

$$P(O_i) = \frac{1}{\binom{n+m}{n}}.$$

The NPI framework thus considers all possible placements of future data points among the observed data, with equal probability assigned to each ordering. Although exact values of future observations within intervals are unspecified, the number of observations per interval is known, and no further assumptions on their distribution are imposed.

NPI provides imprecise probability bounds based on Hill's assumption rather than exact probabilities, making it especially useful for small or uncertain datasets [7,14]. Lower probabilities count orderings where an event must occur, while upper probabilities include orderings where the event may occur [14]. NPI has found successful applications in objective Bayesian inference [17] and is known to be exactly calibrated [18], ensuring its probabilistic inferences align with observed frequencies.

In reproducibility studies, statistical reproducibility measures the probability of obtaining the same hypothesis test outcome upon repetition under identical conditions. Using NPI, one starts by testing the null hypothesis H_0 on an original sample of size n, noting rejection or non-rejection. Then, future samples of size n are predicted by considering all possible orderings of future observations, equally likely among existing data points. For each ordering, one determines if H_0 is certainly rejected, possibly rejected, or possibly not rejected. Lower reproducibility probability is computed by counting orderings that guarantee the same decision as the original test, while upper reproducibility probability includes the possibly consistent cases [8].

While the original and future samples can differ in size, setting $m = n$ is a natural choice to ensure identical testing conditions for fair reproducibility assessment.

However, the number of possible future orderings grows combinatorially with sample size, making exact NPI reproducibility probability (NPI-RP) computations infeasible for moderate to large datasets. To mitigate this, Coolen and Marques [19] proposed a sampling approach, randomly selecting future data arrangements with equal likelihood and independence. This sampling method approximates reproducibility probabilities efficiently, with negligible bias between sampling with and without replacement, given sufficiently large samples.

An alternative computationally efficient method is the NPI bootstrap (NPI-B), introduced by Coolen and Bin Himd [9]. This resampling-based approach estimates reproducibility probabilities for nonparametric tests by generating plausible future samples within NPI's predictive framework, providing estimates bounded by the NPI lower and upper probabilities [8,9]. Unlike classical bootstrap methods, NPI-B explicitly incorporates the prediction of future observations grounded in NPI assumptions, offering a unique frequentist perspective.

The procedure for generating an NPI-B sample for real data is as follows [8,9]:

1. Randomly select an interval $I_j = (x_{(j-1)}, x_{(j)})$, for $j = 1, \ldots, n+1$.
2. If I_j is bounded, sample uniformly within I_j.
3. If I_j is unbounded (i.e., $(-\infty, x_{(1)})$ or $(x_{(n)}, \infty)$), sample from the tail of an appropriate distribution:
   - For real-valued data, sample from the tail of a normal distribution with mean $\mu = \frac{x_{(1)}+x_{(n)}}{2}$ and standard deviation $\sigma = \frac{x_{(n)}-\mu}{\Phi^{-1}\left(\frac{n}{n+1}\right)}$, where Φ^{-1} is the inverse standard normal CDF.
   - For non-negative data $(0, \infty)$, if sampling above $x_{(n)}$, draw from an exponential tail with rate $\lambda = \frac{\ln(n+1)}{x_{(n)}}$.
4. Add the sampled observation to the original dataset, now of size $n+1$.
5. Repeat steps 1–4 to generate m future observations, forming one NPI-B sample.
6. Repeat the entire process N times to generate multiple NPI-B samples.

Coolen and Bin Himd [9] use normal and exponential tail distributions to generate observations in the unbounded intervals, where uniform sampling is not possible, using the exponential tail specifically for non-negative data. These tail distributions are introduced only as a mechanism for generating predictive observations beyond the observed range and do not constitute an assumption that the underlying population follows a normal or exponential distribution.

In the present implementation, however, a truncated normal tail was used for all scenarios considered, including those involving non-negative data (Exponential, Log-normal, and Poisson), rather than switching to an exponential tail for non-negative data as originally proposed. This choice was made for implementational consistency across

scenarios. The tail parameters were calibrated to the observed sample in each case, so that the extreme intervals retain the NPI probability mass of $1/(n+1)$.

The choice of tail distribution can nevertheless influence the predictive samples generated beyond the observed range and, consequently, may affect the resulting RP estimates. This is particularly relevant for the non-negative, right-skewed scenarios (Exponential, Log-normal, Poisson) considered here, where a normal tail may not closely match the shape of the true underlying tail. A formal sensitivity analysis comparing the normal tail used here with the exponential tail originally proposed by Coolen and Bin Himd [9] for non-negative data was not undertaken in the present study and is therefore recognized as a limitation. Investigating this sensitivity would be a useful direction for future research.

The NPI-B approach can be employed to estimate reproducibility probability (NPI-B-RP) by repeatedly performing the statistical test on generated NPI-B samples and measuring consistency with the original test outcome [20]. The procedure, applicable to tests such as Wilcoxon–Mann–Whitney [20] and *t*-test [12], can be summarized as follows.

1. Conduct the statistical test on the original data.
   - Record $T = 1$ if the null hypothesis H_0 is rejected at the chosen significance level α.
   - Record $T = 0$ if H_0 is not rejected.
2. Generate N NPI-B samples from the original data.
   - Apply the same statistical test to each generated sample.
   - For each sample j, record $T_j = 1$ if H_0 is rejected, or $T_j = 0$ otherwise.
3. Calculate the reproducibility probability for this iteration:

$$RP = \frac{\sum_{j=1}^{N} \mathbb{I}_{(T=T_j)}}{N}.$$

4. Repeat Steps 2–3 for h iterations to obtain the set of reproducibility probabilities: $RP_1, RP_2, \ldots, RP_h$.

Remark on Tukey's HSD.

For a single hypothesis test, such as the Wilcoxon–Mann–Whitney test or the *t*-test, the procedure above yields one reproducibility probability per original decision. Tukey's HSD test, by contrast, produces $\binom{k}{2}$ simultaneous pairwise decisions for k groups (three, in the present study), all derived from the same one-way ANOVA fit. When the NPI-B procedure above is applied to Tukey's HSD, step 2 generates N resamples of the groups jointly rather than of a single sample, and each of the three pairwise comparisons in step 2 is then computed from the same set of N joint resamples. Consequently, the resulting reproducibility probabilities for the $Y-X$, $Z-X$, and $Z-Y$ comparisons are not statistically independent of one another: they are three different functionals of a common resampling process at the group level. In Section 4 we report RP separately for each pairwise comparison, consistent with how Tukey's HSD is applied and interpreted in practice; we return to the question of a joint, dependence-aware reproducibility measure across the full set of comparisons in Section 5.

3. Materials and Methods

3.1. Tukey's HSD Test

For k independent groups with sample means $\bar{X}_1, \ldots, \bar{X}_k$, sample sizes $n_1, \ldots, n_k$, and pooled within-group variance MS_W obtained from the one-way ANOVA, Tukey's HSD test statistic for comparing groups i and j is

$$Q_{ij} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_j}\right)/2}},$$

which is compared against the critical value $q_{\alpha,k,N-k}$ of the studentized range distribution with k groups and $N-k$ degrees of freedom, where $N = \sum_{i=1}^{k} n_i$ [21]. The null hypothesis $H_0 : \mu_i = \mu_j$ is rejected at level α if $Q_{ij} > q_{\alpha,k,N-k}$. Equivalently, an adjusted p-value p_{ij} is obtained from the studentized range distribution, and H_0 is rejected whenever $p_{ij} \leq \alpha$; this p-value formulation is used throughout the reproducibility framework below.

3.2. Simulation Design

We conducted a simulation study to investigate the reproducibility probability (RP) of Tukey's Honest Significant Difference (HSD) pairwise decisions under different sample sizes and data-generating conditions. Each simulation scenario consisted of three groups, labeled X, Y, and Z, with group-specific data distributions. The simulation scenarios and group distributions considered in this study are summarized in Table 1.

Table 1. Simulation scenarios and group distributions.

Scenario	Distribution Type	Group Definitions
Normal	Symmetric, equal variance	$X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0.3,1)$, $Z \sim \mathcal{N}(0.6,1)$
Exponential	Right-skewed	$X \sim \text{Exp}(1)$, $Y \sim \text{Exp}(1)+0.3$, $Z \sim \text{Exp}(1)+0.6$
Heavy-tailed	t-distribution	$X \sim t_3$, $Y \sim t_3+0.3$, $Z \sim t_3+0.6$
Heteroscedastic	Unequal variances (normal)	$X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0.3, 1.5^2)$, $Z \sim \mathcal{N}(0.6, 2^2)$
Bimodal	Mixture for X	$X \sim 0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(1,1)$, $Y \sim \mathcal{N}(0.5,1)$, $Z \sim \mathcal{N}(1,1)$
Log-normal	Right-skewed	$X \sim \text{Lognormal}(0,0.5)$, $Y \sim \text{Lognormal}(0.5,0.5)$, $Z \sim \text{Lognormal}(1,0.5)$
Poisson	Discrete counts	$X \sim \text{Poisson}(2)$, $Y \sim \text{Poisson}(4)$, $Z \sim \text{Poisson}(6)$

Because the Poisson scenario generates discrete observations, tied values occur naturally and are incompatible with the strictly ordered observations used in the NPI-B resampling procedure described in Section 2. Therefore, for the Poisson scenario only, a negligible random jitter was applied to the observations before the NPI-B resampling step to break ties. Specifically, independently generated zero-mean normal perturbations with standard deviation 10^{-6} were added to the observations before sorting. This magnitude is negligible relative to the minimum distance of one between distinct Poisson counts and therefore does not materially change the observed data or the intervals used for NPI-B resampling. Importantly, the original Poisson observations without jitter were used to perform the initial ANOVA and Tukey's HSD and to obtain the original p-values. The jitter was used only internally for generating the NPI-B predictive samples. The same ANOVA–Tukey procedure was then applied to the resulting predictive samples. Thus, the tie-breaking step affects only the construction of the NPI-B predictive samples and not the original statistical decisions.

Beyond the Normal scenario, the remaining six scenarios represent different data-generating conditions, including right-skewness (Exponential and Log-normal), heavy tails (t-distribution), unequal variances (Heteroscedastic), multimodality (Bimodal), and discreteness (Poisson). These scenarios are not intended to recommend the application of the ANOVA–Tukey procedure when its underlying assumptions are clearly violated. Rather, they are deliberately included as controlled data-generating mechanisms to investigate how the reproducibility probability (RP) of the resulting Tukey HSD pairwise decisions

changes under different distributional and variance conditions. The complete analysis procedure, consisting of one-way ANOVA followed by Tukey's HSD, is therefore applied consistently across all simulation scenarios. This design allows us to assess the sensitivity of the reproducibility of Tukey HSD decisions to departures from the conditions under which the procedure is conventionally justified.

Each scenario was simulated using sample sizes $n = 10, 30, 60$, and 100 per group. For each configuration, we generated 100 original datasets. The level of significance was set to $\alpha = 0.05$. For the reproducibility assessment described below, we used $N = 1000$ NPI-B resamples per iteration and $h = 100$ Monte Carlo repetitions of the resampling procedure, so that each reported RP value in Section 4 is the average of 100 independent RP estimates, each itself based on 1000 NPI-B resamples.

Throughout this paper, the term effect size refers to the magnitude of the true differences between group means specified in the simulation scenarios, rather than to a standardized effect size measure such as Cohen's d or Cohen's f.

For descriptive assessment of the ability of Tukey's HSD to detect the specified nonzero mean differences, we calculated the empirical rejection rate for each scenario, sample size, and pairwise comparison. This rate was defined as the proportion of the 100 original simulated datasets for which the original Tukey HSD p-value satisfied $p_{\text{orig}} \leq 0.05$. These empirical rejection rates, reported as power alongside the RP results in Section 4, are used only to aid the interpretation of the RP results.

3.3. Reproducibility Framework

For each original dataset:

1. One-way ANOVA was applied, followed by Tukey's Honest Significant Difference (HSD) test for post hoc pairwise comparisons.
2. The original p-values (p_{orig}) from the Tukey test were recorded.
3. $N = 1000$ NPI-B samples were generated jointly for the three groups, following the procedure in Section 2.
4. The same statistical tests (ANOVA followed by Tukey's HSD) were applied to each of the N NPI-B samples, and the resulting bootstrap p-values (p_{boot}) were recorded for each pairwise comparison.
5. The reproducibility probability (RP) was defined, separately for each pairwise comparison, as:
$$\text{RP} = \begin{cases} \dfrac{\#(p_{\text{boot}} > \alpha)}{N}, & \text{if } p_{\text{orig}} > \alpha \\ \dfrac{\#(p_{\text{boot}} \leq \alpha)}{N}, & \text{if } p_{\text{orig}} \leq \alpha \end{cases}$$
6. Steps 3–5 were repeated $h = 100$ times to generate $RP_1, RP_2, \ldots, RP_h$ for each pairwise comparison, and the reported RP value is the average $\bar{RP} = \frac{1}{h}\sum_{l=1}^{h} RP_l$.

4. Results

This section presents the reproducibility probability (RP) of Tukey's HSD test across the seven simulation scenarios, separately for the rejection region ($p_{\text{orig}} \leq 0.05$, Table 2) and the non-rejection region ($p_{\text{orig}} > 0.05$, Table 3), since averaging RP across both regions would obscure opposing trends, as discussed below.

Table 2. Mean reproducibility probability (RP), averaged across simulated datasets whose original Tukey HSD decision was to reject H_0 ($p_{\text{orig}} \leq 0.05$), by scenario, sample size, and pairwise comparison. Values in parentheses denote the standard deviation across these datasets. Cells marked "–" indicate that no simulated dataset fell into the rejection region for that comparison and sample size.

Scenario	Comparison	Sample Size n			
		$n = 10$	$n = 30$	$n = 60$	$n = 100$
Normal (equal var.)	Y–X	0.44 (0.03)	0.63 (0.15)	0.61 (0.09)	0.68 (0.11)
	Z–X	0.57 (0.10)	0.64 (0.13)	0.72 (0.13)	0.87 (0.12)
	Z–Y	0.54 (n=1)	0.60 (0.10)	0.60 (0.10)	0.67 (0.11)
Heteroscedastic	Y–X	–	0.61 (0.14)	0.57 (0.09)	0.56 (0.09)
	Z–X	0.56 (0.08)	0.59 (0.11)	0.59 (0.11)	0.69 (0.14)
	Z–Y	0.57 (0.14)	0.60 (0.07)	0.57 (0.09)	0.61 (0.09)
Heavy-tailed (t_3)	Y–X	0.57 (0.13)	0.56 (0.10)	0.62 (0.11)	0.59 (0.12)
	Z–X	0.66 (0.17)	0.62 (0.13)	0.66 (0.15)	0.67 (0.15)
	Z–Y	0.53 (0.06)	0.59 (0.10)	0.60 (0.15)	0.59 (0.09)
Exponential	Y–X	0.56 (0.06)	0.66 (0.18)	0.65 (0.14)	0.67 (0.13)
	Z–X	0.63 (0.12)	0.66 (0.14)	0.73 (0.16)	0.87 (0.11)
	Z–Y	0.60 (0.09)	0.60 (0.09)	0.58 (0.08)	0.70 (0.12)
Log-normal	Y–X	0.51 (0.05)	0.66 (0.16)	0.83 (0.14)	0.94 (0.08)
	Z–X	0.79 (0.13)	0.99 (0.01)	1.00 (0.00)	1.00 (0.00)
	Z–Y	0.62 (0.11)	0.83 (0.13)	0.95 (0.07)	0.99 (0.02)
Bimodal	Y–X	0.56 (0.06)	0.62 (0.13)	0.66 (0.13)	0.74 (0.13)
	Z–X	0.57 (0.11)	0.75 (0.14)	0.89 (0.11)	0.98 (0.04)
	Z–Y	0.51 (0.08)	0.63 (0.13)	0.73 (0.13)	0.73 (0.12)
Poisson	Y–X	0.56 (0.13)	0.81 (0.15)	0.98 (0.03)	1.00 (0.00)
	Z–X	0.84 (0.11)	1.00 (0.01)	1.00 (0.00)	1.00 (0.00)
	Z–Y	0.65 (0.14)	0.77 (0.16)	0.94 (0.08)	0.99 (0.03)

Table 3. Mean reproducibility probability (RP), averaged across simulated datasets whose original Tukey HSD decision was to not reject H_0 ($p_{\text{orig}} > 0.05$), by scenario, sample size, and pairwise comparison. Values in parentheses denote the standard deviation across these datasets. Cells marked "–" indicate that no simulated dataset fell into the non-rejection region for that comparison and sample size.

Scenario	Comparison	Sample Size n			
		$n = 10$	$n = 30$	$n = 60$	$n = 100$
Normal (equal var.)	Y–X	0.84 (0.08)	0.81 (0.09)	0.74 (0.13)	0.73 (0.11)
	Z–X	0.79 (0.10)	0.72 (0.10)	0.62 (0.08)	0.67 (0.16)
	Z–Y	0.83 (0.08)	0.79 (0.10)	0.77 (0.11)	0.72 (0.11)
Heteroscedastic	Y–X	0.90 (0.06)	0.89 (0.08)	0.83 (0.11)	0.83 (0.12)
	Z–X	0.81 (0.09)	0.76 (0.10)	0.73 (0.10)	0.68 (0.11)
	Z–Y	0.79 (0.07)	0.77 (0.09)	0.77 (0.09)	0.73 (0.12)
Heavy-tailed (t_3)	Y–X	0.87 (0.07)	0.83 (0.09)	0.82 (0.10)	0.79 (0.12)
	Z–X	0.84 (0.08)	0.80 (0.11)	0.74 (0.12)	0.72 (0.11)
	Z–Y	0.85 (0.08)	0.84 (0.08)	0.80 (0.11)	0.78 (0.11)
Exponential	Y–X	0.84 (0.08)	0.80 (0.10)	0.76 (0.10)	0.73 (0.12)
	Z–X	0.80 (0.10)	0.69 (0.11)	0.69 (0.12)	0.49 (n=1)
	Z–Y	0.84 (0.10)	0.82 (0.10)	0.77 (0.11)	0.70 (0.11)
Log-normal	Y–X	0.83 (0.11)	0.73 (0.12)	0.63 (0.07)	–
	Z–X	0.79 (0.15)	–	–	–
	Z–Y	0.83 (0.11)	0.73 (0.12)	0.63 (0.07)	–
Bimodal	Y–X	0.80 (0.09)	0.77 (0.11)	0.73 (0.12)	0.65 (0.12)
	Z–X	0.75 (0.09)	0.66 (0.11)	0.71 (n = 2)	–
	Z–Y	0.87 (0.08)	0.77 (0.12)	0.68 (0.10)	0.65 (0.09)
Poisson	Y–X	0.76 (0.11)	0.69 (0.02)	–	–
	Z–X	0.78 (n = 1)	–	–	–
	Z–Y	0.69 (0.10)	0.62 (0.08)	–	–

Where the standard deviation could not be computed because only one or two simulated datasets fell into the rejection or non-rejection region (as applicable), the num-

ber of cases is reported instead (denoted $n = 1$ or $n = 2$). This applies to cells in both Tables 2 and 3.

The empirical rejection rates (power) of Tukey's HSD across the simulation scenarios, sample sizes, and pairwise comparisons are summarized in Table 4.

Table 4. Empirical rejection rate (power) of Tukey's HSD, defined as the proportion of the 100 simulated datasets for which $p_{\text{orig}} \leq 0.05$, by scenario, sample size, and pairwise comparison. These rates underlie the power-based interpretation of RP discussed in Sections 4 and 5.

Scenario	Comparison	Sample Size n			
		$n = 10$	$n = 30$	$n = 60$	$n = 100$
Normal (equal var.)	Y–X	0.02	0.13	0.24	0.35
	Z–X	0.17	0.54	0.81	0.97
	Z–Y	0.01	0.14	0.18	0.45
Heteroscedastic	Y–X	0.00	0.06	0.05	0.15
	Z–X	0.09	0.23	0.40	0.69
	Z–Y	0.04	0.10	0.09	0.21
Heavy-tailed (t_3)	Y–X	0.05	0.07	0.05	0.12
	Z–X	0.05	0.13	0.28	0.58
	Z–Y	0.05	0.04	0.08	0.14
Exponential	Y–X	0.08	0.10	0.21	0.43
	Z–X	0.19	0.42	0.84	0.99
	Z–Y	0.07	0.14	0.23	0.51
Log-normal	Y–X	0.13	0.50	0.95	1.00
	Z–X	0.96	1.00	1.00	1.00
	Z–Y	0.56	0.96	1.00	1.00
Bimodal	Y–X	0.10	0.30	0.52	0.83
	Z–X	0.38	0.82	0.98	1.00
	Z–Y	0.05	0.18	0.44	0.81
Poisson	Y–X	0.41	0.98	1.00	1.00
	Z–X	0.99	1.00	1.00	1.00
	Z–Y	0.48	0.93	1.00	1.00

4.1. Scenario: Normal Distribution (Equal Variances)

Figure 1 displays the raw relationship between RP and p_{orig} for each pairwise comparison and sample size, and, together with Table 2, presents RP for sample sizes $n = 10, 30, 60$, and 100 under equal-variance normal data. Within the rejection region, RP for the Z–X comparison, which has the largest mean difference (0.6), increases from 0.57 at $n = 10$ to 0.87 at $n = 100$. The Y–X and Z–Y comparisons, with the smaller mean difference of 0.3, show lower and more variable RP throughout, reaching only 0.68 and 0.67, respectively, at $n = 100$. Within the non-rejection region (Table 3), RP declines steadily as n increases for all three comparisons (for instance, from 0.84 to 0.73 for Y–X) reflecting the fact that the datasets remaining in this region at larger n increasingly correspond to atypical samples that failed to detect an effect that is, in fact, present.

To complement the binary rejection/non-rejection summaries reported above, Figure 2 displays RP as a continuous function of p_{orig} for the Normal (equal-variance) scenario. Rather than changing abruptly at $\alpha = 0.05$, RP varies smoothly with p_{orig}: it tends to be highest when p_{orig} is close to either 0 or 1, and lowest for p_{orig} values near the significance threshold, where the original Tukey HSD decision is most sensitive to resampling. This pattern indicates that RP conveys information about the stability of a Tukey HSD decision beyond what is captured by simply classifying p_{orig} as above or below α, and is consistent with the separate treatment of the rejection and non-rejection regions adopted in Tables 2 and 3.

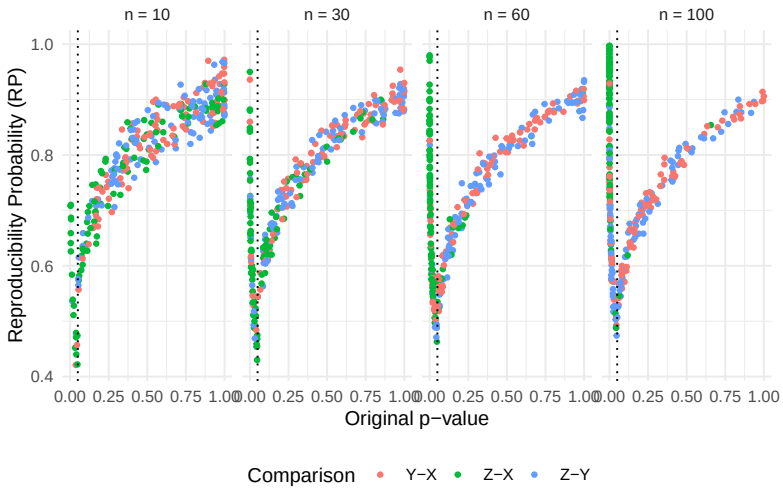


Figure 1. The relationship between the reproducibility probability (RP) and the p-value for Tukey's Honest Significant Difference (HSD) test, when data are sampled from $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0.3,1)$, and $Z \sim \mathcal{N}(0.6,1)$ at a significance level of $\alpha = 0.05$.

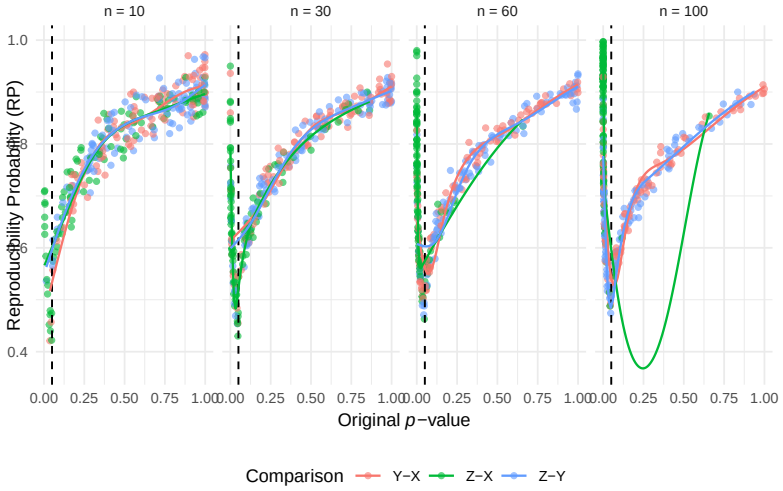


Figure 2. The relationship between the reproducibility probability (RP) and the original Tukey HSD p-value (p_{orig}), for the Normal (equal-variance) scenario, by sample size and pairwise comparison. The dashed vertical line marks the significance threshold $\alpha = 0.05$. Curves show binomial GLM fits (natural spline, 4 degrees of freedom, weighted by the $N = 1000$ NPI-B resamples underlying each RP value), which respect the $[0,1]$ range of RP by construction.

4.2. Scenario: Heteroscedastic (Unequal Variances)

Figure 3 shows the relationship between RP and Tukey's HSD p-values under unequal variances. Introducing unequal variances lowers RP within the rejection region relative to the equal-variance case, particularly for Z–X, which reaches only 0.69 at $n = 100$ (compared with 0.87 under equal variances), despite having the same mean difference. This reflects the combined effect of a large mean difference and a large variance disparity acting in opposite directions. Between the equal-mean-difference comparisons, Z–Y attains marginally higher RP than Y–X at most sample sizes, consistent with Z–Y comparing two groups whose variances (1.5^2 and 2^2) are both larger and more similar to each other than the variance disparity in Y–X (1^2 vs. 1.5^2). In the non-rejection region, RP again declines with n across all three comparisons, following the same pattern observed for the normal scenario, though starting from a slightly higher level (e.g., 0.90 for Y–X at $n = 10$).

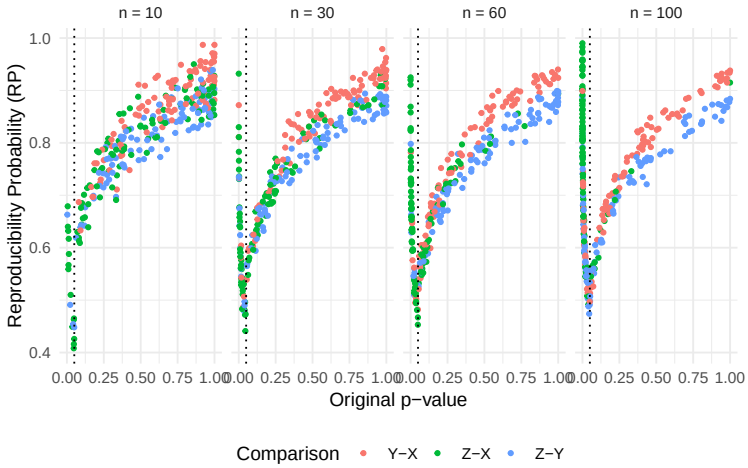


Figure 3. The relationship between RP and Tukey's HSD p-values, when data are sampled from normal distributions with unequal variances: $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0.3, 1.5^2)$, and $Z \sim \mathcal{N}(0.6, 2^2)$ at $\alpha = 0.05$.

4.3. Scenario: Heavy-Tailed Distribution

Figure 4 shows the relationship between RP and Tukey's HSD p-values under heavy-tailed data. The heavy-tailed t_3 scenario shows the weakest reproducibility in the rejection region among all seven scenarios: RP for Z–X ranges only from 0.62 to 0.67 across all sample sizes, showing essentially no improvement with n, and both Y–X and Z–Y remain close to 0.6 throughout. This coincides with the lowest rejection rates observed in any scenario (Table 4): even at $n = 100$, only 58% of datasets reject H_0 for Z–X, compared with 69–100% in every other scenario. The heavy tails of the t_3 distribution thus not only reduce the power of Tukey's HSD but also mean that, when rejection does occur, it is frequently a borderline decision with limited reproducibility. In the non-rejection region, RP again decreases with n, from 0.87 to 0.79 for Y–X, consistent with the pattern found in the previous scenarios.

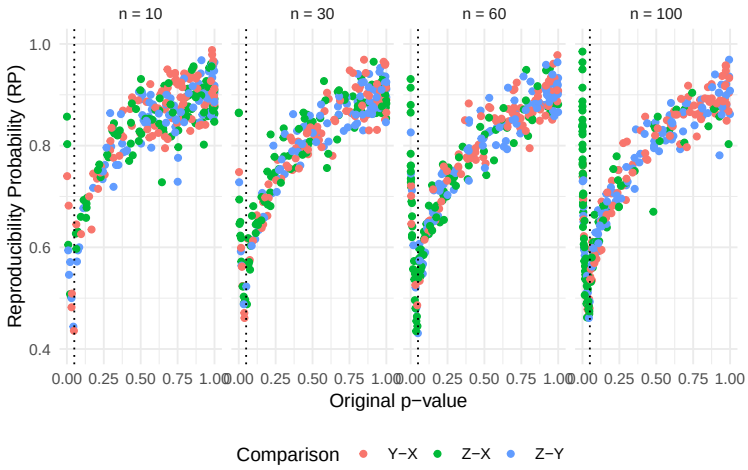


Figure 4. The relationship between the reproducibility probability (RP) and p-values from Tukey's HSD test, when data are sampled from heavy-tailed distributions: $X \sim t_3$, $Y \sim t_3 + 0.3$, and $Z \sim t_3 + 0.6$, with 3 degrees of freedom and significance level $\alpha = 0.05$.

4.4. Scenario: Right-Skewed Distributions

Figure 5 shows the relationship between RP and Tukey's HSD p-values for the exponential scenario. For the Z–X in the rejection region rises from 0.63 at $n = 10$ to 0.87 at

$n = 100$, closely tracking the pattern observed for the equal-variance normal case. The Y–X and Z–Y comparisons remain lower throughout, between 0.56 and 0.70. In the non-rejection region, RP declines with n as in all previous scenarios; at $n = 100$, the Z–X estimate (0.49) is based on a single simulated dataset that failed to reject H_0 out of 100 (Table 3), so this value should be interpreted with caution rather than as a stable estimate.

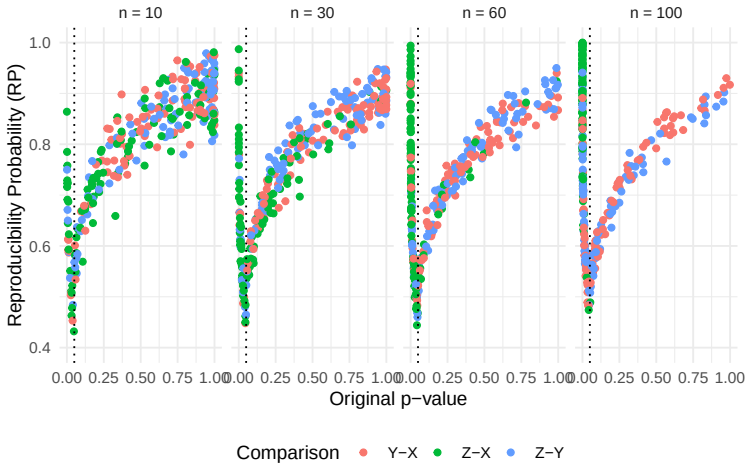


Figure 5. The relationship between the reproducibility probability (RP) and the p-value for Tukey's Honest Significant Difference (HSD) test, when data are sampled from right-skewed distributions: $X \sim \text{Exp}(1)$, $Y \sim \text{Exp}(1) + 0.3$, and $Z \sim \text{Exp}(1) + 0.6$, at a significance level of $\alpha = 0.05$.

Figure 6 shows the relationship between RP and Tukey's HSD p-values for the log-normal scenario. This scenario exhibits the strongest reproducibility in the rejection region of all seven scenarios considered: RP for Z–X reaches 0.99 already at $n = 30$ and 1.00 from $n = 60$ onward, and even the smaller-effect comparisons Y–X and Z–Y reach 0.94 and 0.99, respectively, at $n = 100$. This coincides with very high rejection rates from small sample sizes onward (e.g., 96% for Z–X already at $n = 10$), so that most rejections in this scenario correspond to p-values well below the significance threshold rather than borderline decisions. In the non-rejection region, data become sparse rapidly with increasing n: for Z–X, only 4% of datasets fail to reject H_0 even at $n = 10$, and none do so from $n = 30$ onward, leaving too few observations for a stable RP estimate beyond $n = 10$ for this comparison.

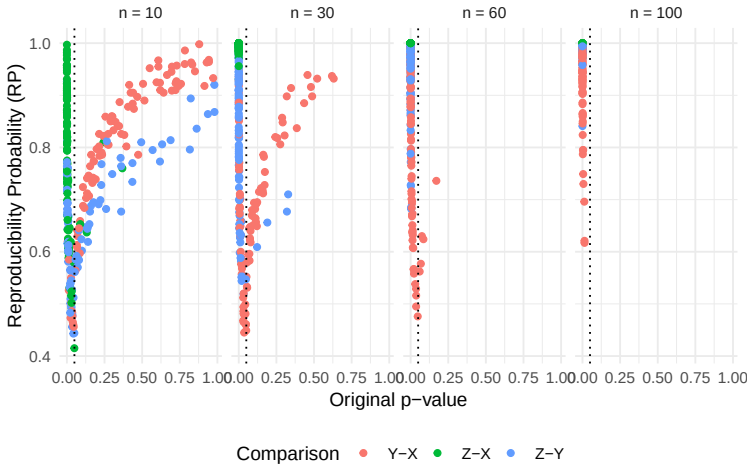


Figure 6. The relationship between NPI-B reproducibility probability (RP) and Tukey's HSD p-values, when data are sampled from log-normal distributions with equal variance: $X \sim \text{Lognormal}(0, 0.5)$, $Y \sim \text{Lognormal}(0.5, 0.5)$, and $Z \sim \text{Lognormal}(1, 0.5)$, at $\alpha = 0.05$.

4.5. Scenario: Bimodal Distribution

Figure 7 shows the relationship between RP and Tukey's HSD p-values for the bimodal scenario. The bimodal scenario exhibits an intermediate pattern between the log-normal/Poisson scenarios and the normal/heteroscedastic/heavy-tailed scenarios. In the rejection region, RP for the Z–X comparison increases from 0.57 at $n = 10$ to 0.98 at $n = 100$, approaching near-perfect reproducibility for the largest-effect comparison. In contrast, RP for the Y–X and Z–Y comparisons remains considerably lower, reaching 0.74 and 0.73, respectively, at $n = 100$.

In the non-rejection region, RP again decreases as n increases. For the Z–X comparison, the RP estimate at $n = 60$ is 0.71 and is based on only two simulated datasets that failed to reject H_0. Therefore, this estimate should be interpreted with caution. At $n = 100$, no datasets remain in the non-rejection region for this comparison, and hence no stable RP estimate is available.

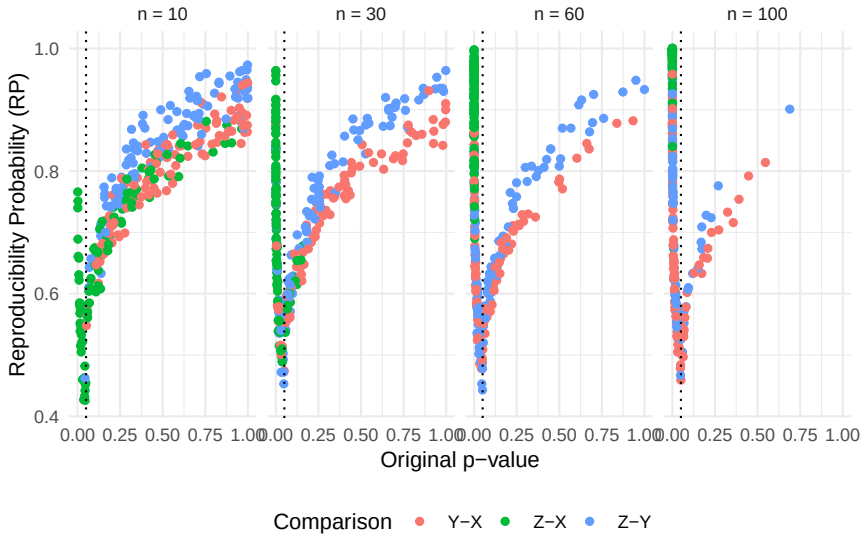


Figure 7. The relationship between RP values and p-values for Tukey's HSD test, when group X is sampled from a bimodal distribution (mixture of $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$), while $Y \sim \mathcal{N}(0.5, 1)$ and $Z \sim \mathcal{N}(1, 1)$, at a significance level of $\alpha = 0.05$.

4.6. Scenario: Discrete Data

Figure 8 shows the relationship between RP and Tukey's HSD p-values for the Poisson scenario. Among the scenarios considered, the Poisson scenario exhibits the highest reproducibility in the rejection region. RP for Z–X reaches 1.00 from $n = 30$ onward, and even Y–X, despite its smaller mean difference, reaches 1.00 at $n = 100$. As in the log-normal case, this coincides with very high rejection rates from small n onward (99% for Z–X already at $n = 10$). This pattern is consistent with the intrinsic mean–variance relationship of the Poisson distribution: groups with larger means (here, Z with rate 6) are also more variable, which increases the magnitude of typical mean differences relative to their standard errors as the rate parameter grows across groups. In the non-rejection region, data again become sparse quickly, with no datasets remaining in this region beyond $n = 30$ for Y–X and Z–Y, and beyond $n = 10$ for Z–X.

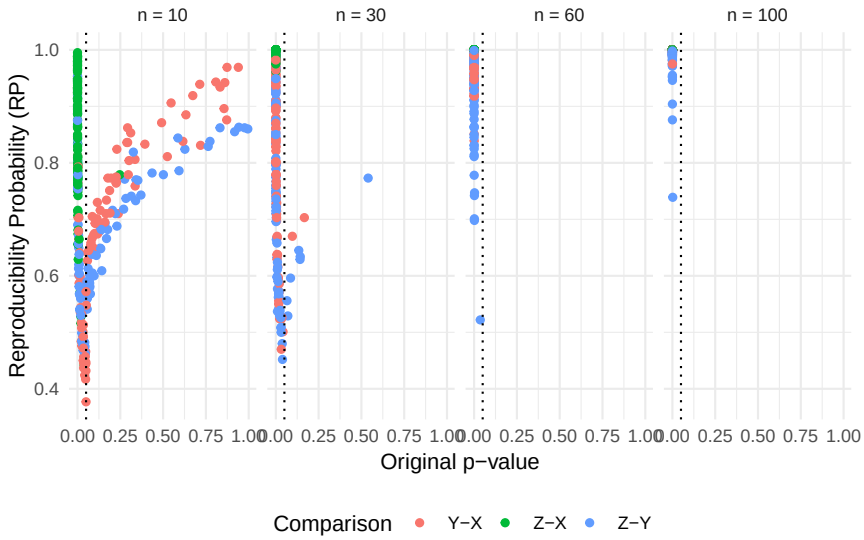


Figure 8. The relationship between RP values and Tukey's HSD p-values, when data are generated from discrete Poisson distributions: $X \sim \text{Poisson}(2)$, $Y \sim \text{Poisson}(4)$, and $Z \sim \text{Poisson}(6)$, at a significance level of $\alpha = 0.05$.

4.7. Synthesis: Two Cross-Scenario Patterns

The simulation results reveal that reproducibility is shaped jointly by statistical power and by the position of the original p-value relative to the significance threshold, rather than by either factor in isolation. Considering the results across all seven scenarios reveals two general principles that provide a unified interpretation of the simulation findings.

First, within the rejection region, RP is not determined by the magnitude of the mean difference alone. Comparisons with identical mean differences can exhibit markedly different reproducibility because the underlying distributional characteristics influence the statistical power of Tukey's HSD. Scenarios in which Tukey's HSD achieves high power even for relatively small sample sizes (log-normal, Poisson, and, to a lesser extent, bimodal data) exhibit substantially higher RP than scenarios with more modest power (equal-variance normal, heteroscedastic, heavy-tailed, and exponential data); see Table 4 for the corresponding empirical rejection rates. This occurs because high-power settings produce rejection decisions associated with p-values well below α, whereas lower-power settings produce a larger proportion of borderline rejections with p-values close to the significance threshold, which are inherently less reproducible under repeated sampling.

Second, within the non-rejection region, RP decreases systematically as sample size increases across every scenario and pairwise comparison considered. Since all simulated scenarios involve genuine non-zero mean differences, increasing the sample size makes failure to reject H_0 an increasingly unusual outcome. Consequently, the datasets remaining in the non-rejection region at larger sample sizes represent a progressively more selective subset of atypical samples in which the true effect was missed by chance. Such atypical non-rejection decisions are, by construction, less likely to be reproduced under resampling than the more typical non-rejection outcomes observed for smaller sample sizes, where modest effects are naturally more difficult to detect.

Taken together, these findings demonstrate that reproducibility probability provides additional insight beyond statistical power and conventional significance testing. Whereas statistical power quantifies the probability of detecting a true effect, RP quantifies the probability that the original statistical decision will be reproduced under future sampling. Consequently, RP serves as a valuable complementary diagnostic for assessing the stability and reliability of post hoc conclusions.

5. Discussion

The results collectively show that RP is governed by more than sample size and effect size alone. Consistent with expectations, RP in the rejection region generally improves with sample size and is highest for the comparison with the largest mean difference across all scenarios. However, the magnitude of RP for a comparison with a fixed mean difference varies considerably across distributional settings: heavy-tailed and heteroscedastic data substantially depress RP relative to the equal-variance normal case, while skewed and discrete distributions with high inherent power (log-normal, Poisson) can produce *higher* RP than the normal case for the same nominal effect size. This indicates that distributional shape and variance structure modify reproducibility through their effect on the distribution of observed p-values relative to the significance threshold, rather than through effect size alone.

The systematic decline of RP with sample size in the non-rejection region, observed consistently across all seven scenarios, is consistent with the pattern we previously observed for the reproducibility of individual hypothesis tests [12,13], and the present study demonstrates that this behavior extends to the pairwise post hoc decisions of Tukey's HSD. The pattern follows directly from the fact that all simulated scenarios in this study involve non-zero true mean differences: as n increases, failing to detect these differences becomes an increasingly atypical outcome, and such atypical outcomes are correspondingly less stable under resampling. This suggests that practitioners should treat a non-significant Tukey HSD result obtained from a large sample with particular caution, since, if a true difference exists, such a result is less likely to be reproduced under future sampling.

Finally, we note that the reproducibility probabilities examined in Section 4 were computed separately for each pairwise comparison, following standard practice for Tukey's HSD. As discussed in Section 2, however, the NPI-bootstrap resamples underlying the Y–X, Z–X, and Z–Y comparisons for a given dataset are not generated independently, since all three comparisons are computed from a common set of resampled groups. A joint, dependence-aware treatment of reproducibility across the full set of pairwise comparisons (for instance, assessing the probability that *all* comparisons are simultaneously reproduced, analogous to family-wise error control in the original test) remains an open question for future research.

6. Real Data Application

To illustrate the practical use of reproducibility probability, a real dataset comprising three experimental groups is considered. The analysis is conducted on the original dataset, reflecting a common practice in applied research. A one-way ANOVA is first applied to assess overall group differences, followed by Tukey's Honest Significant Difference (HSD) test for post hoc pairwise comparisons. Reproducibility probability (RP) is then estimated via the NPI-bootstrap method described in Section 3, in which bootstrap samples are generated conditionally on the observed data. RP quantifies the probability that a given post hoc decision would be reproduced under repeated sampling.

The `PlantGrowth` dataset consists of plant weight measurements from three experimental groups (a control group and two treatment groups), with ten observations per group. The data are approximately symmetric, showing moderate variability and no severe departures from normality. The one-way ANOVA indicates a significant overall group effect, motivating the use of Tukey's HSD for pairwise comparisons. The post hoc analysis identifies a statistically significant difference between the two treatment groups, whereas the remaining comparisons are not significant at the 5% level.

Table 5 reports the original Tukey HSD p-values together with the corresponding reproducibility probabilities for the `PlantGrowth` dataset. Although the comparison between

the two treatment groups (`trt2` and `trt1`) is statistically significant at the 5% level, its reproducibility probability is relatively low (RP = 0.584), indicating instability of the post hoc decision under repeated sampling. In contrast, comparisons involving the control group, which yield p-values farther from the significance threshold, exhibit higher reproducibility probabilities, reflecting more stable non-rejection decisions. This pattern is consistent with the simulation findings of Section 4: a rejection decision with a p-value close to α (here, $p_{\text{orig}} = 0.012$, only marginally below 0.05) tends to be less reproducible than a non-rejection decision with a p-value farther from the threshold, even though the latter corresponds to a non-significant result. This example demonstrates how RP can complement conventional inference by highlighting differences in decision stability in a typical applied setting based on a single observed dataset.

Table 5. Real data illustration using the `PlantGrowth` dataset: original Tukey HSD p-values, corresponding decisions, and reproducibility probabilities (RP) estimated via NPI-bootstrap ($N = 1000$ resamples, averaged over $h = 100$ repetitions). `ctrl` denotes the control group; `trt1` and `trt2` denote the two treatment groups.

Comparison	p_{orig}	Decision	RP
trt1–ctrl	0.3909	Do not reject H_0	0.755
trt2–ctrl	0.1980	Do not reject H_0	0.760
trt2–trt1	0.0120	Reject H_0	0.584

7. Conclusions

This study investigated the reproducibility probability (RP) of pairwise decisions from Tukey's Honest Significant Difference (HSD) procedure using the nonparametric predictive inference bootstrap (NPI-B). The analysis procedure consisted of one-way ANOVA followed by Tukey's HSD for post hoc pairwise comparisons, and the same procedure was applied to the NPI-B predictive samples. Building on our earlier work on the reproducibility of the omnibus ANOVA decision [13], the present study focuses on the reproducibility of the individual pairwise decisions produced by the subsequent Tukey HSD procedure. RP was estimated separately for rejection and non-rejection decisions across seven data-generating scenarios (equal-variance normal, heteroscedastic normal, heavy-tailed (t_3), right-skewed (exponential and log-normal), bimodal, and discrete (Poisson) data) and sample sizes ranging from $n = 10$ to $n = 100$ per group.

Two main findings emerged consistently across the simulation scenarios. First, within the rejection region, RP was not determined solely by whether the original Tukey HSD decision was significant. Rather, RP varied across the continuous range of the original p-values. In general, comparisons with p_{orig} farther below the significance threshold exhibited higher reproducibility, whereas decisions based on p_{orig} values close to $\alpha = 0.05$ were less reproducible. Thus, the original p-value provides additional information about the stability of the Tukey HSD decision beyond the binary rejection/non-rejection classification. Scenarios with high empirical rejection probabilities at small sample sizes, particularly log-normal and Poisson data and, to a lesser extent, bimodal data, generally exhibited higher RP than scenarios with lower empirical rejection probabilities, including equal-variance normal, heteroscedastic, heavy-tailed, and exponential data. This pattern was observed even when the specified population mean difference was the same across scenarios. The heavy-tailed scenario showed the lowest reproducibility among the scenarios considered, with RP for the largest-effect comparison remaining approximately 0.66–0.67 across the sample sizes examined. These empirical rejection probabilities were used as descriptive measures of the ability of the ANOVA–Tukey analysis procedure to detect the specified non-zero pairwise differences under the controlled simulation conditions.

Second, within the non-rejection region, RP decreased systematically as sample size increased across the scenarios and pairwise comparisons examined. Since all simulated scenarios involved genuine non-zero mean differences, non-rejection represented an increasingly less common outcome in those settings where the empirical rejection rate increased with sample size. Consequently, non-rejection decisions could become less stable under predictive resampling as the sample size increased. Consequently, the non-rejection decisions observed at larger sample sizes were less frequently reproduced by the NPI-B predictive samples. This finding indicates that a non-significant Tukey HSD decision in a large sample does not necessarily have a higher reproducibility probability than a non-significant decision obtained from a smaller sample. In particular, when a genuine non-zero difference is present, the reproducibility of a non-rejection decision should be interpreted in relation to the underlying data-generating conditions and sample size.

Taken together, these findings indicate that RP provides information that complements the conventional significance decision based on the Tukey HSD p-value and the empirical rejection probability. Whereas the conventional p-value determines whether a pairwise difference is statistically significant at the specified significance level, RP provides an additional predictive measure of the stability of that pairwise decision under future sampling.

The simulation results should be interpreted in light of the purpose of the non-standard data-generating scenarios. The inclusion of non-normal, heteroscedastic, heavy-tailed, bimodal, and discrete distributions was not intended to recommend the routine use of the ANOVA–Tukey procedure when its underlying assumptions are clearly inappropriate. Rather, these scenarios were deliberately used as controlled data-generating mechanisms to examine how departures from the conditions under which the procedure is conventionally justified affect the reproducibility of its pairwise decisions. Thus, the results describe the behavior of RP for the specified analysis procedure under these controlled conditions and should not be interpreted as evidence that the procedure is preferable to methods designed for data that violate its assumptions.

This study has several limitations that suggest directions for future research. First, the NPI-B predictive samples underlying the three pairwise comparisons are generated jointly from the same resampled groups. Consequently, the three pairwise decisions are dependent. The present study treated each pairwise comparison as the unit of reproducibility, consistent with the pairwise nature of Tukey's HSD, but did not quantify this dependence directly. A natural extension would be to investigate a joint reproducibility measure for the complete set of Tukey comparisons, such as the probability that all pairwise decisions are simultaneously reproduced.

Second, the simulation used 100 original datasets for each combination of scenario and sample size. Although this provides a common simulation size across all configurations, some rejection or non-rejection regions were represented by only one or two original datasets, particularly when the empirical rejection probability was close to 0 or 1. Consequently, RP summaries for these sparsely populated regions should be interpreted cautiously, and their associated variability cannot be estimated reliably. The $N = 1000$ NPI-B resamples and $h = 100$ repetitions improve the precision of the RP calculation for each original dataset but do not compensate for the small number of independent original datasets within a particular decision region. Increasing the number of original datasets, or allocating additional simulation effort to sparsely populated regions, would improve the precision and stability of these estimates.

Another limitation concerns the treatment of the unbounded intervals in the NPI-B construction. The present study uses a truncated normal distribution, with parameters calibrated from the observed sample, to generate predictive observations beyond the ob-

served range, for all scenarios considered. Although this choice is used only for generating predictive observations beyond the observed range and not as an assumption about the underlying population distribution, alternative tail specifications (e.g., an exponential-based extrapolation, which may be more suitable for non-negative, right-skewed data) could potentially affect the resulting RP estimates. A sensitivity analysis comparing different plausible tail distributions was not conducted in this study and would therefore be a useful direction for future research.

Another limitation is that the present simulation considered equal sample sizes across the three groups and fixed population mean differences within each scenario. Local alternatives, in which the magnitude of the population mean differences changes with sample size, were not considered. Likewise, unequal or unbalanced sample sizes across groups were outside the scope of the current study. Investigating these settings in future work would provide a broader assessment of the reproducibility of Tukey's HSD decisions, particularly with respect to how RP behaves under alternatives that become increasingly difficult to detect and under unbalanced sampling designs.

Overall, the results demonstrate that the reproducibility of Tukey HSD pairwise decisions is associated with sample size, the magnitude of the underlying mean difference, variance structure, and distributional shape. The proposed NPI-B framework provides a distribution-free predictive perspective that complements conventional Tukey HSD inference by evaluating not only whether a pairwise decision is statistically significant, but also how likely that decision is to be reproduced under future sampling.

Funding: This research received no external funding. The APC was funded by the author.

Data Availability Statement: The data supporting the findings of this study were generated through simulation and are available from the author upon reasonable request.

Acknowledgments: The author acknowledges the Fast-Track Research Support Program of the Deanship of Graduate Studies and Scientific Research at the University of Bisha.

Conflicts of Interest: The author states that there are no conflicts of interest.

References

1. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*; National Academies Press: Washington, DC, USA, 2019. [CrossRef]
2. Goodman, S.N. A comment on replication, *p*-values and evidence. *Stat. Med.* **1992**, *11*, 875–879. [CrossRef]
3. Senn, S. A comment on 'A comment on replication *p*-values and evidence'. *Stat. Med.* **2002**, *21*, 2437–2444. [CrossRef]
4. Simkus, A.; Coolen-Maturi, T.; Coolen, F.P.A.; Bendtsen, C. Statistical perspectives on reproducibility: Definitions and challenges. *J. Stat. Theory Pract.* **2025**, *19*, 40. [CrossRef]
5. Coolen-Maturi, T.; Elkhafifi, F.F.; Coolen, F.P.A. Three-group ROC analysis: A nonparametric predictive approach. *Comput. Stat. Data Anal.* **2014**, *78*, 69–81. [CrossRef]
6. Baker, R.M.; Coolen-Maturi, T.; Coolen, F.P.A. Nonparametric predictive inference for stock returns. *J. Appl. Stat.* **2017**, *44*, 1333–1349. [CrossRef]
7. Coolen, F.P.A.; Coolen-Maturi, T. Nonparametric predictive inference. In *International Encyclopedia of Statistical Science*, 2nd ed.; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2025.
8. Coolen, F.P.A.; BinHimd, S. Nonparametric predictive inference for reproducibility of basic nonparametric tests. *J. Stat. Theory Pract.* **2014**, *8*, 591–618. [CrossRef]
9. Coolen, F.P.A.; BinHimd, S. Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov–Smirnov test. *J. Stat. Theory Pract.* **2020**, *14*, 26. [CrossRef]
10. Coolen, F.P.A.; Alqifari, H.N. Robustness of nonparametric predictive inference for future order statistics. *J. Stat. Theory Pract.* **2019**, *13*, 12. [CrossRef]
11. Coolen, F.P.A.; Alqifari, H.N. Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT Stat. J.* **2018**, *16*, 167–185.

12. Simkus, A.; Coolen, F.P.A.; Coolen-Maturi, T.; Karp, N.A.; Bendtsen, C. Statistical reproducibility for pairwise *t*-tests in pharmaceutical research. *Stat. Methods Med. Res.* **2022**, *31*, 673–688. [CrossRef]
13. Alshahrani, N.D. The Impact of Preliminary Tests on Statistical Reproducibility of Location Tests. Doctoral Dissertation, Durham University, Durham, UK, 2024.
14. Augustin, T.; Coolen, F.P.A. Nonparametric predictive inference and interval probability. *J. Stat. Plan. Inference* **2004**, *124*, 251–272. [CrossRef]
15. Hill, B.M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J. Am. Stat. Assoc.* **1968**, *63*, 677–691. [CrossRef]
16. Coolen-Maturi, T.; Coolen, F.P.A.; Alqifari, H. Nonparametric predictive inference for future order statistics. *Commun. Stat.-Theory Methods* **2018**, *47*, 2527–2548. [CrossRef]
17. Coolen, F.P.A. On nonparametric predictive inference and objective Bayesianism. *J. Log. Lang. Inf.* **2006**, *15*, 21–47. [CrossRef]
18. Lawless, J.F.; Fredette, M. Frequentist prediction intervals and predictive distributions. *Biometrika* **2005**, *92*, 529–542. [CrossRef]
19. Coolen, F.P.A.; Marques, F.J. Nonparametric predictive inference for test reproducibility by sampling future data orderings. *J. Stat. Theory Pract.* **2020**, *14*, 62. [CrossRef]
20. BinHimd, S. Nonparametric Predictive Methods for Bootstrap and Test Reproducibility. Ph.D. Thesis, Durham University, Durham, UK, 2014.
21. Tukey, J.W. Comparing individual means in the analysis of variance. *Biometrics* **1949**, *5*, 99–114. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.