AIMS Mathematics, 11(1): 957–976.
DOI: 10.3934/math.2026042
Received: 27 October 2025
Revised: 26 December 2025
Accepted: 08 January 2026
Published: 13 January 2026

https://www.aimspress.com/journal/Math

Research article

Statistical reproducibility of correlation tests: Pearson, Spearman, and Kendall

Norah D. Alshahrani*

Department of Mathematics, College of Science, University of Bisha, P.O. Box 551, Bisha 61922, Bisha, Saudi Arabia

* **Correspondence:** Email: ndmuflih@ub.edu.sa.

Abstract: Reproducibility has become a fundamental concern in modern statistical practice, yet its quantitative assessment remains limited for commonly used dependence measures. This study introduces a systematic evaluation of the reproducibility probability (RP), defined as the probability that the same statistical decision would be reached if an experiment were independently replicated under identical conditions. RP was examined for three widely used correlation tests (Pearson, Spearman, and Kendall) across different types of relationships and sample conditions. Through Monte Carlo simulations, RP was shown to provide a meaningful quantitative measure of the stability of statistical decisions across repeated experiments. Results indicated that the underlying relationship between variables, sample size, and noise level influenced reproducibility. In linear relationships, RP increased with both the strength of the true correlation and the sample size. For example, under strong linear dependence ($\rho = 0.9$), RP exceeded 0.95 for $n = 40$ and approached 1.00 for $n = 80$. For weak or null correlations ($\rho = 0$ or $\rho = 0.3$), the tests typically yielded non-significant p-values, and the corresponding RP values were generally above 0.5, reflecting stable decisions in the nonrejection area. The Pearson test demonstrated slightly higher RP in small samples due to its sensitivity to linear dependence, whereas rank-based methods achieved comparable reproducibility as the sample size increased. In contrast, under nonlinear nonmonotonic and piecewise monotonic relationships, reproducibility depended on both sample size and noise intensity. For small samples, all tests displayed highly variable RP values, while for larger samples or higher noise levels, RP values converged across methods. The results emphasized the role of RP as a reliable indicator of correlation test stability and revealed how underlying dependence patterns influenced the reproducibility of statistical results.

Keywords: reproducibility probability (RP); Pearson correlation test; Spearman's rank correlation test; Kendall's tau test
Mathematics Subject Classification: 62F03, 62F05, 62G10, 62G35, 62H20

1. Introduction

Correlation analysis is a fundamental tool in statistics, used to quantify the strength and direction of the relationship between two variables. Among the most widely used methods are the Pearson, Spearman, and Kendall correlation tests. Specifically, the Pearson correlation coefficient is used to determine the linear association between two continuous variables, predicated on the assumption of data normality. Conversely, the Spearman correlation, which relies on a rank transformation of the data, evaluates monotonic relationships and demonstrates greater resilience to deviations from normality and the presence of extreme values. Kendall's tau, similarly a rank-based metric, determines the degree of agreement and disagreement within paired data points, offering a completely nonparametric approach to identify monotonic dependencies.

Despite extensive research that has examined the significance testing of correlations, their reproducibility has received far less attention. Reproducibility probability (RP) quantifies the probability that a statistical decision, whether to reject or not reject the null hypothesis (H_0), would be replicated in an independent repetition of the same experiment. RP therefore provides a direct measure of the stability and dependability of statistical conclusions under repeated sampling.

In recent years, reproducibility in statistical hypothesis testing has gained increased attention as a cornerstone of scientific credibility. Reproducible results demonstrate that statistical conclusions are not mere artifacts of random variation or analytic choices. The National Academies of Sciences, Engineering, and Medicine [1] distinguish reproducibility (obtaining the same results using the same data and methods)from replicability, which concerns consistent findings under newly collected data. RP, by contrast, evaluates the expected stability of a hypothesis-testing decision when new data are collected under the same design.

Goodman [2] was among the first to examine the reproducibility of statistical findings, noting that p-values do not convey the likelihood that a significant result would reoccur. Senn [3] further clarified the conceptual distinction between the p-value and reproducibility probability, emphasizing their fundamentally different roles in statistical inference.

The broader reproducibility crisis in science has highlighted the instability of statistical conclusions under repeated experimentation. Ioannidis [4] showed that many published results fail to replicate, while Gelman and Carlin [5] characterized the prevalence of Type S and Type M errors that arise near decision boundaries. Benjamin et al. [6] proposed lowering the conventional significance threshold to improve replicability, and McShane and Gal [7] argued for moving beyond dichotomous decisions given their inherent instability.

Recent methodological developments have advanced reproducibility assessment. Hedges and Schauer [8] developed statistical methods and optimal design principles for ensembles of replication studies, linking replication analysis to power and study design. Power-oriented reproducibility approaches investigate how hypothesis decisions depend on study power and design characteristics. Simkus and collaborators [9] use predictive, nonparametric methods to quantify statistical reproducibility while incorporating uncertainty about future data. Atmanspacher and Maasen [10] highlighted conceptual challenges that arise when linking reproducibility to statistical evidence.

Applications in engineering further underscore the need to evaluate the stability of correlation-based inference. Zhang et al. [11, 12] demonstrated how linear, and nonlinear dependencies naturally arise in multi-component systems under s-dependent competing risks, where assessing reproducibility is

practically important.

Within this landscape, RP offers a decision-focused, post-study measure that quantifies the expected consistency of hypothesis-testing conclusions under repeated sampling. RP, therefore, complements power analysis, predictive inference, and replication design by directly assessing the stability of statistical decisions.

In this study, RP is defined as the conditional probability that an independent replication of the experiment, using the same design and analysis, yields the same hypothesis test decision to reject or not reject H_0 as the original study, given the observed data. Thus, RP quantifies the post-study stability of a binary decision under repeated sampling.

RP differs from several related concepts in the reproducibility literature. Predictive power is a pre-study quantity based on assumed effect sizes rather than realized data; false positive risk concerns the probability a significant result is false; and test-retest reliability evaluates measurement consistency rather than inferential decisions. RP is closely related to Goodman's replication probability [2], but it is extended by explicitly assessing agreement across both rejection and nonrejection decisions. The proposed framework focuses on estimating RP for Pearson, Spearman, and Kendall tests under diverse association structures, thereby providing targeted insight into the stability of correlation-based inference.

Table 1 provides a conceptual comparison of RP with these related metrics, highlighting their primary object, conditioning, target quantity, and typical applications. This distinction underscores RP's unique role as a decision-focused tool for post-hoc assessment of statistical stability, complementing pre-study planning (e.g., power analysis) and evidential recalibration (e.g., false positive risk).

Table 1. Conceptual comparison of reproducibility metrics.

Metric	Primary Object	Conditioning	Target Quantity	Typical Use
RP (this study)	Binary test decision	Observed data	Probability of same decision on replication	Assessing stability of correlation test outcomes
Predictive Power	Test statistic	Assumed effect and design	Prob. of rejecting H_0	Power analysis, study planning
False Positive Risk	Truth of H_0	p-value + priors	Prob. significant result is false	Reinterpreting evidence from p-values
Goodman's Replication Prob.	Significance outcome	Observed statistic	Prob. of another significant result	Evaluating reproducibility of significance
Test-Retest Reliability	Measurement scores	Repeated measurements	Consistency index (ICC)	Instrument reliability

Despite these valuable contributions, the reproducibility of statistical inference methods, particularly correlation tests, remains inadequately investigated. Correlation analysis is fundamental across numerous empirical sciences; however, there is limited understanding of how frequently correlation test outcomes would be replicated across repeated sampling, or how critical factors such as sample size, underlying dependence structures, and noise influence this stability.

RP is particularly suited to correlation testing because Pearson, Spearman, and Kendall statistics are highly sensitive to sample size, noise levels, and the form of the underlying association. Standard p-values do not reflect how stable a decision would be under repeated sampling, whereas RP directly quantifies the probability of decision concordance.

It is important to note that RP is fundamentally linked to classical properties of hypothesis testing. RP depends on the test's power, type I error rate, and the sampling distribution of the test statistic. When power is low (such as in small samples or under weak effect sizes) the decision boundary is highly sensitive to sampling variability, which reduces RP. In contrast, when power is high, rejection decisions become more stable and RP approaches one. Under the null hypothesis, RP tends to be high because the test statistic usually lies far from the rejection region, leading to consistent non-rejection decisions. Thus, RP can be viewed as a post-study measure that reflects how features of the sampling distribution translate into the expected stability of hypothesis-testing decisions.

This paper aims to evaluate RP of Pearson, Spearman, and Kendall correlation tests across various functional relationships, sample sizes, and noise levels. This provides insight into when correlation test decisions are most and least likely to replicate, offering practical guidance for applied researchers.

This paper is organized as follows. Section 2 outlines the overall methodology adopted in this study. It begins with a formal definition of reproducibility probability (RP) in subsection 2.1 and its interpretation within the framework of hypothesis testing. Subsection 2.2 then describes the resampling-based same-decision approach used to estimate decision stability. Next, subsection 2.3 reviews the Pearson, Spearman, and Kendall correlation tests, emphasizing their theoretical assumptions and properties. Subsection 2.4 details the data-generating mechanisms used to investigate the behavior of RP under varying conditions. Section 3 reports the analysis of the simulation outcomes, summarizing the RP patterns by conditions. Section 4 provides a detailed discussion interpreting these patterns in light of the theoretical properties of the correlation tests. Section 5 applies the RP framework to a real dataset to illustrate its practical use in empirical research. The study's key findings and implications are presented in Section 6.

2. Methodology

2.1. Reproducibility probability (RP)

Reproducibility probability (RP) is defined as the probability that a statistical test yields the same conclusion when the experiment is repeated independently under identical conditions [2].

RP can be formulated either *unconditionally*, based on population-level quantities, or *conditionally*, based on the observed data.

Unconditional RP refers to the probability that two independent replications lead to the same decision under the true data-generating mechanism. If π denotes the test's rejection probability (i.e., its power under the true distribution), then the probability that both replications produce the same

decision is

$$P(\text{same decision}) = \pi^2 + (1 - \pi)^2,$$

a classical result in hypothesis testing [13].

This study focuses on the *conditional* reproducibility probability:

$$RP = P(D^* = D_0 \mid \text{data}),$$

which quantifies the stability of the observed hypothesis test decision given the specific dataset. Conditional RP is directly estimatable via bootstrap resampling and is therefore the relevant measure for applied settings.

High RP values indicate that the observed decision is likely to be reproduced in independent replications, whereas low RP values signal instability and increased susceptibility to sampling variability.

2.2. Bootstrap estimation of reproducibility probability

To estimate RP empirically, a nonparametric plug-in approach based on the paired bootstrap was adopted. In this framework, the unknown joint distribution $F_{X,Y}$ is approximated by the empirical distribution $\widehat{F}_{X,Y}$, and RP (which is defined in terms of new independent samples from $F_{X,Y}$) is estimated by resampling from $\widehat{F}_{X,Y}$. Because the correlation tests considered in this study rely on paired observations (x_i, y_i), the resampling scheme follows the standard pairs-bootstrap procedure [14]. Each bootstrap sample is generated by drawing n pairs (x_i, y_i) with replacement from the observed data, thereby preserving the joint dependence structure between X and Y. This approach parallels case-based bootstrapping in regression models, where the empirical distribution of the observed pairs serves as a nonparametric estimator of the unknown joint distribution $F_{X,Y}$.

Given an observed dataset $\{(x_i, y_i)\}_{i=1}^n$, from the joint distribution of (X, Y), the estimation procedure proceeded as follows:

(1) Compute the test statistic and corresponding p-value from the original data using a chosen correlation method (Pearson, Spearman, or Kendall).
(2) Record the initial decision $D_0 = \mathbb{I}_{(p<\alpha)}$, where α is the chosen significance level; this means $D_0 = 0$ when the null hypothesis is not rejected and $D_0 = 1$ when it is rejected.
(3) Draw B bootstrap samples of size n by resampling pairs (x_i, y_i) with replacement (pairwise bootstrap).
(4) Recalculate the p-value, denoted p_b where $b = 1, \ldots, B$, and determine $D_b = \mathbb{I}_{(p_b<\alpha)}$.
(5) Estimate RP as the proportion of bootstrap decisions that match the original decision:

$$\widehat{RP} = \frac{\sum_{b=1}^{B} \mathbb{I}(D_b = D_0)}{B}. \tag{2.1}$$

Although RP is defined in terms of new independent samples from the true distribution $F_{X,Y}$, the bootstrap estimator replaces $F_{X,Y}$ by the empirical distribution $\widehat{F}_{X,Y}$. Under classical nonparametric bootstrap theory [15, 16], resampling from $\widehat{F}_{X,Y}$ consistently approximates sampling from $F_{X,Y}$ when

(1) the sample size n is sufficiently large,
(2) the dependence within pairs (X, Y) is preserved,

(3) the statistic is a sufficiently smooth functional of $F_{X,Y}$.

Correlation statistics satisfied these conditions under mild assumptions, making the paired bootstrap a theoretically justified method for approximating the probability of decision agreement under repeated sampling.

The bootstrap estimator

$$\widehat{RP} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(D_b = D_0)$$

is unbiased for the bootstrap analogue of RP, RP^*, conditional on the observed data, and has conditional Monte Carlo variance:

$$\mathrm{Var}(\widehat{RP} \mid \widehat{F}_{X,Y}) = \frac{RP^*(1 - RP^*)}{B},$$

where RP^* denotes the reproducibility probability under $\widehat{F}_{X,Y}$. Bias arises because $\widehat{F}_{X,Y}$ replaces $F_{X,Y}$, but this plug-in bias diminishes as $n \to \infty$. Thus, under the usual bootstrap regularity conditions, the paired-bootstrap estimator is consistent for decision reproducibility.

Bootstrap-based RP estimation depended entirely on the observed sample. When n is small, $\widehat{F}_{X,Y}$ may poorly approximate $F_{X,Y}$, producing unstable or biased RP estimates. Pearson correlation is particularly sensitive to outliers and non-Gaussian distributions, whereas rank-based tests (Spearman, Kendall) are more robust but not immune to finite-sample effects. Consequently, bootstrap RP estimates depend strongly on the original data and should be interpreted as conditional, data-dependent approximations rather than exact replications of sampling from the true distribution.

2.3. Correlation tests considered

Three widely used tests for association between two continuous variables were considered: Pearson's product-moment correlation [17], Spearman's rank correlation [18], and Kendall's tau [19]. All are applied to paired observations (x_i, y_i), $i = 1, \ldots, n$, to test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ using their usual large-sample null distributions.

Pearson's test is based on the sample correlation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

which, under H_0 and standard regularity conditions, yields the t-statistic:

$$T_P = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}.$$

Spearman's test applies the same construction to the marginal ranks $R(x_i)$ and $R(y_i)$, producing a correlation coefficient ρ_S and corresponding t-statistic:

$$T_S = \frac{\rho_S\sqrt{n-2}}{\sqrt{1-\rho_S^2}},$$

which is approximately t_{n-2} under H_0 for continuous distributions without ties [18]. Discrete data with ties require permutation tests or tie corrections.

Kendall's

$$\tau_a = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

standardizes to

$$Z_K = \frac{\tau_a}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \overset{approx}{\sim} \mathcal{N}(0,1)$$

for large n without ties [19]. With ties, τ_b and tie-adjusted variance are required.

Table 2 summarizes assumptions, robustness, computational complexity, expected RP behavior, and optimal use cases for Pearson, Spearman, and Kendall tests.

Table 2. Comparison of Pearson, Spearman, and Kendall correlation tests.

	Pearson	Spearman	Kendall
Assumptions	Linear relationship; normality assumptions for inference.	Monotonic relationship (not necessarily linear).	Monotonic concordance; pairwise comparison based.
Robustness	Sensitive to outliers and nonlinear patterns; affected by skewness and heavy tails.	Robust to outliers; invariant to monotonic transformations; performance degrades when global monotonicity is violated.	Most robust to noise and outliers; less sensitive to extreme values than Spearman.
Computational Complexity	$O(n)$	$O(n \log n)$	$O(n^2)$
Expected RP Behavior	High RP under strong linear dependence; RP decreases under nonlinear or non-monotonic relationships.	High RP when the association is approximately monotonic; RP decreases when monotonicity is violated or holds only locally (e.g., sinusoidal or piecewise patterns).	RP broadly similar to Spearman but often slightly lower in small samples due to reduced efficiency.
Best For	Strictly linear relationships.	Linear or approximately monotonic relationships; local monotonic segments of nonlinear associations.	Noisy, small-sample, or tied datasets where robustness is critical.

2.4. Simulation design

To examine RP under controlled and interpretable conditions, a simulation framework was designed to span a broad range of dependence structures, noise levels, and sample sizes commonly encountered in empirical studies. The goal is to evaluate how RP behaves when the underlying association varies from linear to strongly nonlinear, from low to high noise, and when sample sizes range from very small to moderately large.

Four functional relationships were selected to represent linear and nonlinear associations, with a focus on globally non-monotonic and piecewise monotonic dependence structures. These models are widely used as benchmark functions in the dependence-measure literature, and have been employed in simulation studies evaluating nonlinear association statistics such as MIC [20], distance correlation [21], and other modern dependence tests [22, 23].

- Linear: $Y = \rho X + \sqrt{1-\rho^2}\,\varepsilon$, where $X, \varepsilon \sim \mathcal{N}(0,1)$. This model provides a baseline scenario where Pearson correlation is optimal and reflects the most common assumption in applied correlation analyses.
- Quadratic: $Y = X^2 + \varepsilon$ with $X \sim \mathcal{U}(-2,2)$. This nonlinear, symmetric, non-monotonic pattern (U-shaped) is widely used in simulation benchmarks for dependence measures [20, 21] and appears in biomarker dose–response curves and environmental risk models.
- Sinusoidal: $Y = \sin(X) + \varepsilon$ with $X \sim \mathcal{U}(-2\pi, 2\pi)$. Sinusoidal relationships are classic examples of periodic associations and are routinely used to evaluate methods for detecting nonlinear dependence [20, 22].
- Absolute value (piecewise): $Y = |X| + \varepsilon$, with $X \sim \mathcal{U}(-2,2)$. This piecewise, magnitude-driven pattern is another standard benchmark for evaluating nonlinear and non-monotonic dependence [20, 23], reflecting threshold-like or regime-switching behaviour.

These models spanned qualitatively different association structures (strictly linear, nonlinear nonmonotonic, and piecewise) which allows for assessment of the robustness of Pearson, Spearman, and Kendall tests across diverse settings.

The values $\rho \in \{0, 0.3, 0.6, 0.9\}$ were chosen to reflect widely used benchmarks for weak, moderate, and strong associations, following the conventional effect size classifications discussed by Cohen [24]. This range enabled investigation of RP across low-, moderate-, and high-power regimes.

Noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$ represented low, medium, and high noise conditions. This helps to see how stable the results are when measurement errors increase. These noise levels reflected real-world measurement errors often found in biomedical, psychological, and environmental studies, and enabled the investigation of how increasing noise lowers the reproducibility probability across different dependence structures.

Sample sizes $n \in 10, 40, 80$ were selected to represent small, moderate, and larger finite-sample regimes typically discussed in the statistical literature. Very small samples such as $n = 10$ are characteristic of exploratory or pilot studies, where as sample sizes around 10–12 are commonly recommended [25]. Medium sample sizes in the range of 30–50 are frequently used in simulation studies evaluating the behavior of correlation estimators, including Pearson, Spearman, and Kendall [26]. Larger samples, such as $n = 80$, fall within the range where classical large-sample theory for rank-based correlations begins to provide accurate approximations [27]. Considering these

sample sizes allowed us to examine how RP behaves across underpowered, moderately powered, and large-powered regimes

Each parameter configuration was evaluated using $M = 100$ Monte Carlo replications and $B = 1000$ bootstrap iterations per replication. These values were chosen based on preliminary convergence checks: Increasing M and B further produced negligible changes in RP estimates, indicating sufficient numerical stability. These choices are consistent with simulation-based reproducibility studies and ensure that Monte Carlo variability remains small relative to the observed patterns.

For each simulated dataset, the original p-value and its corresponding RP estimate were computed for the Pearson, Spearman, and Kendall correlation tests. This pairing of p-values with their RP values enabled direct comparison of decision stability across linear and nonlinear settings, noise levels, and sample sizes.

Table 3 summarizes all components of the simulation framework, including the selected values, their methodological purpose, and their empirical relevance.

Table 3. Summary of simulation components, selected values, and their methodological and empirical motivations.

Component	Choices	Purpose / Expected Properties	Empirical Relevance
Dependence function	Linear, quadratic, sinusoidal, piecewise $\mid X \mid$	Represent linear, nonlinear non-monotonic associations; assess how curvature and regime changes affect RP	Biomarker dose-response (quadratic), periodic/seasonal effects (sinusoidal), threshold or regime-switching behavior (piecewise)
Correlation level ρ	$0, 0.3, 0.6, 0.9$	Span no, weak, moderate, and strong dependence; cover low- and high-power regimes for correlation tests	Typical effect sizes reported in applied correlation studies; standard benchmarks in simulation work
Noise level σ_e	$0.2, 0.5, 1, 2$	Vary signal-to-noise ratio from low to high noise; study robustness of RP to measurement error and unobserved variability	Reflects measurement variability in psychological, biomedical, and observational data sets
Sample size n	$10, 40, 80$	Represent small, moderate, and larger finite samples; examine finite-sample behavior and approach to asymptotics	Pilot or exploratory studies ($n \approx 10$), medium-size experiments ($n \approx 40$), and more typical studies ($n \approx 80$)

3. Analysis and visualization

For each test and data scenario, the estimated $\widehat{RP}$ values were plotted against their corresponding p-values. The resulting RP-p plots revealed how the stability of a test decision varies across different sample sizes. In these plots, different colours represent each correlation test (Pearson, Spearman, or Kendall). The vertical dashed line represents the significance threshold $\alpha = 0.05$, which separates the rejection region ($p < \alpha$) from the non-rejection region of the null hypothesis. The horizontal dotted line at $RP = 0.5$ serves as a reference for the stability of statistical decisions. Points above this line indicated relatively stable decisions (i.e., the same conclusion is likely to be reached in repeated samples), whereas points near or below this level suggest higher variability and lower reproducibility of the test outcome. RP was plotted against the original p-value for each simulation (dots), with LOESS smoothing curves (lines) added to highlight the general trend.

Figure 1 shows the reproducibility probability (RP) values for Pearson, Spearman, and Kendall correlation tests under different linear correlation strengths $\rho \in \{0, 0.3, 0.6, 0.9\}$, using $X \sim \mathcal{N}(0, 1)$ and $Y = \rho X + \sqrt{1-\rho^2}\,\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$, for different sample sizes 10, 40 and 80.

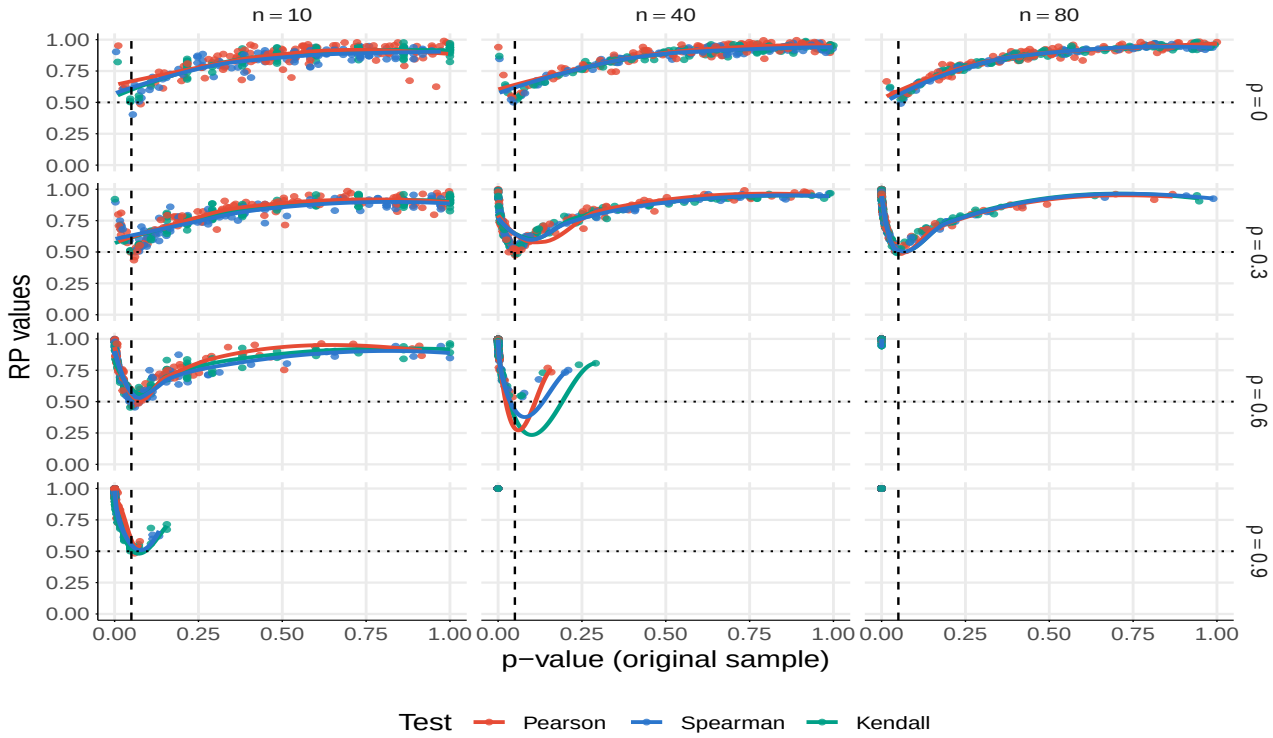

Figure 1. Relationship between reproducibility probability (RP) and p-value for the Pearson, Spearman, and Kendall correlation tests under different linear correlation strengths $\rho \in \{0, 0.3, 0.6, 0.9\}$, using $X \sim \mathcal{N}(0, 1)$ and $Y = \rho X + \sqrt{1-\rho^2}\,\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$, at a significance level of $\alpha = 0.05$.

Across all scenarios, the RP values exhibited a consistent pattern: RP tended to decrease as the p-values approached the significance threshold α and increased when the p-values were farther from this boundary, this pattern noticed in many studies such as [28–31].

When the true correlation was weak ($\rho = 0$ or 0.3), most p-values fell in the non-rejection region, and the corresponding RP values were generally above 0.5. These RP values also increased with sample size, indicating more stable non-rejection decisions for larger n. Pearson tended to show slightly higher RP values than Spearman and Kendall in the non-rejection region for $n = 10$ and $n = 40$, although this difference diminished as the sample size increased. As the true correlation became strong, the tests frequently produced very small p-values, especially for moderate and large sample sizes. In these settings, RP values were close to 1, reflecting consistently significant decisions in repeated samples. For $n = 10$ when $\rho = 0.9$, and $n = 40$ when $\rho = 0.6$, Pearson tended to show slightly lower RP values compared to Spearman and Kendall in the non-rejection region, but slightlly higher RP values within the rejection region.

RP variability was highest when the p-values were close to the significance level α, particularly in small samples. As n increased, the RP values became smoother and more similar in all three correlation tests.

Figure 2 presents the RP values for the Pearson, Spearman, and Kendall correlation tests under a nonmonotonic quadratic relationship where $X \sim \mathcal{U}(-2, 2)$ and $Y = X^2 + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$. Results are shown for different sample sizes $n \in \{10, 40, 80\}$ and noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$.

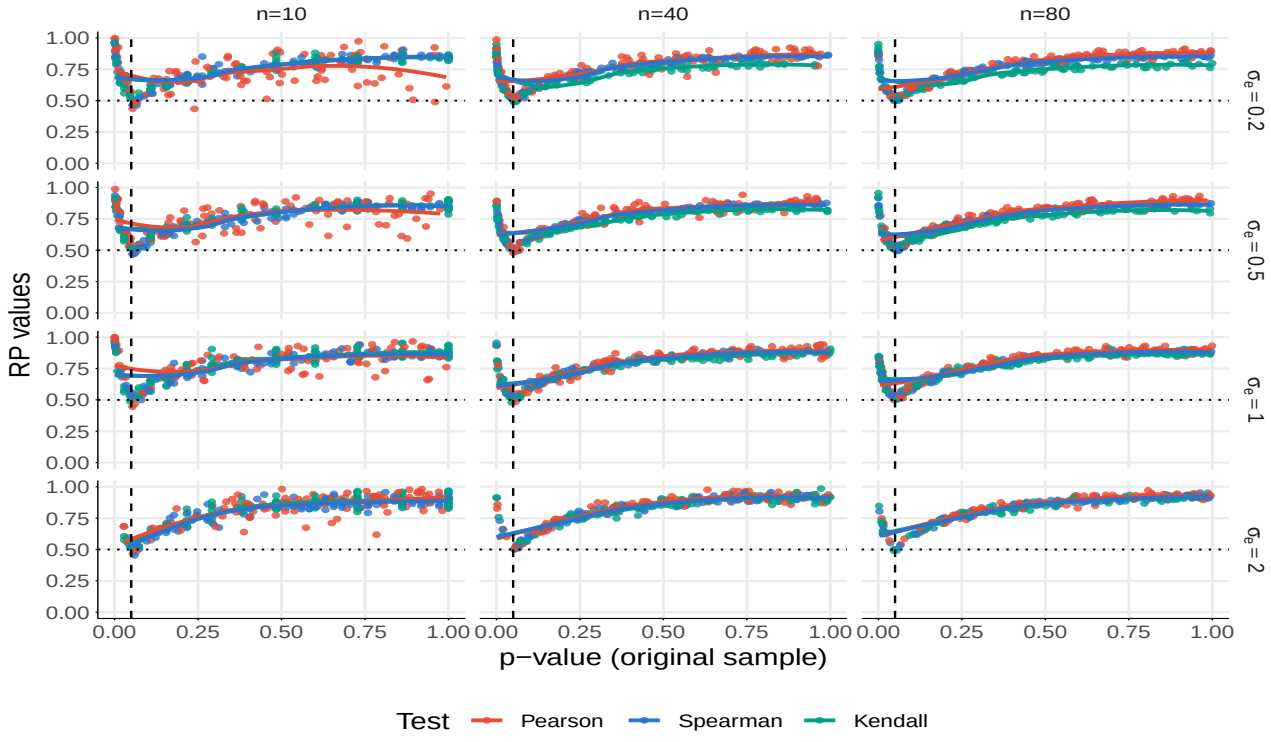

Figure 2. Relationship between reproducibility probability (RP) and p-value for Pearson, Spearman, and Kendall under a non-monotonic quadratic model with $X \sim \mathcal{U}(-2, 2)$ and $Y = X^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$. Results are shown for sample sizes $n \in \{10, 40, 80\}$ and noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$ at $\alpha = 0.05$.

The reproducibility patterns depended strongly on both sample size and noise level. For small sample sizes, all three tests yield nearly similar RP values with high variability, particularly for the

Pearson test, which showed unstable performance due to its sensitivity to nonlinearity. As the sample size increased and noise remained low, the reproducibility probability (RP) for the Pearson test becomes slightly higher, followed by Spearman and then Kendall tests. When both the sample size and noise level increase, the RP values across all tests converge and become approximately similar, suggesting that under higher noise and large samples, the differences among the correlation measures diminish and all tests exhibit comparable reproducibility.

Figure 3 presents RP values for the Pearson, Spearman, and Kendall correlation tests under a nonmonotonic sinusoidal relationship, where $X \sim \mathcal{U}(-2\pi, 2\pi)$ and $Y = \sin(X) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$. The results are shown for different sample sizes $n \in \{10, 40, 80\}$ and noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$. Across all correlation methods, RP values were approximately similar, with most of the corresponding p-values located in the non-rejection region. For small samples or high noise levels, all three methods yield predominantly nonsignificant results, and RP values higher than 0.5 in the non-rejection area. As the sample size increased and the noise level decreased, a larger proportion of p-values appeared in the rejection region across all three methods. This produced consistently high RP values in the rejection region, while the occasional p-values that remained near the significance threshold exhibited lower RP.

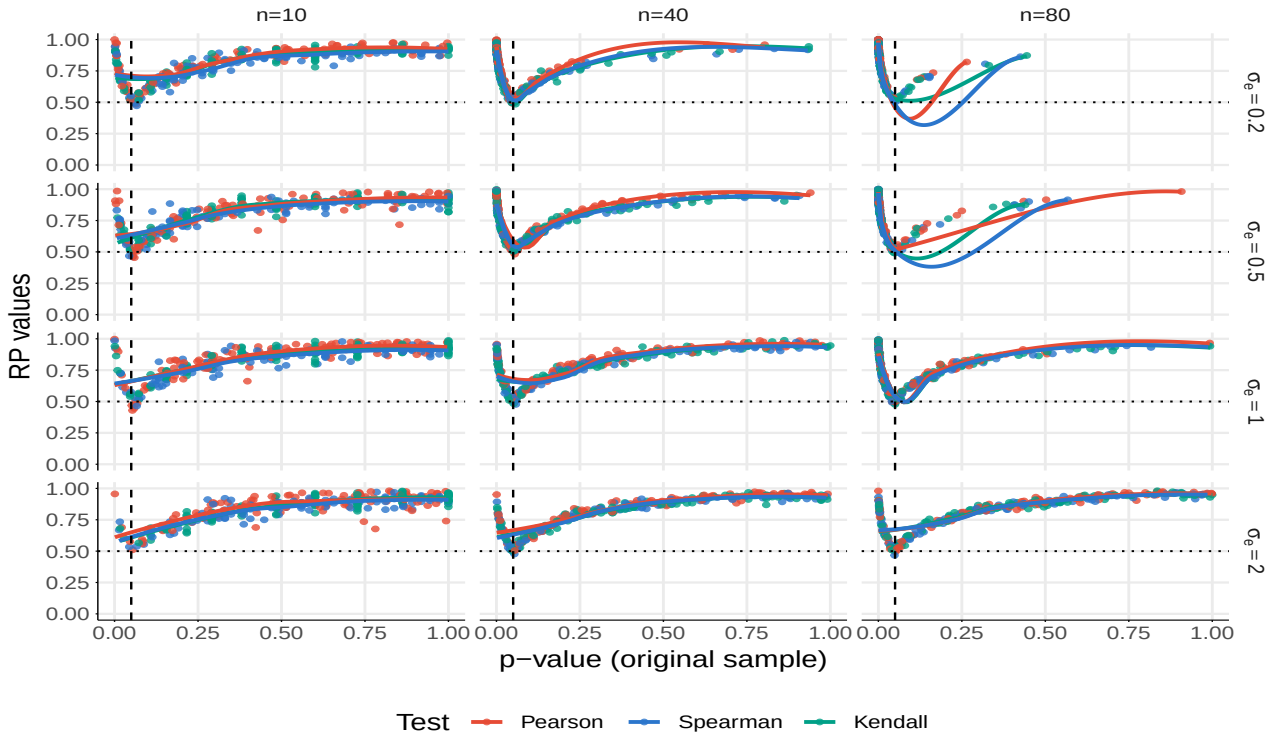

Figure 3. Relationship between reproducibility probability (RP) and p-value for the Pearson, Spearman, and Kendall correlation tests under a nonmonotonic nonlinear model, with $X \sim \mathcal{U}(-2\pi, 2\pi)$ and $Y = \sin(X) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$, at $\alpha = 0.05$.

Figure 4 shows RP values for Pearson, Spearman, and Kendall correlation tests under the piecewise model, $Y = |X| + \varepsilon$, where $X \sim \mathcal{U}(-2, 2)$ and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$. The results are shown for different sample sizes $n \in \{10, 40, 80\}$ and different levels of noise strengths $\sigma_e \in \{0.2, 0.5, 1, 2\}$. Most of the p-values

remained in the non-rejection region across all tests, sample sizes, and noise levels. At low noise and moderate and high sample sizes, Pearson displayed slightly higher RP values, while Spearman and Kendall occasionally yield smaller p-values. Under high noise, the RP values became nearly identical in all three correlation tests.

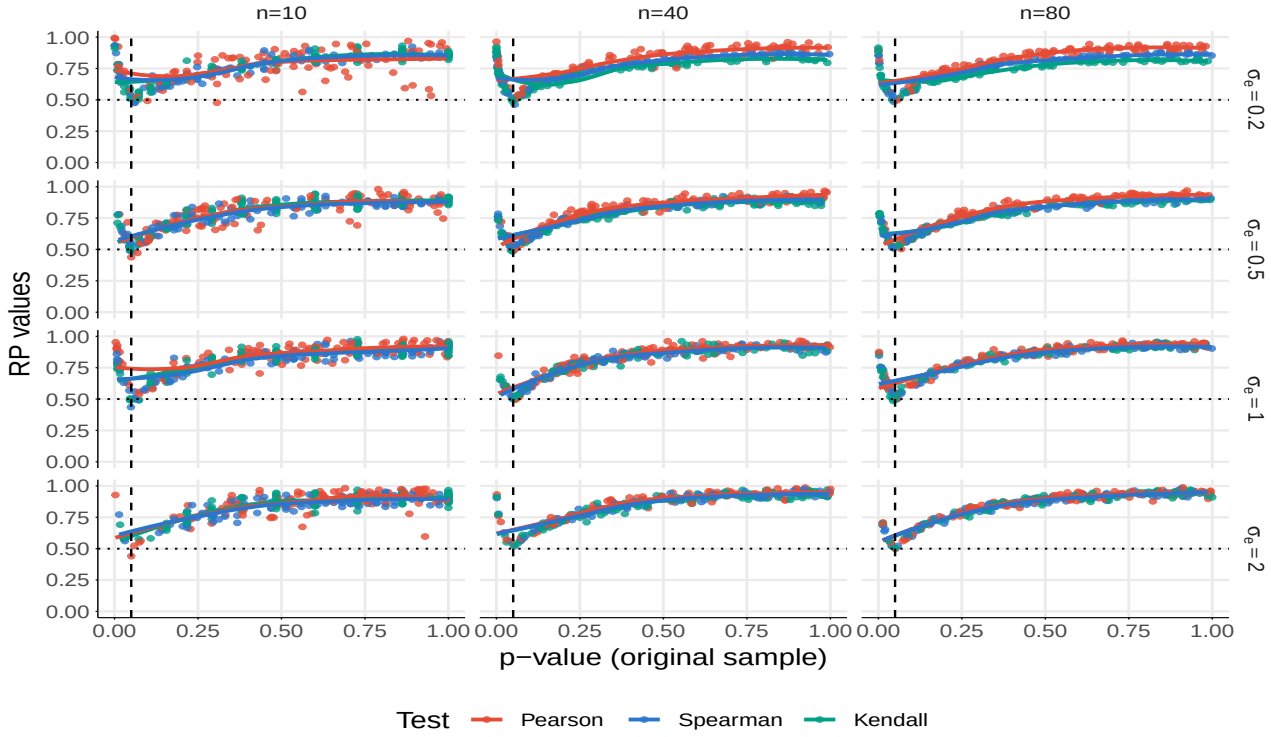

Figure 4. Relationship between reproducibility probability (RP) and p-value for the Pearson, Spearman, and Kendall correlation tests under the piecewise model, $Y = |X| + \varepsilon$, where $X \sim \mathcal{U}(-2, 2)$ and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$. Results are shown for sample sizes $n \in \{10, 40, 80\}$ and noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$ at a significance level of $\alpha = 0.05$.

4. Discussion

The reproducibility patterns observed in the simulations reflect fundamental properties of correlation test statistics and their sampling distributions. RP is lowest when the observed statistic lies near its critical threshold, where small perturbations from repeated sampling can easily change the decision, and increases as the statistic moves farther from the rejection boundary. Importantly, RP depends on the sampling distribution of the test statistic rather than on the p-value itself. Low RP values often occur near $p = \alpha$ because the corresponding statistic is close to the critical region, but the nonlinear mapping between the statistic and its p-value, as well as the asymmetry of finite-sample distributions, implies that the minimum RP does not necessarily align exactly with $p = \alpha$; in several panels of Figures 1 and 3, the empirical RP minimum occurs slightly to the left or right of the nominal threshold.

Under linear dependence, Pearson correlation aligns closely with the linear data-generating mechanism and therefore tends to yield more concentrated sampling distributions and higher RP than

Spearman and Kendall, particularly in small to moderate samples and under strong linear signals. In very small samples, however, a few nonrejection outcomes for Pearson can exhibit noticeably lower RP than the rank-based tests because the Pearson statistic is more sensitive to sampling fluctuations, whereas Spearman and Kendall, being rank-based, produce slightly more stable decisions even when they occasionally fail to detect the strong linear association. As n increases, the sampling distributions of all three statistics tighten around their expectations, leading to convergence of RP values across tests. High RP for large $|\rho|$ reflects high power and consistently significant decisions, while high RP for weak ρ in the non-rejection region reflects consistently non-significant outcomes far from the rejection boundary.

For nonlinear designs, the quadratic, sinusoidal, and $|x|$ models highlight how nonmonotonic or piecewise-monotonic structure interacts with the tests' assumptions. In the quadratic model, curvature induces moderate linear association, so Pearson often attains slightly higher RP at moderate and large n, while Spearman and Kendall show lower RP because global monotonicity is violated. In the sinusoidal design, repeated oscillations make the global correlation approximately zero, so all methods produce mostly non-significant p-values and uniformly high RP for non-rejection; occasional detections of local linear segments at larger n and low noise lead to pockets of high RP in the rejection region. For the $Y = |X| + \varepsilon$ model, symmetry implies zero population correlation for all three measures; any apparent nonzero correlation in low-noise, moderate-n settings arises from sampling within one branch, and increasing noise quickly obscures both curvature and rank orderings, driving all methods toward high non-rejection RP. Overall, the simulations reveal clear interaction effects: Larger n reduces sampling variability and stabilizes RP, higher noise obscures structure and pushes all methods toward uniformly high non-rejection RP, and alignment between the true functional form and the test's assumptions determines which test attains higher RP in low-noise regimes.

4.1. Connection between conditional RP and power

Although reproducibility probability (RP) is not itself a power measure, conditional RP is indirectly linked to classical notions of power through the sampling distribution of the test statistic. Conditional RP,

$$RP_{\text{cond}} = P(D^* = D_0 \mid \text{observed data}),$$

where D^* denotes the decision in a hypothetical replication under the same design, depends on how far the realized test statistic lies from the rejection boundary relative to its sampling variability. Under standard regularity conditions (continuous distributions, smooth test statistics, and large-sample approximations), this variability follows the same sampling distribution that defines the test's power. However, unlike power, conditional RP conditions on the realized test statistic rather than on the underlying model parameters. When the observed statistic lies far from the critical value, small perturbations due to resampling are unlikely to change the decision, leading to high RP. In contrast, when the observed statistic is close to the rejection boundary, even minor sampling fluctuations can change the decision, resulting in low RP.

This mechanism explains the empirical patterns observed in the simulations: RP increases with sample size and effect magnitude because both reduce relative sampling variability, while RP attains its minimum near the decision boundary. Importantly, this relationship is post-study and data-dependent. Unlike power, which is a pre-study, population-level quantity, conditional RP generally does not admit a closed-form expression as a function of (n, ρ, σ) and therefore cannot be used directly for sample

size determination. Instead, RP complements power by quantifying the stability of a specific realized decision under repeated sampling with the same design.

4.2. Practical interpretation and use of conditional RP

The reproducibility probability considered in this study is a *conditional*, *post-study* measure intended to complement, rather than replace, conventional hypothesis testing. In practice, RP should be reported alongside the test statistic and p-value when the stability of a dichotomous decision (reject vs. not reject H_0) is of interest. RP is particularly informative when the observed p-value lies close to the significance threshold: A low RP then indicates that the decision is sensitive to sampling variability and may change under repeated sampling, whereas a high RP in either the rejection or non-rejection region indicates a decision likely to be reproduced under the same design.

When RP and the p-value convey seemingly conflicting information, RP should be interpreted as a measure of decision stability rather than evidential strength. For example, a small p-value accompanied by low RP suggests an unstable rejection decision near the critical boundary, whereas a moderate p-value with high RP indicates a consistently reproducible non-rejection outcome. In this framework, RP plays a diagnostic and descriptive role: It quantifies how sensitive a given test decision is to repeated sampling under the same design. Translating RP into formal design rules (such as determining the sample size required to achieve a target RP level) would require an *unconditional*, population-level formulation of RP linked explicitly to power, and such extensions are therefore beyond the scope of the present study.

4.3. Limitations and future directions

The simulation framework in this study focused on smooth, continuous data-generating mechanisms commonly used as benchmarks in the dependence-measurement literature, allowing systematic assessment of how RP responds to changes in sample size, noise, and functional form. Several practically important scenarios were not included. Heteroscedastic noise, heavy-tailed distributions, outliers, and discrete data that induce ties can substantially affect the sampling distributions of correlation statistics: Heteroscedasticity and asymmetric noise increase sampling variability and tend to lower RP; heavy tails reduce the stability of Pearson statistics; and ties alter the null distributions of Spearman and Kendall and require tie-adjusted inference. Another key assumption in the simulations was independence of observations; clustering, temporal autocorrelation, or spatial dependence can alter both test statistics and their variability, often reducing RP. Extending RP analysis to these settings represents an important direction for future work.

The case $\rho = 0$ in nonlinear models was intentionally included as a baseline scenario. Although the global correlation is zero, the underlying functional dependence may be strong, illustrating that RP reflects reproducibility of the *test decision*, not the presence or strength of structural dependence. When tests have low power against nonlinear alternatives, RP may remain high simply because non-rejection is consistently reproduced. A further limitation is that RP was reported only through point estimates. Because RP is itself a Monte Carlo estimator, it has sampling variability that depends on the number of bootstrap replications B and on the proximity of the test statistic to the rejection boundary. While confidence intervals for RP can be computed (e.g., via a secondary bootstrap or binomial standard errors), they are computationally intensive for large simulation grids; interval estimation is therefore

left for future work.

More broadly, reproducibility assessment in data analysis extends beyond correlation tests. Correlation coefficients are only one component of exploratory analysis. Graphical tools such as scatterplots, smoothers, and residual diagnostics are crucial for identifying nonlinear or non-monotonic structure that may affect reproducibility, and modern dependence measures (e.g., distance correlation or mutual-information–based statistics) can complement RP when assessing more complex associations. Incorporating RP alongside these exploratory and inferential tools can provide a more comprehensive understanding of the stability of scientific conclusions.

5. Real data application

To illustrate the practical use of RP in an empirical setting, the relationship between `petal.length` and `petal.width` from the classical *Iris* dataset was analyzed. These variables exhibited a strong but nonlinear and species-dependent relationship, making them suitable for comparing the performance of Pearson, Spearman, and Kendall correlation tests. For each test, the correlation estimate, the corresponding p-value, a 95% confidence interval (parametric or bootstrap), and RP based on $B = 1000$ bootstrap samples were computed.

Table 4 reports the correlation estimates, confidence intervals, the corresponding p-value, and RP values for Pearson, Spearman, and Kendall methods. All three tests detect a very strong positive association, with estimates of 0.96 (Pearson), 0.94 (Spearman), and 0.81 (Kendall), and extremely small p-values ($< 10^{-40}$). The corresponding bootstrap confidence intervals are narrow for all three measures, indicating a highly stable estimation across resamples. RP values range from 0.88 to 0.95, showing that the decision to reject H_0 would be reproduced in more than 88% of repeated samples of the same size.

Table 4. Correlation estimates, confidence intervals, and reproducibility probabilities for the *Iris* dataset

Method	Estimate	p-value	95% CI	RP
Pearson	0.963	4.68×10^{-86}	[0.952, 0.973]	0.95
Spearman	0.938	8.16×10^{-70}	[0.916, 0.952]	0.92
Kendall	0.807	2.44×10^{-44}	[0.773, 0.839]	0.88

To complement the real-data analysis, an empirical simulation study based on the *Iris* dataset was conducted. Rather than assuming a parametric model, the original sample of size 150 was treated as an empirical population, and subsamples of sizes $n \in 10, 40, 80$ were repeatedly drawn. For each subsample, the correlation estimate, its corresponding p-value, and the conditional RP were computed using the paired bootstrap. This procedure evaluates how the stability of the correlation-test decision would change if the same study were carried out with smaller or larger samples drawn from the same underlying relationship. In a second step, artificial noise of varying magnitudes $\sigma_e \in \{0.2, 0.5, 1, 2\}$ was added to the `petal.width` variable to examine how measurement error affects reproducibility. For each combination of (n, σ_e), 100 resampled datasets were generated, the three correlation tests were computed, and their conditional RP was estimated via the paired bootstrap. This design allowed us to study how reproducibility degrades when noise obscures the underlying biological signal.

Figure 5 illustrates how sample size and noise jointly influenced RP in this real-data setting. Small subsamples exhibited greater variability and lower RP values near the significance boundary, whereas larger subsamples yielded highly stable decisions with RP values close to one. Nearly all p-values fell far below the significance threshold $\alpha = 0.05$, which explains the consistently high RP values: For this dataset, the decision to reject H_0 would be reproduced in the vast majority of repeated samples. This reproducibility was particularly strong when the noise level was low and the sample size was moderate or large.

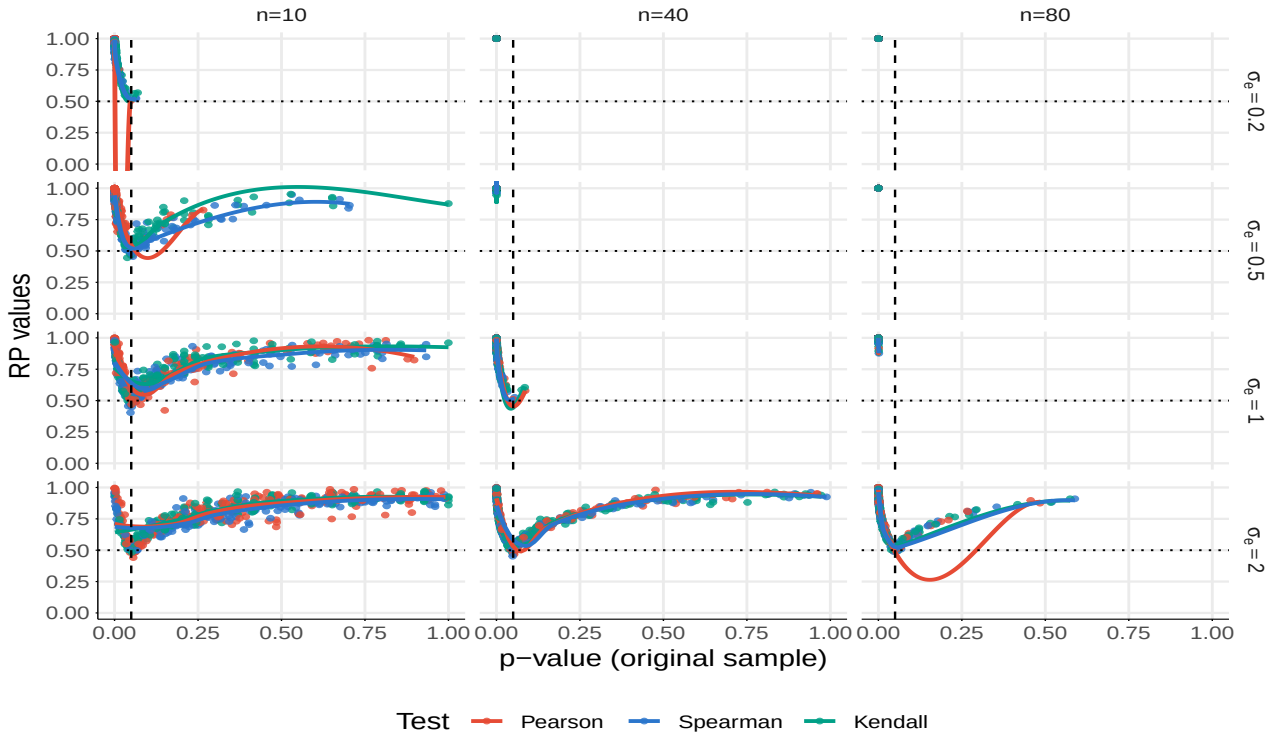

Figure 5. Relationship between reproducibility probability (RP) and p-value for the Pearson, Spearman, and Kendall correlation tests for the *Iris* dataset (`Petal.Length` vs. `Petal.Width`). Results are shown for sample sizes $n \in \{10, 40, 80\}$ and noise levels $\sigma_e \in \{0.2, 0.5, 1, 2\}$ at a significance level of $\alpha = 0.05$.

Overall, the *Iris* example demonstrated how RP complements standard correlation analysis. Although the effect size was strong and the p-values were extremely small, statistical significance alone does not indicate how stable the conclusion would be under repeated sampling. The RP values, which ranged from 0.88 to 0.95, confirmed that the detected association was not only statistically significant but also highly reproducible in studies of similar design.

6. Conclusions

This study examined the reproducibility probability (RP) of three commonly used correlation tests, Pearson, Spearman, and Kendall, under a variety of data generating mechanisms and sample conditions. Through simulations, RP was shown to provide a meaningful quantitative measure of

decision stability across repeated experiments. The results showed that reproducibility is strongly influenced by the underlying relationship between variables, the presence of noise, and sample size.

In linear relationships, RP increased with both the strength of the true correlation and the sample size. When the underlying relationship was strongly linear, all correlation tests, Pearson, Spearman, and Kendall, showed highly reproducible decisions, with RP values approaching one. The Pearson test exhibited slightly higher RP in smaller samples, reflecting its greater sensitivity to linear dependence, while rank-based methods became comparable as the sample size increased. These results confirmed that, under linear associations, reproducibility primarily depended on the magnitude of the correlation and the available data rather than on the choice of the correlation test.

While in the nonlinear, non-monotonic, and piecewise settings, reproducibility was affected by both sample size and noise level. For small samples, all correlation tests exhibited similar yet highly variable RP values, indicating unstable decision reproducibility. As the sample size increased and noise remained low, the Pearson test tended to yield slightly higher RP values due to its sensitivity to curved dependencies, whereas the Spearman and Kendall tests, being rank-based, displayed greater robustness to nonlinearity but comparatively lower RP in such conditions. With increasing noise and larger sample sizes, the RP values across all tests converged, suggesting that differences between linear and rank-based correlations became negligible under higher variability or data-rich environments.

The real-data analysis supported the simulation findings. Smaller subsamples exhibited greater variability and less stable reproducibility across the three correlation tests, whereas larger samples produced more consistent and reliable outcomes. As the sample size increased, the RP values for Pearson, Spearman, and Kendall became nearly identical, indicating that under practical data conditions, reproducibility improves with sample size and the differences among the correlation measures diminish. This demonstrates that RP provides a useful assessment of decision stability beyond conventional p-values, both in simulated and real-world settings.

RP provides applied researchers with a valuable tool for assessing the stability of statistical decisions beyond traditional probability values. By determining the probability of a statistically significant observational outcome being repeated under identical conditions, RP helps distinguish robust results from those highly sensitive to sample variance. Practitioners can use RP to prioritize findings for further verification, guide study design choices such as sample size, and complement effect size and power analyses to enhance scientific reliability. Integrating RP into standard analytical workflows promotes transparent, evidence-based decision-making and reinforces reproducible research practices.

Use of Generative-AI tools declaration

AI-assisted language tools (ChatGPT) were used for language polishing and clarity improvement.

Acknowledgments

The author is thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

Conflict of interest

The author declares no conflicts of interest in this paper.

References

1. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and replicability in science*, Washington: The National Academies Press, 2019.
2. S. N. Goodman, A comment on replication, p-values and evidence, *Stat. Med.*, **11** (1992), 875–879. https://doi.org/10.1002/sim.4780110705
3. S. Senn, A comment on replication p-values and evidence, S. N. Goodman, Statistics in Medicine, 1992; 11: 875–879, *Statist. Med.*, **21** (2002), 2437–2444. https://doi.org/10.1002/sim.1072
4. J. P. A. Ioannidis, Why most published research findings are false, *PLoS Med.*, **2** (2005), e124. https://doi.org/10.1371/journal.pmed.1004085
5. A. Gelman, J. Carlin, Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors, *Perspect. Psychol. Sci.*, **9** (2014), 641–651. https://doi.org/10.1177/1745691614551642
6. D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, et al., Redefine statistical significance, *Nat. Hum. Behav.*, **2** (2018), 6–10. https://doi.org/10.1038/s41562-017-0189-z
7. B. B. McShane, D. Gal, Statistical significance and the dichotomization of evidence, *J. Am. Stat. Assoc.*, **112** (2017), 885–895. https://doi.org/10.1080/01621459.2017.1289846
8. L. V. Hedges, J. M. Schauer, The design of replication studies, *J. R. Statist. Soc. Ser. A*, **184** (2021), 868–886. https://doi.org/10.1111/rssa.12688
9. A. Simkus, T. Coolen-Maturi, F. P. A. Coolen, C. Bendtsen, Statistical perspectives on reproducibility: Definitions and challenges, *J. Stat. Theory Pract.*, **19** (2025), 40. https://doi.org/10.1007/s42519-025-00459-x
10. H. Atmanspacher, S. Maasen, *Reproducibility: Principles, problems, practices, and prospects*, John Wiley & Sons, 2016. https://doi.org/10.1002/9781118865064
11. L. Zhang, X. Chen, A. Khatab, Y. An, Optimizing imperfect preventive maintenance in multi-component repairable systems under s-dependent competing risks, *Reliab. Eng. Syst. Safe.*, **219** (2022), 108177. https://doi.org/10.1016/j.ress.2021.108177
12. L. Zhang, X. Chen, A. Khatab, Y. An, X. Feng, Joint optimization of selective maintenance and repairpersons assignment problem for mission-oriented systems operating under s-dependent competing risks, *Reliab. Eng. Syst. Safe.*, **242** (2024), 109796. https://doi.org/10.1016/j.ress.2023.109796
13. E. L. Lehmann, J. Romano, *Testing statistical hypotheses*, New York: Springer, 2005. https://doi.org/10.1007/0-387-27605-X
14. J. Fox, *Applied regression analysis and generalized linear models*, Sage publications, 2015.
15. B. Efron, R. Tibshirani, *An introduction to the bootstrap*, New York: Chapman and Hall/CRC, 1994. https://doi.org/10.1201/9780429246593

16. A. C. Davison, D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, 1997.
17. K. Pearson, Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia, *Philos. Trans. A Math. Phys. Eng. Sci.*, **187** (1986), 253–318. https://doi.org/10.1098/rsta.1896.0007
18. C. Spearman, The proof and measurement of association between two things, *Am. J. Psychol.*, **15** (1904), 72–101. https://doi.org/10.2307/1412159
19. M. G. Kendall, A new measure of rank correlation, *Biometrika*, **30** (1938), 81–93. https://doi.org/10.2307/2332226
20. D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, et al., Detecting novel associations in large data sets, *Science*, **334** (2011), 1518–1524. https://doi.org/10.1126/science.1205438
21. G. J. Székely, M. L. Rizzo, N. K. Bakirov, Measuring and testing dependence by correlation of distances, *Ann. Statist.*, **35** (2007), 2769–2794. https://doi.org/10.1214/009053607000000505
22. R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, *J. Am. Stat. Assoc.*, **107** (2012), 1129–1139. https://doi.org/10.1080/01621459.2012.695654
23. R. Heller, Y. Heller, M. Gorfine, A consistent multivariate test of association based on ranks of distances, *Biometrika*, **100** (2013), 503–510. https://doi.org/10.1093/biomet/ass070
24. J. Cohen, *Statistical power analysis for the behavioral sciences*, New York: Routledge, 1988. https://doi.org/10.4324/9780203771587
25. S. A. Julious, Sample size of 12 per group rule of thumb for a pilot study, *Pharm. Stat.*, **4** (2005), 287–291. https://doi.org/10.1002/pst.185
26. D. G. Bonett, T. A. Wright, Sample size requirements for estimating pearson, kendall and spearman correlations, *Psychometrika*, **65** (2000), 23–28. https://doi.org/10.1007/BF02294183
27. M. G. Kendall, *Rank correlation methods*, Griffin, 1948.
28. N. D. Alshahrani, T. Coolen-Maturi, F. P. A. Coolen, On statistical reproducibility of normality and equality of variances tests, *J. Stat. Theory Pract.*, **19** (2025), 81. https://doi.org/10.1007/s42519-025-00495-7
29. F. P. A. Coolen, S. BinHimd, Nonparametric predictive inference for reproducibility of basic nonparametric tests, *J. Stat. Theory Pract.*, **8** (2014), 591–618. https://doi.org/10.1080/15598608.2013.819792
30. A. Simkus, F. P. A. Coolen, T. Coolen-Maturi, N. A. Karp, C. Bendtsen, Statistical reproducibility for pairwise t-tests in pharmaceutical research, *Stat. Methods Med. Res.*, **31** (2021), 673–688. https://doi.org/10.1177/09622802211041765
31. A. Aldawsari, T. Coolen-Maturi, F. P. A. Coolen, Parametric predictive bootstrap method for the reproducibility of hypothesis tests, *J. Stat. Theory Pract.*, **19** (2025), 21. https://doi.org/10.1007/s42519-025-00438-2

AIMS Press