Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2025, 13(4), 440-466
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

MVN Q-Test II: A Comparison of the MVN H-Test with the Chi-Square Approximation and Bootstrap Versions of the Q-Test

José Moral de la Rubia

School of Psychology, Universidad Autónoma de Nuevo León, Monterrey, Mexico
Email: jose.morald@uanl.edu.mx

How to cite this paper: Moral de la Rubia, J. (2025) MVN Q-Test II: A Comparison of the MVN H-Test with the Chi-Square Approximation and Bootstrap Versions of the Q-Test. *Journal of Data Analysis and Information Processing*, **13**, 440-466.
https://doi.org/10.4236/jdaip.2025.134026

Received: August 2, 2025
Accepted: September 23, 2025
Published: September 26, 2025

Copyright © 2025 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

In a previous article, an R script was developed and divided into three parts to implement the multivariate normality (MVN) Q-test based on both the chi-square approximation and the bootstrap approach, using either the Shapiro-Wilk W statistic (QSWa and QSWb) or the Shapiro-Francia W' statistic (QSFa and QSFb). Royston's H-test was included as a supplementary MVN test. The aim of this study is to compare the hit rate and statistical power of the four Q-test variants and the H-test using 200 samples drawn from multivariate standard normal distributions and 200 samples from multivariate t-distributions with five degrees of freedom. The simulations vary in sample size (50, 75, 100, 125, 150, 200, 250, and 500), number of variables (from 2 to 6), and homogeneous inter-variable correlation (0, 0.3, 0.5, 0.7, and 0.9). The H-test outperformed QSWb and QSFb, but not QSWa in the multivariate normal samples or QSFa in the multivariate t-distribution samples. QSFb performed better than QSWb. It is concluded that the bootstrap approach is conservative under the null hypothesis of multivariate normality. However, when the assumption of independence is violated, the bootstrap approach is theoretically more appropriate than QSWa and QSFa. A 10% significance level is recommended for QSFb in terms of hit rate, but in terms of statistical power, only when rejecting the null hypothesis.

Keywords

Multivariate Normality, Parametric Bootstrap, Non-Parametric Bootstrap, Bootstrap P-Value, Bootstrap Power

1. Introduction

A previous article aimed to develop an R script for implementing the multivariate

normality Q-test [1]. Based on the original formulation, a version of the Q-test using a chi-square approximation was implemented [2]. Additionally, a bootstrap version was included, following the recommendations of the original study. Both versions of the Q-test incorporate two variants: one based on Royston's standardization [3] of Shapiro-Wilk univariate normality statistic [4], and the other based on Royston's standardization [5] of Shapiro-Francia statistic [6]. These four variants are referred to as QSWa, QSFa, QSWb, and QSFb, respectively.

The univariate normality tests proposed by Shapiro and Wilk [4] and by Shapiro and Francia [6] are based on the correlation between empirical and theoretical quantiles under the null hypothesis of normality. The former employs a more complex procedure for computing the theoretical quantiles, which in its original version [4] restricts its application to samples of 3 to 50 observations, and to samples of 3 to 2000 with Royston's standardization [5]. The latter uses a simpler method for obtaining the theoretical quantiles, resulting in fewer limitations [6]; with Royston's standardization [7], it can be applied to samples ranging from 5 to 5000 observations.

The objective of this study is to compare the hit rate and mean power across the four variants of the Q-test [1] [2] and the Royston's H-test [7]. A total of 200 random samples were generated from multivariate normal (MVN) distributions and another 200 from multivariate t (MVT) distributions with five degrees of freedom. The simulations varied in sample size (50, 75, 100, 125, 150, 200, 250, and 500), number of variables (2, 3, 4, 5, and 6), and the homogeneous inter-variable correlation (0, 0.3, 0.5, 0.7, and 0.9).

Distribution type and these three factors are known determinants of statistical power in multivariate tests [8]. For simplicity, the study focused on two distribution types: one under which the null hypothesis of multivariate normality should be retained (MVN), and one under which it should be rejected (MVT).

The R program was selected because it is one of the most comprehensive statistical software packages, open access, and continuously developed and reviewed by the mathematical community [9] [10]. Royston's H-test [7] was included because it is one of the most widely used and powerful tests of multivariate normality [11] [12]. Moreover, in the original study [2], the Q-test was found to outperform Mardia's K^2-test [13] in terms of hit rate and mean power, and its performance closely approximated that of Royston's H-test [7].

Additionally, to explore the possibility of harmonizing decisions based on *p*-values obtained from the chi-square approximation of the Q-test with those derived from its bootstrap counterpart, hit rates and mean power were compared at significance levels of 0.05 and 0.10 using the bootstrap version of the Q-test. The bootstrap method produces a flatter sampling distribution with heavier tails than the standard chi-square distribution, as it corresponds to a generalized chi-square distribution—typical of quadratic forms involving correlated Gaussian vectors. Accordingly, a 5% significance level may be used for the chi-square-based Q-test variant, while a 10% level is preferable for the bootstrap-based variant [1].

2. Method

A total of 400 samples were generated. Specifically, 200 samples of sizes 50, 75, 100, 125, 150, 200, 250, and 500 were created for each of 2, 3, 4, 5, and 6 variables, with homogeneous correlations of 0, 0.3, 0.5, 0.7, and 0.9. These samples were drawn from a multivariate t-distribution with five degrees of freedom (MVTD). Another 200 samples with the same specifications were drawn from a multivariate standard normal distribution (MVND). All samples were analyzed using the four Q-test variants (SWa, SWb, SFa, and SFb) [1] [2] and Royston's H-test [7].

The multivariate t-distribution with five degrees of freedom was chosen as the primary alternative to normality because its pronounced leptokurtosis, resulting from heavy tails, represents a common and substantial deviation from the multivariate normal distribution. For simplicity, the study focused on only alternative for multivariate non-normality.

For the generation of these 400 samples, the R package mvtnorm [14] was used, with a fixed seed (123) to ensure reproducibility. **Appendix** presents the script used to generate two samples, each consisting of 50 observations. Both samples contain two variables with a correlation coefficient of 0.3. One sample was drawn from a multivariate t-distribution with five degrees of freedom, while the other was drawn from a multivariate standard normal distribution.

Since the samples were generated using the same seed, the hit rate and mean power values for the five MVN tests constitute repeated measurements. Therefore, statistical tests for repeated measures were applied. In these analyses, the hit rate is defined as the proportion of correct decisions made based on the p-value, using a significance level of 5%. For samples drawn from multivariate normal distributions, the correct decision is to retain the null hypothesis, whereas for samples from multivariate t-distributions, the correct decision is to reject the null hypothesis.

Confidence intervals (CI) for the hit rates were calculated using Wilson's method [15], as implemented in the R DescTools package [16]. Differences in hit rates among the five MVN tests were assessed using Cochran's Q test [17]. Effect sizes were estimated using the eta-squared coefficient, with values of approximately 0.01, 0.06, and 0.14 conventionally interpreted as small, medium, and large effects, respectively. In practical terms, larger eta-squared values indicate a greater proportion of the variance in hit rates attributable to the type of MVN test, thus reflecting stronger differences in performance among the tests [18].

Pairwise comparisons of hit rates were conducted using paired-sample Z-tests with a shared estimated standard error [19]. To control the family-wise error rate [20], the Bonferroni's correction was applied [21]. Effect sizes for these comparisons were calculated using a Cohen's d analogue, defined as $d = z/\sqrt{n}$ [19]. The resulting d values were interpreted according to Cohen's (1988) thresholds: [0, 0.2) = trivial, [0.2, 0.5) = small, [0.5, 0.8) = medium, and ≥ 0.8 = large [22]. All omnibus and pairwise comparisons were performed using SPSS, version 27 [23]. These analyses were conducted on both the pooled sample and the individual sam-

ples based on distribution type (MVT and MVN).

To assess the effect of the significance level (0.05 versus 0.10) on the hit rate when applying the bootstrap variant of the Q-test—based on either the Shapiro-Wilk W or the Shapiro-Francia W' statistics—Rosenthal's r effect size was used, calculated from the z-statistic of the McNemar test ($r = |z|/\sqrt{n}$).

In the pooled sample, mean power—specifically, power derived from the MVT distributions and its complement from the MVN distributions—was compared using repeated measures analysis of variance (ANOVA). Sample size, number of variables, and homogeneous inter-variable correlation were included as covariates in the general linear model (GLM), while distribution type (MVT vs. MVN) was treated as a between-subjects factor.

Sphericity—*i.e.*, the homogeneity of variances of the differences between all possible pairs of repeated measures—was evaluated using Mauchly's test [24]. Due to a substantial violation of this assumption ($\varepsilon < 0.7$), the Huynh-Feldt epsilon correction [25] was applied to the analysis of within-subject effects. Additionally, Pillai's trace [26] was employed for multivariate testing, as recommended for repeated measures ANOVA [27].

Effect sizes for each model component were estimated using partial eta-squared. Pairwise comparisons were conducted via paired-sample Student's t-tests with Bonferroni's correction [21]. Corresponding effect sizes were computed using Cohen's d [22].

Bootstrap confidence intervals for the mean statistical powers of the five MVN tests (repeated measures), as well as their correlations with the three covariates (in the GLM), were computed using the Bias-Corrected and accelerated (BCa) percentile method based on 1000 resamples [28]. These correlations represent Pearson product-moment coefficients [29].

Given that the assumption of multivariate normality for the repeated-measure components—as assessed by Royston's H-test [7]—and the assumption of univariate normality for the residuals—as tested by the Shapiro-Wilk W-test [3] [4]—were both violated, Friedman's test [30] was additionally applied. Pairwise comparisons were conducted using Conover's test [31]. Effect size for the omnibus test was estimated via Kendall's W coefficient [32], with thresholds of approximately 0.1, 0.3, and 0.5 interpreted as small, moderate, and large effects, respectively. Effect sizes for pairwise comparisons were calculated using the rank-biserial correlation [33], where values around 0.1, 0.3, and 0.5 similarly represent small, medium, and large effects. In practical terms, higher coefficients for either measure indicate stronger associations between the type of MVN test and the observed differences in statistical power, reflecting more pronounced performance gaps [22].

Mean power was compared once again for the 200 samples drawn from the multivariate t-distribution with five degrees of freedom and the 200 samples from the multivariate standard normal distribution, focusing specifically on power values. These mean comparisons were conducted using the JASP software, version

0.19.2 [34].

To assess the effect of significance level (0.05 versus 0.10) on the mean power of the bootstrap variant of the Q-test, based on either the Shapiro-Wilk W or the Shapiro-Francia W' statistics, a paired t-test was used. The effect size was measured using the Hedges-Olkin g statistic, and BCa confidence intervals were computed for the mean difference due to the violation of the normality assumption.

3. Results

3.1. Hit Rate Comparison

3.1.1. Differences among Five MVN Test and Their Pairwise Comparisons

Based on 400 samples drawn from two distributions (MVT and MVN), Royston's H-test achieved the highest hit rate (*HR* = 0.95; 95% Wilson-type *CI*: 0.924, 0.967). Its confidence interval overlapped with those of two Q-test variants based on the chi-square approximation: QSFa (*HR* = 0.93; 95% Wilson-type *CI*: 0.901, 0.951) and QSWa (*HR* = 0.90; 95% Wilson-type *CI*: 0.867, 0.926). See Table 1 and Figure 1 for additional details.

The QSWb test exhibited the lowest hit rate (*HR* = 0.84; 95% Wilson-type *CI*: 0.801, 0.873), with its confidence interval overlapping that of the QSFb test (*HR* = 0.88; 95% Wilson-type *CI*: 0.844, 0.908). However, the upper limit of the QSWb interval was below the lower bounds of both the QSFa test (*HR* = 0.93; 95% *CI*: 0.901, 0.951) and the H test (*HR* = 0.95; 95% *CI*: 0.924, 0.967). Likewise, the upper bound of the QSFb interval was lower than the lower bound of the H-test confidence interval. For further details, refer to Table 1 and Figure 1.

For the 200 samples drawn from the multivariate t-distribution, the results mirrored those obtained from the pooled sample. In contrast, the five confidence intervals corresponding to the 200 samples from the multivariate standard normal distribution were found to overlap and were significantly higher than those derived from the multivariate t-distribution. Refer to Table 1 and Figure 1 for further details.

Table 1. Hit rates and 95% Wilson-type confidence intervals of the five MVN tests.

MNV test	Pooled (*n* = 400)			MVT (*n* = 200)			MVN (*n* = 200)		
	HR	*LL*	*UL*	*HR*	*LL*	*UL*	*HR*	*LL*	*UL*
QSWa	0.90	0.867	0.926	0.825	0.766	0.871	0.975	0.943	0.989
QSWb	0.84	0.801	0.873	0.68	0.612	0.741	1	0.981	1
QSFa	0.93	0.901	0.951	0.87	0.816	0.910	1	0.981	1
QSFb	0.88	0.844	0.908	0.75	0.686	0.805	1	0.981	1
H	0.95	0.924	0.967	0.915	0.868	0.946	0.99	0.964	0.997

Note. MVN tests: QSWa = chi-square approximation Q-test variant using Shapiro-Wilk W statistics, QSWb = bootstrap Q-test variant using Shapiro-Wilk W statistics, QSFa = chi-square approximation Q-test variant using Shapiro-Francia W' statistics, QSFb = bootstrap Q-test variant using Shapiro-Francia W' statistics, H = Royston's multivariate normality H-test. Distribution: MVT = multivariate t-distribution with five degrees of freedom and MVN = multivariate standard normal distribution. Statistics: *n* = sample size, *HR* = hit rate, 95%Wilson-type confidence interval: *LL* = lower limit, *UL* = upper limit.

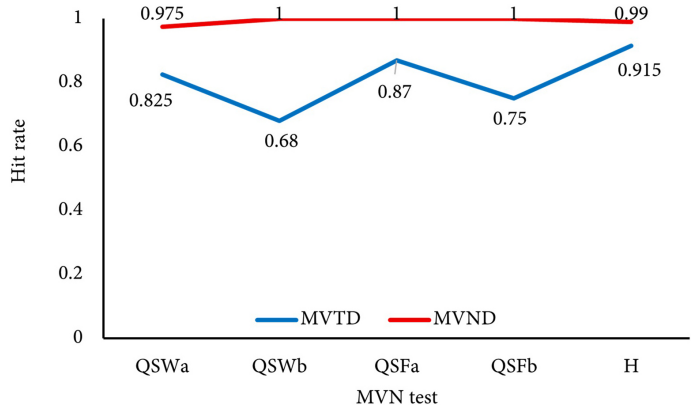

Figure 1. Stacked line chart showing the hit rates of five multivariate normality tests, with separate lines representing the distribution from which each random sample was drawn (MVTD = multivariate t-distribution with five degrees of freedom and MVND = multivariate standard normal distribution).

According to Cochran's omnibus test, there was a significant difference in hit rates among the five multivariate normality tests within the pooled sample (Q = 96.529, asymptotic $p < 0.001$, exact $p < 0.000$), with a medium effect size as indicated by the eta-squared coefficient ($0.06 < \eta^2 = 0.0603 < 0.14$). Of the 10 possible pairwise comparisons, 6 showed statistically significant differences in hit rates using the paired Z-test with Bonferroni's correction, all reflecting small effect sizes ($0.2 \leq d < 0.5$). No significant differences were found between the following test pairs: QSWa vs. QSFa ($p_{Bonf} = 0.073$), QSWa vs. QSFb ($p_{Bonf} = 0.552$), QSWb vs. QSFb ($p_{Bonf} = 0.073$), and QSFa vs. H ($p_{Bonf} = 1$). Refer to Table 2 for detailed results.

Table 2. Pairwise comparisons of hit rates in the pooled sample of 400 observations of 5-tuples.

Test 1 − Test 2	$HR_1 - HR_2$	se	z	p	p_{Bonf}	d	Int.
QSWa − QSWb	0.060	0.013	4.602	< 0.001	< 0.001	0.231	s
QSWa − QSFa	−0.035	0.013	−2.684	0.007	0.073	−0.135	t
QSWa − QSFb	0.025	0.013	1.917	0.055	0.552	0.096	t
QSWa − H	−0.053	0.013	−4.027	< 0.001	0.001	−0.204	s
QSWb − QSFa	−0.095	0.013	−7.286	< 0.001	< 0.001	−0.365	s
QSWb − QSFb	−0.035	0.013	−2.684	0.007	0.073	−0.135	t
QSWb − H	−0.113	0.013	−8.628	< 0.001	< 0.001	−0.435	s
QSFa − QSFb	0.06	0.013	4.602	< 0.001	< 0.001	0.231	s
QSFa − H	−0.018	0.013	−1.342	0.180	1	−0.069	t
QSFb − H	−0.078	0.013	−5.944	< 0.001	< 0.001	−0.300	s

Note. $HR_1 - HR_2$ = difference between hit rates, se = standard error, z = paired Z-test statistic, p = two-tailed asymptotic p-value, p_{Bonf} = Bonferroni-adjusted two-tailed p-value for an alpha level of 0.050, Cohen's $d = z/\sqrt{n}$ = effect size statistic, Int = interpretation of the effect size: [0, 0.2) trivial (t), [0.2, 0.5) small (s), [0.5, 0.8) medium (m), and ≥ 0.8 large (l).

When comparing the hit rates of the five multivariate normality tests based on the 200 samples drawn from the multivariate t-distribution, the differences were statistically significant (Q = 115.836; asymptotic $p < 0.001$; exact $p < 0.001$), with

a large effect size ($\eta^2 = 0.145 > 0.14$). Out of the 10 possible pairwise comparisons, 8 showed significant differences. No significant difference was observed between QSWa and QSFa, or between QSWa and H ($HR_1 - HR_2 = -0.045$; $z = -1.822$; $p = 0.068$; $p_{Bonf} = 0.685$; Cohen's $d = -0.127$ for both comparisons). The effect sizes were small in the statistically significant differences, except for three of them. The effect sizes were medium ($0.5 \leq$ Cohen's $d < 0.8$) in the difference between QSWb and H (Cohen's $d = -0.665$) and between QSWb and QSFa (Cohen's $d = -0.537$). In contrast, the effect size was trivial (Cohen's $d = -0.198 < 0.2$) in the difference between QSWb and QSFb.

The difference among the five MVN tests was not significant when comparing the hit rate of the 200 samples drawn from the multivariate standard normal distribution ($Q = 13.714$, asymptotic $p = 0.008$, exact $p = 0.007$; $\eta^2 = 0.017$).

3.1.2. Effect of Significance Level on Q-Test Hit Rate (Bootstrap Version)

In the pooled sample, when the significance level was increased to 0.10 to compensate for the marked right-tail asymmetry in the bootstrap sampling distribution of the Q-test statistic, the hit rate of QSWb was 0.85 (95% Wilson-type *CI*: 0.812, 0.882), which was not significantly different from that obtained at a sinificance level of 0.05 ($HR_{0.05} = 0.84$ vs. $HR_{0.10} = 0.85$; $d = -0.010$, 95% Wilson-type *CI*: −0.023, 0.002, two-tailed mid-p binomial p-value = 0.063 > 0.05).

When the significance level was increased to 0.10, the hit rate of QSFb rose to 0.8975 (95% Wilson-type *CI*: 0.864, 0.924), which was significantly higher than the value obtained at 0.05 ($HR_{0.05} = 0.875$ vs. $HR_{0.10} = 0.8975$; $d = -0.0225$, 95% Wilson-type *CI*: [−0.040, −0.007], two-tailed mid-p binomial p-value = 0.002 < 0.05). The effect size of the significance level on the hit rate was small (Rosenthal's $r = |z|/\sqrt{n} = 3/20 = 0.15$, based on McNemar's test).

The same result was obtained with the 200 samples drawn from multivariate t-distributions with five degrees of freedom. When using a significance level of 0.05 or 0.10, the QSWb test showed no significant difference in the hit rate ($HR_{0.05} = 0.68$ vs. $HR_{0.10} = 0.70$; d = −0.020, 95% Wilson-type *CI*: −0.043, 0.002, two-tailed mid-p binomial p-value = 0.063 > 0.05). However, when the significance level was increased to 0.10, the hit rate of QSFb significantly increased ($HR_{0.05} = 0.75$ vs. $HR_{0.10} = 0.795$; $d = -0.045$, 95% Wilson-type *CI*: −0.078, −0.014, two-tailed mid-p binomial p-value = 0.002 < 0.05). The effect size of the significance level on the hit rate was small (Rosenthal's $r = |z|/\sqrt{n} = 3/\sqrt{200} \approx 0.212$, based on McNemar's test).

Using a significance level of 0.05 or 0.10, no significant difference in the hit rate was observed for QSWb and QSFb when 200 samples were drawn from multivariate normal distributions. The hit rate in both cases was 1.

3.2. Mean Power Comparison among the Five MVN Tests in the Pooled Sample

3.2.1. Repeated Measures ANOVA

Using the 400 data points per MVN test obtained from samples drawn from multivariate t-distributions with five degrees of freedom (MVTD) and from multivar-

iate standard normal distributions (MVND)—pooled sample—the highest mean statistical power (for MVTD) or its complement (for MVND) was found in the QSFa test. Its 95% BCa confidence interval (0.849; 95% BCa *CI*: 0.827, 0.869) overlapped with those of the other tests, except for QSWa (0.800; 95% BCa *CI*: 0.772, 0.824). However, when the distributions were analyzed separately, numerous crossovers in mean power were observed. This indicates a clear interaction between the test type and the distribution type (Table 3).

Table 3. Mean power values, 95% BCa confidence intervals, and descriptive statistics.

Dist	Test	*n*	*m*	*LL*	*UL*	*sd*	*cv*
MVT	QSWa	200	0.917	0.882	0.943	0.015	0.235
MVN		200	0.685	0.64	0.716	0.021	0.438
Total		400	0.800	0.772	0.824	0.285	0.357
MVT	QSWb	200	0.758	0.719	0.79	0.021	0.394
MVN		200	0.857	0.836	0.876	0.01	0.158
Total		400	0.807	0.781	0.828	0.236	0.293
MVT	QSFa	200	0.939	0.91	0.96	0.013	0.189
MVN		200	0.760	0.725	0.786	0.017	0.31
Total		400	0.849	0.827	0.869	0.227	0.267
MVT	QSFb	200	0.806	0.77	0.838	0.019	0.342
MVN		200	0.868	0.846	0.889	0.01	0.16
Total		400	0.837	0.813	0.858	0.220	0.263
MVT	H	200	0.948	0.915	1.044	0.029	0.429
MVN		200	0.732	0.709	0.755	0.011	0.221
Total		400	0.841	0.820	0.884	0.328	0.390

Note. MVN tests: QSWa = chi-square approximation Q-test variant using Shapiro-Wilk W statistics, QSWb = bootstrap Q-test variant using Shapiro-Wilk W statistics, QSFa = chi-square approximation Q-test variant using Shapiro-Francia W' statistics, QSFb = bootstrap Q-test variant using Shapiro-Francia W' statistics, H = Royston's multivariate normality H-test. Statistics: *n* = sample size, *m* = arithmetic mean, 95% Bias-Corrected and accelerated (BCa) percentile confidence interval (CI): *LL* = lower limit, *UL* = upper limit, *sd* = standard error, and *cv* = Pearson's coefficient of variation.

When comparing the mean power (for MVTD) or its complement (for MVND)—pooled sample—among the five MVN tests applied to samples drawn from multivariate distributions, a repeated measures ANOVA was conducted using sample size (8 levels), number of variables (5 levels), and homogeneous correlation among variables (5 levels) as covariates, and distribution type (2 levels) as a between-subjects factor. The assumption of sphericity was violated (Mauchly's test: $W = 0.012$, χ^2 [9, $N = 400$] = 1749.684, $p < 0.001$). After applying the Huynh-Feldt correction ($\varepsilon = 0.483$) to the within-subjects effects, the mean difference was significant (F[1.933, 763.725] = 13.58, $p < 0.001$), with a small effect size ($0.01 < \eta_p^2 = 0.033 < 0.06$). Refer to Table 4 for further details. The effect of test type on statistical power was further supported by a multivariate analysis using Pillai's trace ($V = 0.097$, F[4, 392] = 10.583, $p < 0.001$), indicating a medium effect size ($0.06 < \eta_p^2 = 0.097 < 0.14$) and full statistical power ($\phi = 1$). See Table 5 for more details.

Table 4. Within-subjects effects using the Huynh-Feldt sphericity correction in the pooled sample of 400 observations of 5-tuples.

Effect	*SS*	*df*	*MS*	*F*	*p*	η_p^2
MVN_tests	1.728	1.933	0.894	13.58	< 0.001	0.033
MVN_tests × *n*	0.835	1.933	0.432	6.564	0.002	0.016
MVN_tests × *k*	2.835	1.933	1.466	22.286	< 0.001	0.053
MVN_tests × *r*	0.483	1.933	0.25	3.797	0.024	0.010
MVN_tests × Distribution	10.404	1.933	5.381	81.777	< 0.001	0.172
Residuals	50.252	763.725	0.066			

Note. Principal effect or repeated-measures factor: MVN_tests (five variables: QSWa, QSWb, QSSFa, QSFb, and H). Interactions between the type of MVN test and three covariates: *n* = sample size (eight levels: 50, 75, 100, 125, 150, 200, 250, and 500); *k* = number of variables (five levels: 2, 3, 4, 5, and 6); and homogeneous inter-variable correlation (five levels: *r* = 0, 0.3, 0.5, 0.7, and 0.9). Between-subjects factor: distribution (two levels: multivariate t distribution and multivariate standard normal distribution). Statistics: *SS* = type III sum of squares; *df* = degrees of freedom; *MS* = mean square; *F* = test statistic; *p* = probability value under the null hypothesis of equal means; η_p^2 = partial eta squared.

Table 5. Multivariate analysis using Pillai's trace in the pooled sample of 400 observations of 5-tuples.

Effect	*V*	*F*	*df*_H_0	*df*_error	*p*	η_p^2	ϕ
MVN_tests	0.097	10.583b	4	392	<0.001	0.097	1
MVN_tests × *n*	0.101	10.976b	4	392	<0.001	0.101	1
MVN_tests × *k*	0.224	28.315b	4	392	<0.001	0.224	1
MVN_tests × *r*	0.034	3.461b	4	392	0.009	0.034	0.857
MVN_tests × Distribution	0.407	67.169b	4	392	<0.001	0.407	1

Note. Within-subjects design: MVN_Test. Design: Intersection + *n* + *k* + *r*. + Dist. *V* = Pillai's trace, F = test statistic, *df*_H_0 = degrees of freedom for the null hypothesis of equal means, *df*_ error = error degrees of freedom, *p* = probability value under the null hypothesis of equal means, η_p^2 = partial eta squared, and ϕ = statistical power.

The interactions of test type with sample size (F[1.933, 763.725] = 6.564, p = 0.002, η_p^2 = 0.016), number of variables (F[1.933, 763.725] = 22.286, p < 0.001, η_p^2 = 0.053), homogeneous inter-variable correlation (F[1.933, 763.725] = 3.797, p = 0.024, η_p^2 = 0.010), and distribution type (F[1.933, 763.725] = 81.777, p < 0.001, η_p^2 = 0.172) were statistically significant, with small effect sizes except for the interaction with distribution type, which showed a large effect (Table 4). These interactions were also significant in the multivariate test, with a small effect size for the interaction with homogeneous correlation (η_p^2 = 0.034), medium for the interaction with sample size (η_p^2 = 0.101), and large for the interactions with number of variables (η_p^2 = 0.224) and distribution type (η_p^2 = 0.407). See Table 5.

Correlations between sample size and statistical power were positive, with three reaching statistical significance, ranging from 0.114 for the H-test to 0.308 for the

QSWb test. Two statistically significant correlations between power and the number of variables were negative and moderate, and both appeared in the Q-test variants based on the chi-square approximation. Similarly, two significant negative but small correlations were observed between power and the level of homogeneous inter-variable correlation in the two bootstrap variants of the Q-test (see Table 6). Overall, power increased with larger sample sizes, lower inter-variable correlations, and fewer variables. The Q-test benefited more from these three factors than the H-test.

Table 6. Pearson product-moment correlation coefficients and 95% BCa confidence intervals in the pooled sample of 400 observations of 5-tuples.

MVN	Sttaistical power		
Test	*n*	*k*	*r*
QSWa	0.092 (−0.03, 0.207)	−0.349* (−0.429, −0.265)	−0.017 (−0.119, 0.074)
QSWb	0.308* (0.251, 0.358)	−0.083 (−0.179, 0.017)	−0.152* (−0.230, −0.065)
QSFa	0.086 (−0.031, 0.195)	−0.300* (−0.383, −0.211)	−0.001 (−0.097, 0.093)
QSFb	0.278* (0.222, 0.326)	−0.020 (−0.122, 0.086)	−0.153* (−0.234, −0.069)
H	0.114* (0.034, 0.270)	−0.022 (−0.189, 0.065)	−0.014 (−0.110, 0.063)

Note. *n* = sample size; *k* = number of variables; *r* = homogeneous inter-variable correlation. Results from the BCa confidence interval sampling simulation are based on 1000 samples. *The asterisk indicates a statistically significant correlation in a two-tailed test at the 0.05 significance level, as zero was not included in the confidence interval.

The between-subjects effects of sample size ($F[1, 395] = 24.405$, $p < 0.001$, $\eta_p^2 = 0.058$), number of variables ($F[1, 395] = 20.431$, $p < 0.001$, $\eta_p^2 = 0.049$), and distribution ($F[1, 395] = 28.067$, $p < 0.001$, $\eta_p^2 = 0.066$) were significant, with small effect sizes for the first two and a medium effect size for the last. However, the between-subjects effect of homogeneous inter-variable correlation was not significant ($F[1, 395] = 3.241$, $p = 0.073$, $\eta_p^2 = 0.008$). Refer to Table 7 for more details.

Table 7. Between-subjects effects in the pooled sample of 400 observations of 5-tuples.

Effect	*SS*	*df*	*MS*	*F*	*p*	η_p^2
n	3.772	1	3.772	24.405	< 0.001	0.058
k	3.157	1	3.157	20.431	< 0.001	0.049
r	0.501	1	0.501	3.241	0.073	0.008
Distribution	4.338	1	4.338	28.067	< 0.001	0.066
Residuals	61.046	395	0.155			

Note. *SS* = type III sum of squares; *df* = degrees of freedom; *MS* = mean square; *F* = test statistic; *p* = probability value under the null hypothesis of equal means; η_p^2 = partial eta squared.

When pairwise comparisons were conducted using Student's t-test for paired samples, with Bonferroni's correction applied to control the familywise error rate, significant differences were found in 4 out of the 10 tests. Mean power was significantly higher in both variants of the Q-test (chi-square approximation and bootstrap approach) when calculated using Shapiro-Wilk's W statistic compared to Shapiro-Francia's W' statistic. The effect size was small for the difference between QSWa and QSFa (Cohen's $d = -0.207$, 95% *CI*: −0.258, −0.156), and trivial for the remaining three tests. In contrast, mean power was equivalent when comparing the H-test to the four Q-test variants, as well as when comparing the chi-square approximation and bootstrap versions of the Q-test, regardless of whether they were calculated using the Shapiro-Wilk W or the Shapiro-Francia W' statistics. See Table 8 and Figure 2.

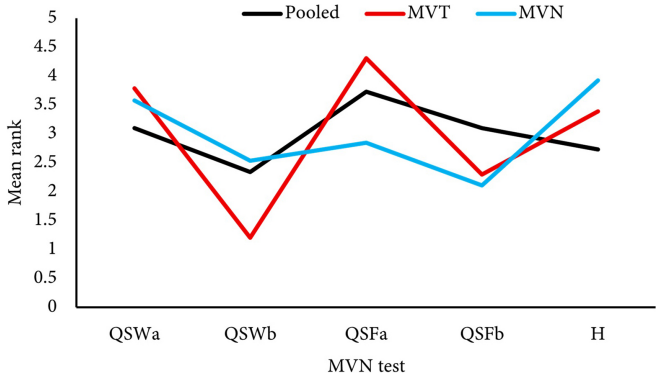

Figure 2. Mean ranks of statistical power of the five MVN tests in the 400 pooled observations, in 200 observations from MVT distributions, and 200 observations from MVN distributions.

Table 8. Post hoc pairwise comparisons of means between the five MVN tests in the pooled sample of 400 observations of 5-tuples.

Comparisons		Differences				t-test		Cohen's d		
Test 1	Test 2	*m*	*LL*	*UL*	*se*	*t*	p_{Bonf}	*d*	*LL*	*UL*
QSWa	QSWb	−0.007	−0.039	0.026	0.011	−0.597	1	−0.029	−0.166	0.108
	QSFa	−0.049	−0.06	−0.038	0.004	−12.537	<0.001	−0.207	−0.258	−0.156
	QSFb	−0.037	−0.07	−0.004	0.012	−3.122	0.019	−0.155	−0.295	−0.014
	H	−0.041	−0.084	0.003	0.015	−2.637	0.087	−0.171	−0.356	0.013
QSWb	QSFa	−0.042	−0.072	−0.013	0.01	−4.049	<0.001	−0.178	−0.304	−0.053
	QSFb	−0.030	−0.037	−0.022	0.003	−11.443	<0.001	−0.126	−0.159	−0.092
	H	−0.034	−0.083	0.016	0.018	−1.926	0.548	−0.143	−0.352	0.067
QSFa	QSFb	0.012	−0.017	0.042	0.011	1.188	1	0.053	−0.073	0.178
	H	0.008	−0.034	0.051	0.015	0.567	1	0.036	−0.142	0.213
QSFb	H	−0.004	−0.054	0.046	0.018	−0.229	1	−0.017	−0.226	0.192

Note. Comparisons = Test 1 − Test 2; Differences: *m* = arithmetic mean; *LL* = lower limit and *UL* = upper limit of the 95% asymptotic confidence interval for the mean difference; se = standard error. t-test: *t* = test statistic; p_{Bonf} = Bonferroni-adjusted two-tailed *p*-value for an alpha level of 0.05, calculated from the t distribution with 395 degrees of freedom. Cohen's *d* = effect size statistic; *LL* = lower limit and *UL* = upper limit of the 95% asymptotic confidence interval for *d*.

3.2.2. Friedman Test

It is worth noting that not only was the assumption of sphericity violated, but the residuals also deviated significantly from normality (Shapiro-Wilk test: W = 0.865, $p < 0.001$ for QSWa; $W = 0.858$, $p < 0.001$ for QSWb; $W = 0.795$, $p < 0.001$ for QSFa; $W = 0.813$, $p < 0.001$ for QSFb; and $W = 0.377$, $p < 0.001$ for H). Furthermore, the joint distribution of the five repeated measures did not follow a multivariate normal distribution (Royston $H = 0.691$, $p < 0.001$). According to the Friedman omnibus test, the null hypothesis of equality of central tendency across the five tests was also rejected (χ^2 [4, $N = 400$] = 181.580, $p < 0.001$), with a small effect size (Kendall $W = 0.113$).

Post hoc comparisons using the Conover test showed that 9 out of the 10 pairwise differences were statistically significant, both with and without Bonferroni's correction. The only non-significant difference was between QSWa and QSFb, whose mean power was statistically equivalent. Based on rank-biserial correlations, the effect of test type on power was very large for the comparison between QSWa and QSFa (r_{rb} = −0.794), large between QSWb and QSFb (r_{rb} = −0.670), and medium between QSFa and H (r_{rb} = 0.445). Effect sizes were small for the remaining comparisons, except for a trivial effect in the comparison between QSFb and H. Among the five tests, the Q-test variants based on the chi-square approximation (QSWa and QSFa) showed the highest statistical power, with QSFa outperforming QSWa. See Table 9.

Table 9. Conover's post hoc comparisons of mean rank sums of the five MVN tests in the pooled sample of 400 observations of 5-tuples.

Test 1	Test 2	W_i	W_j	t	df	p	p_{Bonf}	r_{rb}
QSWa	QSWb	1241.5	935.5	7.507	1596	< 0.001	< 0.001	0.102
	QSFa	1241.5	1491.5	6.133	1596	< 0.001	< 0.001	−0.794
	QSFb	1241.5	1239	0.061	1596	0.951	1	−0.012
	H	1241.5	1092.5	3.655	1596	< 0.001	0.003	0.152
QSWb	QSFa	935.5	1491.5	13.64	1596	< 0.001	< 0.001	−0.252
	QSFb	935.5	1239	7.446	1596	< 0.001	< 0.001	−0.670
	H	935.5	1092.5	3.852	1596	< 0.001	0.001	−0.092
QSFa	QSFb	1491.5	1239	6.195	1596	< 0.001	< 0.001	0.138
	H	1491.5	1092.5	9.789	1596	< 0.001	< 0.001	0.445
QSFb	H	1239	1092.5	3.594	1596	< 0.001	0.003	0.040

Note. W_i = sum of ranks for Test 1, W_j = sum of ranks for Test 2, t = t-test statistic, df = degrees of freedom, p = asymptotic two-tailed p-value from the t distribution with 1596 degrees of freedom, p_{Bonf} = Bonferroni-adjusted p-value for an alpha level of 0.05, r_{rb} = rank-biserial correlation based on individual signed-rank tests, used as a measure of size effect.

3.2.3. Effect of Significance Level on Q-Test Mean Power (Bootstrap Version)

In the pooled sample, the bootstrap statistical power—particularly for multivariate t-distributions and its complement in multivariate standard normal distributions—of the Q-test calculated using the Shapiro-Wilk W statistic was significantly higher ($md = 0.021$, 95% BCa CI: 0.012, 0.029; two-tailed $p_{boot} < 0.001$) when

applying a 5% significance level (ϕ_{boot} = 0.807, 95% BCa *CI*: 0.783, 0.830) than when applying a 10% significance level (ϕ_{boot} = 0.786, 95% BCa *CI*: 0.763, 0.810). The effect size of the significance level on power was small (Hedges-Olkin g = 0.255, 95% asymptotic *CI*: 0.155, 0.355).

Additionally, the bootstrap statistical power—particularly for multivariate t-distributions and its complement in multivariate standard normal distributions—of the Q-test calculated using the Shapiro-Francia W' statistic was significantly higher (md = 0.020, 95% BCa *CI*: 0.011, 0.028); two-tailed p_{boot} < 0.001) when applying a 5% significance level (ϕ_{boot} = 0.837, 95% BCa *CI*: 0.816, 0.856) than when using a 10% significance level (ϕ_{boot} = 0.817, 95% BCa *CI*: 0.797, 0.838). The effect size of the significance level on potency was small (Hedges-Olkin g = 0.263, 95% asymptotic *CI*: 0.163, 0.362). For both variants of the Q-test, 5% was a better option than 10% at a significance level.

3.3. Power Comparison across 200 Samples Drawn from the MVT Distribution

3.3.1. Repeated Measures ANOVA

Based on the 200 data points per MVN test, derived from samples drawn from the multivariate t-distributions with five degrees of freedom, the highest mean statistical power was observed in the H-test. Its confidence interval overlapped with those of the two Q-test variants based on chi-square approximation. In contrast, the lowest mean power was found in the QSWb, whose interval overlapped with that of QSFb; notably, the upper bounds of both intervals were lower than the lower bounds of the intervals for the other three MVN tests. Refer to Table 3 and Figure 3 for details.

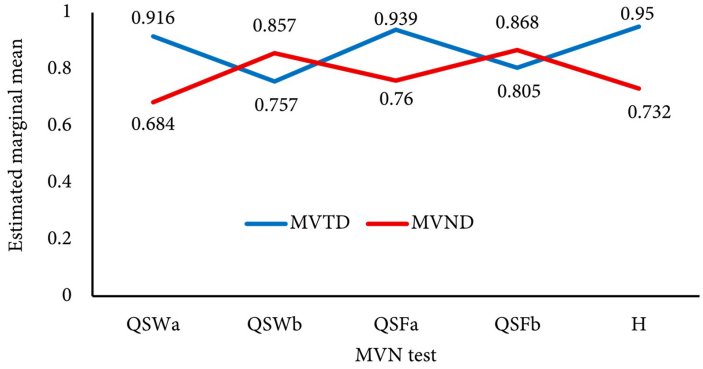

Figure 3. Stacked line chart showing the mean power (mean[ϕ] for MVTD and mean[1−ϕ] for MVND) of five multivariate normality tests, with separate lines representing the distribution from which each random sample was drawn (MVTD = multivariate t distribution with five degrees of freedom and MVND = multivariate standard normal distribution).

The assumption of sphericity was not met (W = 0.003, χ^2 [9, N = 200] = 1129.079, p < 0.001). After applying the Huynh-Feldt correction (ε = 0.343) to within-subject effects, the mean difference among the MVN tests was not statistically significant (F[1.374, 269.287] = 1.164, p = 0.299), resulting in a trivial effect

size (η_p^2 = 0.0004). However, the multivariate analysis using the Pillai's trace revealed a significant difference (V = 0.103, F[4, 193] = 5.511, p < 0.001), with a medium effect size (0.06 < η_p^2 = 0.103 < 0.14) and a very high statistical power (ϕ = 0.975). See Table 10 for more details.

The interactions of test type with sample size (F[1.374, 269.287] = 8.288, p = 0.002, η_p^2 = 0.041, ω_p^2 = 0.017), number of variables (F[1.374, 269.287] = 10.656, p < 0.001, η_p^2 = 0.052, ω_p^2 = 0.022), and homogeneous inter-variable correlation (F[1.374, 269.287] = 5.233, p = 0.014, η_p^2 = 0.026, ω_p^2 = 0.010) were statistically significant, with a small effect size. These interactions were also significant in the multivariate test, but with a medium effect size for the interactions with number of variables (η_p^2 = 0.134) and homogeneous inter-variable correlation (η_p^2 = 0.131), and a large effect size for the interaction with sample size (η_p^2 = 0.264). See Table 10. for details.

Table 10. Multivariate analysis using Pillai's trace in the sample of 200 observations of 5-tuples drawn from multivariate t-distributions with five degrees of freedom.

Effect	V	F	df_H_0	df_error	p	η_p^2	ϕ
MVN_tests	0.097	10.583b	4	392	<0.001	0.103	0.975
MVN_tests × n	0.101	10.976b	4	392	<0.001	0.264	1
MVN_tests × k	0.224	28.315b	4	392	<0.001	0.134	0.996
MVN_tests × r	0.034	3.461b	4	392	0.009	0.131	0.996

Note. Within-subjects design: MVN_Test. Design: Intersection + n + k + r. + Dist. V = Pillai's trace, F = test statistic, df_H_0 = degrees of freedom for the null hypothesis of equal means, df_ error = error degrees of freedom, p = probability value under the null hypothesis of equal means, η_p^2 = partial eta squared, and ϕ = statistical power.

Correlations between sample size and statistical power were positive, ranging from 0.109 in the H-test to 0.443 in the QSWb-test. Two significant, low-magnitude negative correlations between power and the number of variables were observed, both in the bootstrap variants of the Q-test. Similarly, correlations between statistical power and homogeneous inter-variable correlation were negative, ranging from −0.095 in the H-test to −0.353 in the QSFb-test (see Table 11). Overall, statistical power increased with larger sample sizes, weaker inter-variable correlations, and fewer variables—factors that most favored the two bootstrap versions of the Q-test, whereas the H-test benefited the least from these conditions.

Table 11. Pearson correlation coefficients and 95% BCa confidence intervals in the sample of 200 observations of 5-tuples drawn from multivariate t-distributions with five degrees of freedom.

MVN	Statistical power		
test	n	k	r
QSWa	0.261* (0.199, 0.321)	0.019 (−0.131, 0.182)	−0.232* (−0.352, −0.089)
QSWb	0.443* (0.378, 0.506)	−0.209* (−0.333, −0.073)	−0.343* (−0.444, −0.229)

Continued

QSFa	0.235* (0.178, 0.294)	0.034 (−0.124, 0.192)	−0.204* (−0.320, −0.062)
QSFb	0.383* (0.317, 0.444)	−0.203* (−0.325, −0.070)	−0.353* (−0.443, −0.251)
H	0.109* (0.021, 0.345)	0.114 (−0.059, 0.216)	−0.095* (−0.301, −0.034)

Note. n = sample size; k = number of variables; r = homogeneous inter-variable correlation. Results from the BCa confidence interval sampling simulation are based on 1,000 samples. *The asterisk indicates a statistically significant correlation in a two-tailed test at the 0.05 significance level, as zero was not included in the confidence interval.

The between-subjects effect of the sample size (F[1, 196] = 30.865, $p < 0.001$, $\eta_p^2 = 0.136$) and homogeneous inter-variable correlation (F[1, 196] = 22.461, $p < 0.001$, $\eta_p^2 = 0.103$) were significant, both showing medium effect sizes. However, the between-subjects effect of the number of variables was not significant (F[1, 196] = 0.812, $p = 0.369$, $\eta_p^2 = 0.004$).

When pairwise comparisons were conducted using Student's t-test for paired samples with Bonferroni's correction applied to control the familywise error rate, significant differences were found in 6 out of the 10 tests. Mean power was equivalent between the two Q-test variants based on chi-square approximation and H-test, as well as between the two bootstrap variants of the Q-test. Consequently, mean power was significantly higher for the chi-square approximation variants of the Q-test and the H-test compared to the two bootstrap variants of the Q-test. The effect size of small for the difference between QSWa and QSFb (Cohen's $d = 0.423$, 95% CI: 0.2, 0.646) and medium for the other five comparisons (0.5 < Cohen's d < 0.8). See Table 12 for details.

Table 12. Post hoc pairwise comparisons of means between the five MVN tests in the sample of 200 observations of 5-tuples drawn from multivariate t-distributions with five degrees of freedom.

Comparisons		Differences			t-test		Cohen's d statistic		
Test 1	Test 2	m	LL	UL	t	p_{Bonf}	d	LL	UL
QSWa	QSWb	0.160	0.104	0.215	8.041	<0.001	0.609	0.377	0.840
	QSFa	−0.022	−0.078	0.034	−1.124	1	−0.085	−0.3	0.130
	QSFb	0.111	0.055	0.167	5.592	<0.001	0.423	0.2	0.646
	H	−0.033	−0.089	0.023	−1.674	0.945	−0.127	−0.342	0.089
QSWb	QSFa	−0.182	−0.238	−0.126	−9.165	<0.001	−0.694	−0.93	−0.457
	QSFb	−0.049	−0.104	0.007	−2.449	0.145	−0.185	−0.402	0.031
	H	−0.193	−0.249	−0.137	−9.715	<0.001	−0.735	−0.975	−0.496
QSFa	QSFb	0.133	0.077	0.189	6.716	<0.001	0.508	0.282	0.735
	H	−0.011	−0.067	0.045	−0.55	1	−0.042	−0.256	0.173
QSFb	H	−0.144	−0.2	−0.088	−7.266	<0.001	−0.550	−0.779	−0.321

Note. Comparisons = Test 1 − Test 2; Differences: m = arithmetic mean; LL = lower limit and UL = upper limit of the 95% asymptotic confidence interval for the mean difference; t-test: t = test statistic (pooled standard error of 0.020); p_{Bonf} = Bonferroni-adjusted two-tailed p-value for an alpha level of 0.05, calculated from the t distribution with 784 degrees of freedom. Cohen's d = effect size statistic; LL = lower limit and UL = upper limit of the 95% asymptotic confidence interval for d.

3.3.2. Friedman Test

Not only was the assumption of sphericity violated, but the residuals also showed significant deviations from normality ($W = 0.709$, $p < 0.001$ for QSWa; $W = 0.906$, $p < 0.001$ for QSWb; $W = 0.626$, $p < 0.001$ for QSFa; $W = 0.888$, $p < 0.001$ for QSFb; and $W = 0.312$, $p < 0.001$ for H). Moreover, the joint distribution of the five repeated measures did not conform to multivariate normality ($H = 0.636$, $p < 0.001$). Based on the Friedman omnibus test, the null hypothesis of equal central tendency in power across the five MVN tests was also rejected (χ^2 [4] = 564.628, $p < 0.001$), yielding a large effect size (Kendall $W = 0.706$), rather than a medium one.

Post hoc comparisons using the Conover test revealed that the 10 differences were statistically significant, even after applying the Bonferroni's correction. According to the biserial-rank correlation, the effect size of the test type on statistical power was very large ($|r_{br}| > 0.7$), except for a large effect in the comparison between QSFa and H ($r_{br} = 0.526$) a small effect in the comparison between QSWa and H ($r_{br} = 0.221$). The chi-square approximation variants of Q-test demonstrated the highest power, while the bootstrap variants showed the lowest. Furthermore, the Q-test calculated using the Shapiro-Francia W' statistic was more powerful than the version based on the Shapiro-Wilk W statistic. See Table 13 and Figure 2.

Table 13. Conover's post hoc pairwise comparisons of mean rank sums between the five MVN tests in the sample of 200 observations of 5-tuples drawn from multivariate t-distributions with five degrees of freedom.

Test 1	Test 2	W_i	W_j	t	df	p	p_{Bonf}	r_{rb}
QSWa	QSWb	757	242.5	32.023	796	< 0.001	< 0.001	0.963
	QSFa	757	861.5	6.504	796	< 0.001	< 0.001	−0.818
	QSFb	757	460.5	18.454	796	< 0.001	< 0.001	0.864
	H	757	678.5	4.886	796	< 0.001	< 0.001	0.221
QSWb	QSFa	242.5	861.5	38.527	796	< 0.001	< 0.001	−0.997
	QSFb	242.5	460.5	13.568	796	< 0.001	< 0.001	−0.989
	H	242.5	678.5	27.137	796	< 0.001	< 0.001	−0.968
QSFa	QSFb	861.5	460.5	24.958	796	< 0.001	< 0.001	0.949
	H	861.5	678.5	11.39	796	< 0.001	< 0.001	0.526
QSFb	H	460.5	678.5	13.568	796	< 0.001	< 0.001	−0.867

Note. W_i = sum of ranks for Test 1, W_j = sum of ranks for Test 2, t = t-test statistic, df = degrees of freedom, p = asymptotic two-tailed p-value from the t distribution with 796 degrees of freedom, p_{Bonf} = Bonferroni-adjusted p-value for an alpha level of 0.05, r_{rb} = rank-biserial correlation based on individual signed-rank tests, used as a measure of size effect.

3.3.3. Effect of Significance Level on Q-Test Mean Power (Bootstrap Version)

In the 200 samples drawn from multivariate t-distributions, the bootstrap statistical power of the Q-test calculated using the Shapiro-Wilk W statistic (QSWb) was significantly lower ($md = -0.044$, 95% BCa CI: 0.051, −0.037; two-tailed $p_{boot} < 0.001$) when applying a 5% significance level ($\phi_{boot} = 0.757$, 95% BCa CI: 0.715,

0.798) compared to a 10% significance level (ϕ_{boot} = 0.801, 95% BCa *CI*: 0.763, 0.838). Therefore, the 10% level proved more power than the 5% level for rejecting the null hypothesis. The effect size of the significance level on statistical power was large (Hedges_Olkin g = −0.962, 95% asymptotic *CI*: −1.128, −0.793).

The bootstrap statistical power of the Q-test, calculated using the Shapiro–Francia W' statistics (QSFb), was significantly lower (md = −0.039, 95% BCa *CI*: −0.045, −0.032; two-tailed $p_{boot} < 0.001$) when applying a 5% significance level (ϕ_{boot} = 0.805, 95% BCa *CI*: 0.765, 0.844) compared to a 10% significance level (ϕ_{boot} = 0.844, 95% BCa *CI*: 0.807, 0.878). Therefore, the 10% level is preferable to the 5% level for rejecting the null hypothesis. The effect size of the significance level on statistical power was large (Hedges-Olkin g = −0.878, 95% asymptotic *CI*: −1.041, −0.715).

3.4. Power Comparison across 200 Samples Drawn from the MVN Distribution

3.4.1. Repeated Measures ANOVA

For samples drawn from a multivariate normal distribution, the null hypothesis of normality ($p \geq \alpha = 0.05$) should be retained, and statistical power should be less than 0.5 ($\phi < 0.5$). The closer the power is to zero, the better the performance of the statistical test under true normality.

Based on the 200 data points per MVN test, derived from samples drawn from multivariate standard normal distributions, the lowest mean power was observed in the QSFb. Its confidence interval overlapped with that of QSWb, and the upper bounds of both intervals were lower than the lower bounds of the intervals of the other three tests, whose confidence intervals overlapped with one another. See Table 3 and Figure 3 for details.

The assumption of sphericity was violated (W = 0.041, χ^2 [9, N = 200] = 621.211, $p < 0.001$). After applying the Huynh-Feldt correction (ε = 0.527) to within-subject effects, the mean difference among the MVN tests was statistically significant (F[2.107, 413.058] = 64.354, $p < 0.001$), with a large effect size (η_p^2 = 0.247). Multivariate analysis using the Pillai's trace also indicated a significant difference (V = 0.385, F[4, 193] = 30.161, $p < 0.001$), with a large effect size (η_p^2 = 0.385 > 0.14) and full statistical power (ϕ = 1).

The interactions of test type with sample size (F[2.107, 413.058] = 3.093, p = 0.044) and number of variables (F[2.107, 413.058] = 169.264, $p < 0.001$) were statistically significant. The effect size of the former was small (η_p^2 = 0.016), while the latter showed a large effect (η_p^2 = 0.463). In contrast, the interaction between test type and homogeneous inter-variable correlation was not significant (F[2.107, 413.058] = 0.316, p = 0.741, η_p^2 = 0.002).

In the multivariate analysis using Pillai's trace, the interaction between test type and number of variables remained significant and presented a large effect size (V = 0.652, F[4, 193] = 90.471, $p < 0.001$, η_p^2 = 0.652). However, the interaction between test type and sample size was not significant (V = 0.028, F[4, 193] = 1.386, p = 0.240, η_p^2 = 0.028), nor was the interaction between test type and homogene-

ous inter-variable correlation (V = 0.004, F[4, 193] = 0.197, p = 0.940, η_p^2 = 0.004).

The three significant correlations between sample size and statistical power were negative and low, ranging from -0.103 for the QSWb test to -0.188 for the H test. The correlations between statistical power and the number of variables were also low and negative for the two bootstrap variants of the Q-test, but positive for the other three tests—high for the two Q-test variants based on chi-square approximation, and moderate for the H-test. The five correlations between statistical power and the homogeneous correlation among variables were negative and low, ranging from −0.134 for QSWa to −0.224 for QSWb (Table 14). A larger sample size reduced the statistical power of the two bootstrap Q-test variants and the H-test. In contrast, a higher inter-variable correlation increased the statistical power of both the Q and H tests. A smaller number of variables benefited the two Q-test variants based on chi-square approximation and, to a lesser extent, the H-test. Conversely, a larger number of variables favored the two bootstrap variants of the Q-test. See Table 14 for more details.

Table 14. Pearson product-moment correlation coefficients and 95% BCa confidence intervals in the sample of 200 observations of 5-tuples drawn from MVN distributions.

MVN	Statiistical power		
test	n	k	r
QSWa	0.013 (−0.148, 0.183)	0.680* (0.607, 0.740)	−0.134* (−0.263, −0.008)
QSWb	−0.103* (−0.197, −0.006)	−0.171* (−0.286, −0.040)	−0.224* (−0.352, −0.064)
QSFa	0.011 (−0.136, 0.170)	0.604* (0.528, 0.672)	−0.152* (−0.296, −0.020)
QSFb	−0.125* (−0.208, −0.035)	−0.337* (−0.436, −0.229)	−0.212* (−0.338, −0.053)
H	−0.188* (−0.289, −0.076)	0.375* (0.257, 0.483)	−0.184* (−0.327, −0.026)

Note. n = sample size; k = number of variables; r = homogeneous inter-variable correlation. Results from the BCa confidence interval sampling simulation are based on 1000 samples. *The asterisk indicates a statistically significant correlation in a two-tailed test at the 0.05 significance level, as zero was not included in the confidence interval.

The between-subjects effects of the number of variables (F[1, 196] = 54.620, p < 0.001, η_p^2 = 0.218) and the homogeneous inter-variable correlation (F[1, 196] = 13.191, p < 0.001, η_p^2 = 0.063) were significant with large and medium effect sizes, respectively. However, the between-subjects effect of the sample size was not significant (F[1, 196] = 1.479, p = 0.225, η_p^2 = 0.007).

When conducting pairwise comparisons using Student's t-test for paired samples with Bonferroni's correction to control the familywise error rate, significant differences were observed in 8 out of the 10 tests. The mean power of the QSWb test was equivalent to that of the QSFb test, while the QSFa test showed equivalent power to the H-test. The effect size of the test type on statistical power was large

in three comparisons (QSWa vs. QSWb, QSWa vs. QSFb, and QSFb vs. H), favoring the bootstrap variant of the Q-test; medium in three others (QSWb vs. H, QSFa vs. QSFb, and QSWb vs. QSFa); and small in the remaining two (QSWa vs. QSFa and QSWa vs. H). The two bootstrap variants of the Q-test demonstrated to be statistically equivalent and had the lowest mean power values, indicating better performance. See Table 15.

Table 15. Post hoc pairwise comparisons of means between the five MVN tests in the sample of 200 observations of 5-tuples drawn from MVN distributions.

Comparisons		Differences			t-test		Cohen's d statistic		
Test 1	Test 2	*m*	*LL*	*UL*	*t*	p_{Bonf}	*d*	*LL*	*UL*
QSWa	QSWb	0.173	0.141	0.206	14.992	<0.001	1.051	0.801	1.3
	QSFa	0.076	0.043	0.109	6.577	<0.001	0.461	0.251	0.67
	QSFb	0.184	0.152	0.217	15.947	<0.001	1.118	0.862	1.373
	H	0.048	0.016	0.081	4.166	<0.001	0.292	0.089	0.495
QSWb	QSFa	−0.097	−0.13	−0.065	−8.415	<0.001	−0.590	−0.806	−0.374
	QSFb	0.011	−0.022	0.044	0.955	1	0.067	−0.132	0.266
	H	−0.125	−0.158	−0.093	−10.826	<0.001	−0.759	−0.985	−0.532
QSFa	QSFb	0.108	0.076	0.141	9.37	<0.001	0.657	0.437	0.877
	H	−0.028	−0.06	0.005	−2.41	0.162	−0.169	−0.369	0.031
QSFb	H	−0.136	−0.169	−0.104	−11.781	<0.001	−0.826	−1.057	−0.594

Note. Comparisons: Test 1 − Test 2; Differences: *m* = arithmetic mean; *LL* = lower limit and *UL* = upper limit of the 95% asymptotic confidence interval for the mean difference; t-test: *t* = t-test statistic (pooled standard error of 0.012); p_{Bonf} = Bonferroni-adjusted two-tailed *p*-value for an alpha level of 0.05, calculated from the t distribution with 784 degrees of freedom. Cohen's *d* = effect size statistic; *LL* = lower limit and *UL* = upper limit of the 95% asymptotic confidence interval for *d*.

3.4.2. Friedman Test

Not only was the assumption of sphericity violated, but the residuals also showed significant deviations from normality ($W = 0.978$, $p = 0.003$ for QSWa; $W = 0.903$, $p < 0.001$ for QSWb; $W = 0.934$, $p < 0.001$ for QSFa; $W = 0.912$, $p < 0.001$ for QSFb; and $W = 0.966$, $p < 0.001$ for H). Additionally, the joint distribution of the five repeated measures did not conform to multivariate normality ($H = 0.861$, $p < 0.001$). Friedman's omnibus test also rejected the null hypothesis of equal central tendency in statistical power across the five MVN tests (χ^2 [4] = 179.323, $p < 0.001$), with a small effect size (Kendall $W = 0.224$).

Post hoc comparisons using the Conover test revealed that all 10 differences were significant without correction for the Type I error rate (false positives). However, after applying Bonferroni's correction, the differences between QSWa and H ($p = 0.116$) and between QSWb and QSFa ($p = 0.241$) were no longer significant. According to the biserial-rank correlation, the effect size of test type on statistical power was large for the comparison between QSWa and QSFa ($r_{br} = 0.845$), medium for four comparisons (QSWa vs. QSWb, QSWa vs. QSFb, QSWb vs. H, and QSFb vs. H), and small for the remaining three (QSWb vs. QSFb, QSFa vs. QSFb, and QSFa vs. H). The two bootstrap variants of the Q-test showed the lowest

power (*i.e.*, best performance), with power being lower when the Q-test was based on the Shapiro-Francia W' statistic than when based on the Shapiro-Wilk W statistic. See Table 16 and Figure 2 for details.

Table 16. Conover's post hoc pairwise comparisons of mean rank sums between the five MVN tests in the sample of 200 observations of 5-tuples drawn from MVN distributions.

Test 1	Test 2	W_i	W_j	t	df	p	p_{Bonf}	r_{rb}
QSWa	QSWb	715.5	507	7.48	796	< 0.001	< 0.001	0.596
	QSFa	715.5	570	5.22	796	< 0.001	< 0.001	0.845
	QSFb	715.5	421.5	10.547	796	< 0.001	< 0.001	0.583
	H	715.5	786	2.529	796	0.012	0.116	-0.090
QSWb	QSFa	507	570	2.260	796	0.024	0.241	-0.444
	QSFb	507	421.5	3.067	796	0.002	0.022	0.333
	H	507	786	10.009	796	< 0.001	< 0.001	-0.683
QSFa	QSFb	570	421.5	5.327	796	< 0.001	< 0.001	0.456
	H	570	786	7.749	796	< 0.001	< 0.001	-0.361
QSFb	H	421.5	786	13.076	796	< 0.001	< 0.001	-0.692

Note. W_i = sum of ranks for Test 1, W_j = sum of ranks for Test 2, t = t-test statistic, df = degrees of freedom, p = asymptotic two-tailed p-value from the t distribution with 796 degrees of freedom, p_{Bonf} = Bonferroni-adjusted p-value for an alpha level of 0.05, r_{rb} = rank-biserial correlation based on individual signed-rank tests, used as a measure of size effect.

3.4.3. Effect of Significance Level on Q-Test Mean Power (Bootstrap Version)

In the 200 samples drawn from the multivariate standard normal distributions — under which the null hypothesis should be retained—the bootstrap statistical power of the Q-test calculated using the Shapiro-Wilk W statistic was significantly lower (md = −0.086, 95% BCa CI: −0.094, −0.079; two-tailed p_{boot} < 0.001) when using a 5% significance level (ϕ_{boot} = 0.143, 95% BCa CI: 0.124, 0.161) compared to a 10% significance level (ϕ_{boot} = 0.229, 95% BCa CI: 0.205, 0.251). Therefore, the 5% level yielded lower power (better performance) than the 10% level when the null hypothesis should be retained. The effect size of the significance level on statistical power was very large (Hedges-Olkin g = −1.687, 95% asymptotic CI: −1.901, −1.470).

Additionally, the bootstrap statistical power of the Q-test, calculated with the Shapiro–Francia W' statistics, was significantly lower (md = −0.078, 95% BCa CI: −0.085, −0.071; two-tailed p_{boot} < 0.001) when using a 5% significance level (ϕ_{boot} = 0.132, 95% BCa CI: 0.114, 0.151) compared to a 10% significance level (ϕ_{boot} = 0.210, 95% BCa CI: 0.188, 0.233). Therefore, the 5% level is preferable to the 10% level when the null hypothesis should be retained. The size of the effect of the significance level on statistical power was very large (Hedges-Olkin g = −1.605, 95% asymptotic CI: −1.813, −1.394).

4. Discussion

The objective of the study was to compare the hit rate and mean statistical power

across four variants of the Q-test and the H-test (five MVN tests in total). A total of 200 random samples were generated from multivariate normal distributions and another 200 from multivariate t-distributions with five degrees of freedom. Since the same seed (123) was used for sample generation, the MVN tests reflect repeated measures of normality. Accordingly, tests designed for repeated measures were applied. Sample size (eight levels: 50, 75, 100, 125, 150, 200, 250, and 500), number of variables (five levels: 2, 3, 4, 5, and 6), and homogeneous inter-variable correlation (five levels: 0, 0.3, 0.5, 0.7, and 0.9) were included as key factors or covariates in the repeated measures ANOVA of mean power [8].

Unlike in the previous study [2], the chi-square approximation of the Q-test did not subtract the number of *z*' or *z*-values set to zero when determining the degree of freedom. The significance level was set at 5% for the chi-square approximation and the bootstrap variants of the Q-test, consistent with the previous study. In the present study, however, we compared the performance of the two bootstrap variants of the Q-test (SW vs. SF) under two significance levels: the conventional 5% level and an elevated 10% level, which was applied to compensate for the pronounced elongation of the right tail in the sampling distribution of the Q statistic.

On the one hand, differences in hit rate were examined using Cochran's Q test as an omnibus test, followed by paired samples t-tests with estimated joint variance and Bonferroni's correction for pairwise comparisons [35]. This analysis was conducted on both the pooled sample ($n = 400$) and separately by distribution ($n = 200$ for MVN and $n = 200$ for MVT). On the other hand, differences in mean statistical power were assessed using repeated measures ANOVA. Due to clear violations of the sphericity assumption, multivariate analysis and the Huynh-Feldt epsilon correction were applied to account for within-subject effects. However, since the assumption of multivariate normality was not met for either the error distribution or the repeated measures (*i.e.*, the power estimates for each MVN test), the results of this parametric analysis should be interpreted with caution and given less weight. Conversely, greater emphasis should be placed on the findings from the omnibus nonparametric Friedman test for comparing central tendency, Kendall's W for assessing the effect size of test type on statistical power, and the pairwise comparisons conducted using the Conover test with Bonferroni correction, with effect sizes evaluated via rank-biserial correlation.

The hit rate is equivalent across all five MVN tests when samples are drawn from multivariate normal distributions. However, for samples from multivariate t-distributions and in the pooled sample, both the H-test and Q-test version based on the chi-square approximation calculated using the Shapiro-Francia W' statistic show an advantage, yielding comparable hit rates. In the Q-test, the chi-square approximation outperforms the bootstrap approach, and calculations using the Shapiro-Francia W' statistic outperform those using the Shapiro-Wilk W statistic. The effect size of the MVN test type on the hit rate was large in the 200 samples drawn from the multivariate t-distributions and medium in the pooled sample of 400 observations, as indicated by the eta-squared coefficient. In practical terms,

the choice of MVN test meaningfully affects the hit rate when rejecting the null hypothesis.

Regarding statistical power and based on the non-parametric analysis, the Q-test variants based on the chi-square approximation demonstrated the highest power, while the bootstrap variants of the Q-test applied to samples drawn from the multivariate t-distributions and the pooled sample showed the lowest. Among the Q-tests, the variant computed using the Shapiro-Francia W' statistic (SF) was more powerful than the one calculated using the Shapiro-Wilk W statistic (SW). The H-test exhibited intermediate power, falling between the two Q-test variants based on the chi-square approximation and being more closely aligned with QSWa than QSFa. For samples drawn from multivariate standard normal distributions, the H-test and the Q-test variant based on the chi-square approximation, calculated using the Shapiro-Wilk W statistic, produced the highest power. In contrast, the bootstrap variants of the Q-test consistently showed the lowest power. The effect size of the MVN test type on statistical power was medium for the 200 samples drawn from multivariate t-distributions, and small for both the 200 samples drawn from multivariate normal distributions and the pooled sample of 400 observations assessed using Kendall's W coefficient. In practical terms, the choice of MVN test is relevant to statistical power when rejecting the null hypothesis.

When a 10% significance level is used for the bootstrap variants of the Q-test to compensate for the highly skewed right tail, the hit rate improves significantly when the Shapiro-Francia W' statistic is used and the null hypothesis must be rejected. This effect is observed in the 200 samples drawn from the multivariate t-distributions and the pooled sample, although the effect size of the significance level on the hit rate is small. When the Shapiro-Wilk W statistic is used, or when the null hypothesis of multivariate normality must be retained, no improvement is observed.

Statistical power, when the null hypothesis must be rejected (*i.e.*, for samples drawn from multivariate t-distributions), increases when using a 10% significance level, with a large effect size for both QSWb and QSFb. However, when the null hypothesis must be retained (*i.e.*, for samples drawn from multivariate normal distributions), a 5% significance level yields the best results, showing a large effect size for QSWb and a very large effect size for QSFb. Consequently, the optimal significance level for maximizing power in the pooled sample (including both t and normal distributions) is 5%, where the effect size is small for both QSWb and QSFb.

Based on the analysis of within-subject effects and the correlations between covariates and statistical power, several clear patterns emerge. For samples from multivariate t-distributions and for the pooled sample, power increases with larger sample sizes, lower inter-variable correlations, and fewer variables. These conditions particularly favor the Q-test: a smaller number of variables improves the accuracy of its chi-square approximation, while lower correlations enhance the per-

formance of its bootstrap version. The H-test, by contrast, gains the least from these favorable conditions, making it the least robust option when heavy-tailed data are present.

For samples from multivariate normal distributions, where the null hypothesis should be retained, higher sample sizes generally reduce statistical power—indicating better control of Type I error. In this context, having fewer variables improves the performance of both chi-square-based Q-test variants and the H-test. The bootstrap Q-test variants, however, benefit from the opposite trend: an increased number of variables strengthens their ability to correctly retain the null. Additionally, higher inter-variable correlations consistently improve the performance of all Q-test variants and the H-test when normality holds.

In practical terms, these results suggest that larger samples and simpler data structures—with fewer variables and weaker intercorrelations—are advantageous for detecting departures from multivariate normality, particularly for the Q-test. When normality holds, however, stronger correlations and an appropriate match between test type and dimensionality are key: the chi-square Q-tests and H-test work best with fewer variables, while the bootstrap Q-tests are better equipped to handle the complexity introduced by a larger number of variables. This interaction underscores the importance of tailoring the choice of normality test to the expected data structure.

One limitation of this study is that only the multivariate t-distribution with five degrees of freedom was considered as an alternative to multivariate normality. This distribution significantly deviates from normality due to its pronounced leptokurtosis. Other distributions characterized by asymmetry (e.g., Dirichlet distribution), mesokurtosis (e.g., multivariate uniform distribution), or those closer to normality (e.g., multivariate logistic distribution or multivariate t-distribution with 50 degrees of freedom) were not included [36]. Consistent with the original study [2], the Q-test based on the Shapiro-Francia W' statistic outperformed the Shapiro-Wilk W statistic for the multivariate t-distribution. However, this advantage did not hold for the chi-square distribution in the mentioned study.

5. Conclusion

In this simulation study comparing two multivariate normality tests—one of which includes four variants—under two types of distributions (one favoring the null hypothesis and the other opposing it), the Royston H-test outperformed the bootstrap variants of the Q-test. However, it did not outperform the Q-test variant based on the chi-square approximation when using the Shapiro-Wilk W statistic for multivariate normal samples or the Shapiro-Francia W' statistic for multivariate t-distributed samples. The bootstrap Q-test variants performed best when calculated using the Shapiro-Francia W' statistic. In such cases, a 10% significance level is preferable to a 5% level in terms of hit rate when either retaining or rejecting the null hypothesis, and in terms of statistical power only when rejecting the null hypothesis. This approach is more conservative regarding the null hypothesis

of multivariate normality than both the H-test and the chi-square approximation-based Q-test variants, as evidenced by its lower power in correctly rejecting the null for samples from a multivariate t-distribution. Based on the sampling distribution of the Q-statistic revealed by the bootstrap method, subtracting the number of z or z'-statistics truncated to zero when calculating the degrees of freedom in the chi-square approximation-based Q-test variants is not recommended.

6. Suggestions for Further Research

For the Q-test, the chi-square approach is preferable when the assumption of serial independence holds. This assumption can be evaluated using the Wald-Wolfowitz runs test [37] and the Ljung-Box Q-test [38], both of which are included in the MVT Q-test implementation in R [1]. If the assumption is violated, the bootstrap version is theoretically more appropriate. In such cases, it is recommended to compute the Q-test using the Shapiro-Francia W' statistic and a 10% significance level. The advantage of using a 10% significance level with this variant of the Q-test is reflected in the hit rate, showing a small effect size. In terms of statistical power, however, a 5% level performs better when the null hypothesis is retained with a very large effect size, whereas a 10% level is preferable when the null hypothesis is rejected with a large effect size.

When running the R script for the MVT Q-test, the convergence of results between the Q-test and Royston's H-test provides further evidence supporting the validity of the conclusions. Future research on the MVT Q-test should broaden the range of non-normal multivariate distributions considered, including the multivariate log-normal, Dirichlet, gamma, uniform, contaminated normal, and generalized hyperbolic distributions. It may also incorporate the multivariate normality test developed by Henze and Zirkler [39], which is available in R [40].

Acknowledgements

The author expresses gratitude to the reviewers and editor for their helpful comments.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Moral, J. (2025) MVN Q-Test I: A Bootstrap-Based Implementation in R. *Journal of Data Analysis and Information Processing*, **13**, 389-424. https://doi.org/10.4236/jdaip.2025.134024

[2] Rubia, J.M.L. (2023) Proposal and Pilot Study: A Generalization of the W or W' Statistic for Multivariate Normality. *Open Journal of Statistics*, **13**, 119-169. https://doi.org/10.4236/ojs.2023.131008

[3] Royston, P. (1992) Approximating the Shapiro-Wilk W-Test for Non-Normality. *Statistics and Computing*, **2**, 117-119. https://doi.org/10.1007/bf01891203

[4] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality

(Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.2307/2333709

[5] Royston, P. (1993) A Toolkit for Testing for Non-Normality in Complete and Censored Samples. *The Statistician*, **42**, 37-43. https://doi.org/10.2307/2348109

[6] Shapiro, S.S. and Francia, R.S. (1972) An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, **67**, 215-216. https://doi.org/10.1080/01621459.1972.10481232

[7] Royston, J.P. (1983) Some Techniques for Assessing Multivarate Normality Based on the Shapiro-Wilk W. *Applied Statistics*, **32**, 121-133. https://doi.org/10.2307/2347291

[8] Jobst, L.J., Bader, M. and Moshagen, M. (2023) A Tutorial on Assessing Statistical Power and Determining Sample Size for Structural Equation Models. *Psychological Methods*, **28**, 207-221. https://doi.org/10.1037/met0000423

[9] Braun, W.J. and Murdoch, D.J. (2021) A First Course in Statistical Programming with R. 3rd Edition, Cambridge University Press. https://doi.org/10.1017/9781108993456

[10] Giorgi, F.M., Ceraolo, C. and Mercatelli, D. (2022) The R Language: An Engine for Bioinformatics and Data Science. *Life*, **12**, Article No. 648. https://doi.org/10.3390/life12050648

[11] Anis, W., Kuntoro, K. and Melaniani, S. (2021) Difference of Power Test and Type II Error (β) on Mardia MVN Test, Henze Zikler's MVN Test, and Royston's MVN Test Using Multivariate Data Analysis. *Jurnal Biometrika dan Kependudukan*, **10**, 153-161. https://doi.org/10.20473/jbk.v10i2.2021.153-161

[12] Khatun, N. (2021) Applications of Normality Test in Statistical Analysis. *Open Journal of Statistics*, **11**, 113-122. https://doi.org/10.4236/ojs.2021.111006

[13] Mardia, K.V. (1980) 9 Tests of Unvariate and Multivariate Normality. In: Krishnaiah, P.R., Ed., *Handbook of Statistics* 1: *Analysis of Variance*, Elsevier, 279-320. https://doi.org/10.1016/s0169-7161(80)01011-5

[14] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M. and Hothorn, T. (2025) mvtnorm: Multivariate Normal and t Distributions (Version 1.3-3). https://doi.org/10.32614/CRAN.package.mvtnorm

[15] Wilson, E.B. (1927) Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, **22**, 209-212. https://doi.org/10.1080/01621459.1927.10502953

[16] Signorell, A. (2025) DescTools: Tools for Descriptive Statistics. Version 0.99.6. https://doi.org/10.32614/CRAN.package.DescTools

[17] Cochran, W.G. (1950) The Comparison of Percentages in Matched Samples. *Biometrika*, **37**, 256-266. https://doi.org/10.1093/biomet/37.3-4.256

[18] Tomczak, M. and Tomczak, E. (2014) The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size. *Trends in Sport Sciences*, **1**, 19-25.

[19] Sheskin, D.J. (2011) Handbook of Parametric and Non-Parametric Statistical Procedures. 5th Edition, Chapman and Hall/CRC.

[20] Barnett, M.J., Doroudgar, S., Khosraviani, V. and Ip, E.J. (2022) Multiple Comparisons: To Compare or Not to Compare, That Is the Question. *Research in Social and Administrative Pharmacy*, **18**, 2331-2334. https://doi.org/10.1016/j.sapharm.2021.07.006

[21] Bonferroni, C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3-62.

[22] Cohen, J. (1988) Statistical Power Analysis for Behavioral Sciences. 2nd Edition, Law-

rence Erlbaum Associates.

[23] IBM Corporation (2022) IBM SPSS Statistics for Windows, Version 27.0. IBM Corporation.

[24] Mauchly, J.W. (1940) Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics*, **11**, 204-209. https://doi.org/10.1214/aoms/1177731915

[25] Huynh, H. and Feldt, L.S. (1970) Conditions under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions. *Journal of the American Statistical Association*, **65**, 1582-1589. https://doi.org/10.2307/2284340

[26] Pillai, K.C.S. (1955) Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, **26**, 117-121. https://doi.org/10.1214/aoms/1177728599

[27] Din, I.U. and Hayat, Y. (2021) ANOVA or MANOVA for Correlated Traits in Agricultural Experiments. *Sarhad Journal of Agriculture*, **37**, 1250-1259. https://doi.org/10.17582/journal.sja/2021/37.4.1250.1259

[28] Rousselet, G., Pernet, C.R. and Wilcox, R.R. (2023) An Introduction to the Bootstrap: A Versatile Method to Make Inferences by Using Data-Driven Simulations. *MetaPsychology*, **7**, 1-24. https://doi.org/10.15626/mp.2019.2058

[29] Pearson, K. (1895) Notes on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, **58**, 240-242. https://doi.org/10.1098/rspl.1895.0041

[30] Friedman, M. (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **32**, 675-701. https://doi.org/10.1080/01621459.1937.10503522

[31] Conover, W.J. (1999) Practical Nonparametric Statistics. 3rd Edition, John Wiley and Sons.

[32] Kendall, M.G. and Gibbons, J.D. (1990) Rank Correlation Methods. 5th Edition, Oxford University Press.

[33] In, J. and Lee, D.K. (2024) Alternatives to the P Value: Connotations of Significance. *Korean Journal of Anesthesiology*, **77**, 316-325. https://doi.org/10.4097/kja.23630

[34] JASP Team (2024) (Version 0.19.2) [Computer Software]. https://jasp-stats.org/download/

[35] Meyners, M. and Hasted, A. (2021) On the Applicability of ANOVA Models for CATA Data. *Food Quality and Preference*, **92**, Article ID: 104219. https://doi.org/10.1016/j.foodqual.2021.104219

[36] Calvetti, D. and Somersalo, E. (2023) Continuous and Discrete Multivariate Distributions. In: Calvetti, D. and Somersalo, E., Eds., *Bayesian Scientific Computing*, Springer International Publishing, 35-48. https://doi.org/10.1007/978-3-031-23824-6_3

[37] Wald, A. and Wolfowitz, J. (1943) An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation. *The Annals of Mathematical Statistics*, **14**, 378-388. https://doi.org/10.1214/aoms/1177731358

[38] Ljung, G.M. and Box, G.E.P. (1978) On a Measure of Lack of Fit in Time Series Models. *Biometrika*, **65**, 297-303. https://doi.org/10.1093/biomet/65.2.297

[39] Henze, N. and Zirkler, B. (1990) A Class of Invariant Consistent Tests for Multivariate Normality. *Communications in Statistics—Theory and Methods*, **19**, 3595-3617. https://doi.org/10.1080/03610929008830400

[40] Korkmaz, S. (2025) Package "MVN". Multivariate Normality Tests. https://cran.r-project.org/web/packages/MVN//MVN.pdf

Appendix. Generation of MVT and MVN Samples

```
# Sample of 50 observations in two variables, with a correlation coefficient of 0.3, drawn from a multivariate t-distribution with five degrees of freedom.
  library(mvtnorm)
  n <- 50 # Number of participants
  k <- 2 # Number of variables
  df <- 5 # Degrees of freedom for MVT distribution
  rho <- 0.3 # Homogeneous correlation
  sigma <- matrix(rho, nrow = k, ncol = k) # Correlation matrix
  diag(sigma) <- 1 # Correlation matrix
  set.seed(123) # Seed is set for reproducibility
  mvt_data_2v <- rmvt(n = n, sigma = sigma, df = df)
  x1 <- mvt_data_2v[,1]
  x2 <- mvt_data_2v[,2]
  original_data <- data.frame(x1, x2) # sample data
  print(original_data)

  # Sample of 50 observations in two variables, with a correlation coefficient of 0.3, drawn from a multivariate standard normal distribution.
  library(mvtnorm)
  n <- 50 # Number of observations
  k <- 2 # Number of variables
  rho <- 0.3 # Homogeneous correlation
  sigma <- matrix(rho, nrow = k, ncol = k) # Correlation matrix
  diag(sigma) <- 1 # Correlation matrix
  set.seed(123) # Seed is set for reproducibility
  mvn_data <- rmvnorm(n = n, sigma = sigma)
  x1 <- mvn_data[,1]
  x2 <- mvn_data[,2]
  original_data <- data.frame(x1, x2)
  print(original_data)
```