Comparing Normality Tests through Monte Carlo Simulation and Real-World Application: Power, Type I Error and Surat Air Quality Index Evidence

Shital S. Patel
Department of Statistics
Veer Narmad South Gujarat University,
Udhna-Magdalla Road, Surat – 395007, Gujarat, India
Email: sspatel@vnsgu.ac.in

Abstract:

Choosing an appropriate check of the Gaussian assumption is a live practical problem, since the merits of competing tests shift with sample size and the shape of the departure. This paper combines a review of four procedures, Shapiro–Wilk, Anderson–Darling, Kolmogorov–Smirnov with the Lilliefors correction, and Jarque–Bera, with two original components. First, a Monte Carlo experiment of five thousand replications evaluates Type I error and power at sample sizes twenty to one hundred against skewed, heavy-tailed, contaminated and platykurtic alternatives. Shapiro–Wilk is most powerful against asymmetry; Jarque–Bera gains slightly against heavy tails but collapses against platykurtic data, and Lilliefors trails throughout. Second, the tests are applied to forty-nine real half-hourly Air Quality Index readings for Surat city; all four reject the Gaussian hypothesis, matching the simulation's predictions near n = 50. Practical selection guidance is given for applied researchers.

1. Introduction:

Few ideas shape day-to-day statistical practice as strongly as the Gaussian model. Whenever an analyst fits an ordinary regression, runs a t-type comparison, or carries out an analysis of variance, the machinery of inference presumes that the raw measurements, or the residuals left after model fitting, behave like draws from a bell-shaped population (Khatun, 2021; Singh & Masuku, 2014). When that presumption fails, the consequences propagate: estimators lose efficiency, stated error rates no longer hold, and forecasts drift away from reality (Das & Imon, 2016). Verifying the distributional condition ahead of the main analysis is therefore a matter of methodological hygiene. Indeed, whether formal normality testing is necessary at all remains debated (Tsagris & Pandis, 2021), which makes the choice of procedure all the more consequential.

A large simulation literature compares the available checking procedures, with Razali and Wah (2011) and Yap and Sim (2011) among the most cited benchmarks, alongside broader Monte Carlo evaluations by Farrell and Rogers-Stewart (2006), Öztuna, Elhan and Tüccar (2006), Romão, Delgado and Costa (2010) and Noughabi and Arghami (2011), and the book-length treatment of Thode (2002). Applied assessments in clinical and veterinary research likewise examine how test performance varies with sample size (Kwak & Park, 2019; Le Boedec, 2016). Yet a gap separates that literature from applied practice in at least two respects. First, published power studies rarely connect their findings back to a concrete applied decision on a genuinely short real-world series, leaving practitioners unsure how the abstract rankings translate into their own settings. Second, environmental monitoring data, whose distributional character drives the choice between parametric and rank-based trend methods, is almost absent

from these benchmarks even though pollution series are notoriously skewed. The present paper addresses both points by pairing a fresh Monte Carlo experiment with an application to air quality measurements from Surat city.

Surat supplies a particularly instructive setting for such an application. It ranks among India's fastest-growing industrial centres, hosting extensive textile and diamond processing alongside heavy vehicular traffic, and its ambient air quality is tracked continuously through public monitoring dashboards. Regulatory summaries, exceedance analyses and trend assessments for such cities routinely employ parametric tools, so the distributional behaviour of the underlying AQI series is not an academic curiosity but a direct determinant of valid environmental inference. The study accordingly proceeds in three stages: a compact review of four workhorse procedures, Shapiro–Wilk (SW), Kolmogorov–Smirnov in its Lilliefors form (KS-L), Anderson–Darling (AD) and Jarque–Bera (JB); an original simulation experiment measuring their Type I error and power across sample sizes and departure shapes; and a fully worked application to a short Surat AQI series, interpreted in the light of the simulation evidence.

2. Materials and Methods

2.1 Statistical basis of the four procedures.

Each procedure evaluates the same pair of competing claims: under H_0 the data originate from some Gaussian population, whereas under H_1 the parent distribution departs from the Gaussian family in an unspecified way. A computed significance probability below the working threshold (here $\alpha = 0.05$) counts as evidence against H_0 (Yang & Berdine, 2021). Methodologically, the four procedures come from three distinct traditions: order-statistic correlation (SW), distance between distribution functions (KS-L and AD), and sample moments (JB).

Shapiro–Wilk.

Shapiro and Wilk (1965) built their statistic W by setting two rival variance estimators against one another, one derived from the ordered observations and their expected Gaussian order statistics, the other from the usual sum of squared deviations. W is confined to the unit interval, and only values well short of unity speak against the Gaussian claim. Its weak spots are heavily tied data and extremely short series.

Kolmogorov–Smirnov with the Lilliefors adjustment.

Building on Kolmogorov (1933), Smirnov (1948) turned the quantity D, the largest vertical gap separating the sample's empirical distribution from the fitted Gaussian curve, into a usable goodness-of-fit criterion. Whenever the mean and dispersion are estimated from the sample,

the textbook critical points overstate the p-value; Lilliefors (1967) supplied corrected reference points, used throughout this paper.

Anderson–Darling.

Anderson and Darling (1952) proposed, and Stephens (1974) operationalised, a squared-distance criterion that gives amplified weight to the tails of the distribution, sharpening sensitivity to tail departures relative to the Kolmogorov–Smirnov statistic.

Jarque–Bera.

Formalised for regression residuals by Jarque and Bera (1987) and traceable to the omnibus construction of Bowman and Shenton (1975), the JB statistic combines squared standardised skewness and excess kurtosis into one quantity, asymptotically chi-square on two degrees of freedom. Its asymptotic basis makes it unreliable in small samples (Hatem et al., 2022).

2.2 Monte Carlo simulation design.

For every combination of sample size $n \in \{20, 30, 50, 100\}$ and parent distribution, 5,000 independent samples were generated by Monte Carlo sampling (Metropolis & Ulam, 1949) and each procedure applied at $\alpha = 0.05$. Five parents were studied: standard Gaussian (Type I error), lognormal (moderate skew), Student's t with three degrees of freedom (heavy tails), a 90:10 contaminated Gaussian mixture (outliers; Huber, 1964; Tukey, 1960), and the uniform distribution (light tails). Lilliefors critical values were obtained by simulation with 20,000 null replications per sample size. All computation used Python 3 (scipy.stats) with a fixed random seed for reproducibility.

2.3 Case study data and procedure.

The working sample comprises forty-nine consecutive half-hourly AQI (US standard) readings for Surat, Gujarat, recorded between 11:01 on 22 April 2026 and 11:01 on 23 April 2026, retrieved from the public monitoring dashboard operated by AQI.in (Purelogic Labs India) (AQI.in, 2026): 78, 72, 80, 82, 77, 77, 77, 89, 83, 84, 78, 82, 80, 87, 92, 90, 94, 96, 102, 95, 88, 89, 85, 83, 87, 76, 83, 80, 89, 80, 78, 77, 83, 87, 91, 92, 99, 110, 113, 125, 126, 117, 108, 96, 95, 77, 83, 129, 84. Descriptive statistics are given in Table 3. Analysis was carried out in Python 3 and reproduced independently in Jamovi.

Table 1. Summary measures for the Surat AQI sample (n = 49)

Measure	Value	Interpretation
Sample size (n)	49	Near the n = 50 simulation grid point
Mean	89.90	Moderate AQI level (US scale)
Median	87.00	Mean above median → right skew

Std. deviation	13.77	Marked intraday variability
Skewness	1.359	Strong positive skew (> 1)
Excess kurtosis	1.194	Heavier-than-Gaussian right tail

3. Results

3.1 Type I error.

Table 1 about here. SW, KS-L and AD hold the nominal level closely at every sample size. JB is markedly undersized in short series, rejecting a true Gaussian null only 2.8% of the time at n = 20 and 4.0% at n = 100, foreshadowing power losses.

3.2 Power.

Table 2 about here. Figure 1 about here. Against skewed data SW leads at every sample size, reaching 52% power at n = 20 where KS-L and JB manage barely 34%. Against heavy tails and contamination JB overtakes SW modestly from n = 30, though never by more than five percentage points. Against the platykurtic uniform, JB fails almost totally (power below 1% up to n = 50), while SW climbs from 21% to nearly 100%. KS-L finishes last against every alternative. The gradient between n = 20 and n = 50 is steep: samples near fifty observations give the tests a realistic chance of detecting departures typical of environmental data, whereas samples of twenty leave them close to blind.

Table 2. Empirical Type I error at nominal α = 0.05 (5,000 replications)

N	SW	KS-L	AD	JB
20	0.053	0.054	0.053	0.028
30	0.047	0.050	0.047	0.031
50	0.047	0.046	0.045	0.035
100	0.046	0.043	0.042	0.040

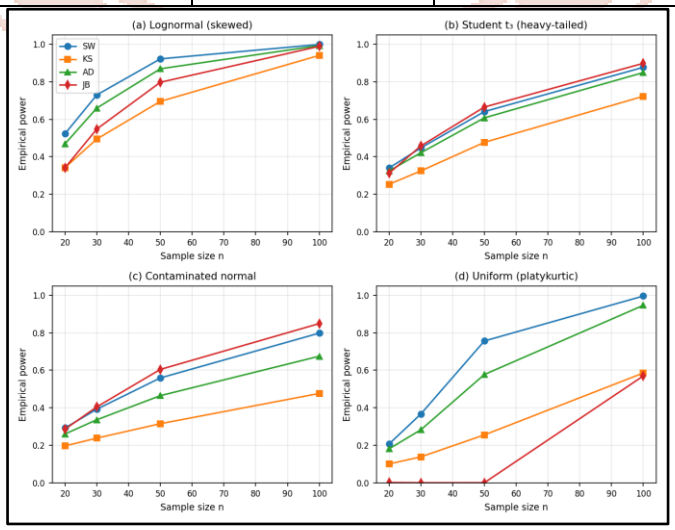


Figure 1. Empirical power of the four procedures across sample sizes and alternatives (5,000 replications).

3.3 Case study outcome.

Table 4 about here. All four procedures reject the Gaussian hypothesis for the Surat AQI sample. The sample sits essentially at the n = 50 grid point, where the simulation recorded power of 0.92 (SW), 0.87 (AD), 0.80 (JB) and 0.70 (KS-L) against a moderately skewed parent, matching the strength ordering observed here. A logarithmic transformation does not restore Gaussianity (log-scale Shapiro–Wilk $p < 0.001$), and a lag-one autocorrelation of 0.66 indicates the half-hourly readings are serially dependent, a caveat for interpretation.

Table 4. Outcomes of the four procedures on the Surat AQI sample (n = 49)

Test	Statistic	p-value	Decision (α = 0.05)	Remark
Shapiro–Wilk	W = 0.8512	< 0.001	Reject H_0	Strongest signal, as the power study predicts
Kolmogorov–Smirnov (Lilliefors)	D = 0.1587	0.003	Reject H_0	D exceeds the simulated 5% point 0.1258
Anderson–Darling	A^2 = 2.3994	< 0.005 (A^2 > 0.740)	Reject H_0	More than three times the 5% critical value
Jarque–Bera	JB = 17.986	< 0.001	Reject H_0	Driven by skewness 1.36 and excess kurtosis 1.19

4. Discussion

The simulation findings align with established benchmarks. Razali and Wah (2011) ranked SW first, AD second, Lilliefors third and uncorrected KS last across a broad sweep of alternatives; the present experiment reproduces that ordering for skewed data and extends it by quantifying JB's shortfall at realistic environmental sample sizes. Yap and Sim (2011) reported that moment-based procedures compete strongly against symmetric heavy-tailed alternatives, consistent with JB's small margin over SW for t_3 and contaminated data from n = 30; they also noted JB's weakness against light-tailed parents, confirmed here in stark form. The advice of Yang and Berdine (2021), that numerical and graphical evidence should be weighed together, gains concrete support: the four rejections, their ordering by strength, and the failed logarithmic remedy together yielded a richer diagnosis than any single statistic.

Echoing Öztuna et al. (2006) and Mendes and Pala (2003), JB's undersizing in short series means its retentions carry less evidential weight than the nominal level suggests. The case study and simulation jointly bracket the practical problem: at twenty observations the tests operate near blindness, while at forty-nine real observations they detect a genuinely skewed series

unanimously. Intraday environmental series also carry serial dependence that formal checks do not accommodate, a limitation any honest application should disclose.

Practical recommendations. With fewer than about thirty observations, treat any retention of the Gaussian hypothesis as provisional. Adopt Shapiro–Wilk as the default check, particularly for skewed data such as pollution or income measurements. Where heavy tails are the specific concern and n exceeds thirty, Jarque–Bera adds modest power but should not be trusted against light-tailed data or in short series. Avoid the Kolmogorov–Smirnov family unless required, and then only in the Lilliefors form. Always pair a formal test with a quantile-quantile plot.

This paper has combined a review, an original Monte Carlo experiment, and an environmental application to show how four normality tests behave in the short, imperfect series applied researchers actually encounter. The experiment confirmed Shapiro–Wilk's leadership against asymmetric departures and exposed Jarque–Bera's finite-sample undersizing and severe light-tail blindness. The Surat AQI application showed forty-nine real half-hourly readings, right-skewed by a morning pollution episode, rejected unanimously by all four tests in exactly the order the power study predicts. Future work should extend to daily-resolution series spanning full seasonal cycles and to dependence-robust checking procedures.

References

- Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. The Annals of Mathematical Statistics, 23(2), 193–212.
- AQI.in. (2026). Surat air quality index (AQI): Real-time air pollution dashboard. Purelogic Labs India. Retrieved July 17, 2026, from https://www.aqi.in/in/dashboard/india/gujarat/surat
- Bowman, K. O., & Shenton, L. R. (1975). Omnibus test contours for departures from normality based on √b1 and b2. Biometrika, 62(2), 243–250.
- Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. American Journal of Theoretical and Applied Statistics, 5(1), 5–12.
- Farrell, P. J., & Rogers-Stewart, K. (2006). Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. Journal of Statistical Computation and Simulation, 76(9), 803–816.
- Hatem, G., Zeidan, J., Goossens, M., & Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. BAU Journal - Science and Technology, 3(2), Article 7.

- Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), 163–172.
- Khatun, N. (2021). Applications of normality test in statistical analysis. Open Journal of Statistics, 11, 113–122.
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4, 83–91.
- Kwak, S. G., & Park, S.-H. (2019). Normality test in clinical research. Journal of Rheumatic Diseases, 26(1), 5–11.
- Le Boedec, K. (2016). Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size. Veterinary Clinical Pathology, 45(4), 648–656.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318), 399–402.
- Mendes, M., & Pala, A. (2003). Type I error rate and power of three normality tests. Pakistan Journal of Information and Technology, 2(2), 135–139.
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. Journal of the American Statistical Association, 44(247), 335–341.
- Noughabi, H. A., & Arghami, N. R. (2011). Monte Carlo comparison of seven normality tests. Journal of Statistical Computation and Simulation, 81(8), 965–972.
- Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turkish Journal of Medical Sciences, 36(3), 171–176.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21–33.
- Romão, X., Delgado, R., & Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. Journal of Statistical Computation and Simulation, 80(5), 545–591.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611.
- Singh, A. S., & Masuku, M. B. (2014). Assumption and testing of normality for statistical analysis. American Journal of Mathematics and Mathematical Sciences, 3(1), 169–175.

- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 19(2), 279–281.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347), 730–737.
- Thode, H. C. (2002). Testing for normality. New York, NY: Marcel Dekker.
- Tsagris, M., & Pandis, N. (2021). Normality test: Is it really necessary? American Journal of Orthodontics and Dentofacial Orthopedics, 159(4), 548–549.
- Yang, S., & Berdine, G. (2021). Normality tests. The Southwest Respiratory and Critical Care Chronicles, 9(37), 87–90.
- Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), 2141–2155.