ST open

EDITORIAL

The myth of normality testing in biomedical research

Darko Kero

University of Split School of Medicine, Graduate Study of Dental Medicine, Split, Croatia

Correspondence to:
Darko Kero, University of Split School of Medicine, Graduate Study of Dental Medicine, 21000 Split, Croatia
dkero@mefst.hr

Cite as:
Kero D. The myth of normality testing in biomedical research. ST-OPEN. 2026;7:e2026.2519.1.

DOI:
https://doi.org/10.48188/so.7.2

Testing data for normality before applying parametric statistics has become a routine procedure in biomedical research. This commentary argues that such tests provide little inferential value, may mislead analytical choices, and reflect outdated thinking from a pre-computational era. Parametric procedures are robust to modest departures from normality, and the Central Limit Theorem makes most normality checks unnecessary. According to this Theorem, the sampling distribution of the mean approaches a normal shape as sample size increases, regardless of the original distribution of the data. Outlier panic and indiscriminate data ranking further undermine the meaning of measurement and prediction. The obsession with distributional purity has replaced the logic of inference with statistical rigor. It is time to abandon this ritual and refocus on design, representativeness, and modeling – the true pillars of inference.

Keywords: biomedical statistics; Kolmogorov–Smirnov; nonparametric inference; normality testing; parametric inference; Shapiro–Wilk

The problem

These familiar lines appear in the “Statistical analysis” subsections of countless papers: “Data were tested for normality using the Shapiro–Wilk test” or “Kolmogorov–Smirnov test was applied to test the normality of data distribution.” They sound scientific, but provide no useful information for choosing or justifying inferential methods. Regardless of whether a test is significant, the underlying research question remains unchanged, even though the choice of statistical test may lead to different formal decisions. Normality testing has become a statistical reflex – performed automatically, rarely justified, but widely misunderstood. Several questions instantly come to mind: do we not know that parametric procedures such as the t-test and ANOVA do not require the observed raw data to be perfectly normal (*1*-*3*)? Have we forgotten what the Central Limit Theorem tells us about the nor-

mality of the distribution of sample means (*4*)? Why do we choose to assume nothing when the Central Limit Theorem supposes underlying normality in most practical datasets (*5*, *6*)?

While the familiar lines about Shapiro–Wilk and Kolmogorov–Smirnov testing continue to appear as the proverbial *Lorem ipsum* in the ever-growing body of biomedical research papers, we must also ask how well these tests serve their stated purpose. The truth is that when samples are small, normality tests lack power; when samples are large, they detect trivial deviations (*7*). Normality tests lack power in small samples and become overly sensitive in large samples, detecting trivial deviations with little practical relevance. Testing whether a small sample comes from a "normal population" assumes that the population's true distribution is known (*8*), which is rarely the case. In practice, testing conformity to an unknown population distribution shape is a circular exercise – it answers a question no one can verify. Furthermore, in statistical modeling – where distributional assumptions are often invoked – what matters is the behavior of model residuals or errors, not the supposed normality of raw data (*9*, *10*). This reliance on normality reflects an epistemological confusion between the geometry of data and the logic of inference. We do not generalize from shapes; we generalize from sampling principles. Yet, many researchers in the biomedical field continue to prioritize distributional form over sound inferential reasoning.

Reconsidering representativeness

A common confusion occurs when normality is mistaken for representativeness. In practice, researchers and reviewers often equate representativeness with sample size, as if a larger sample guarantees inferential validity (*11*). It does not. A large, but biased sample can still misrepresent the population, while a small, but well-drawn sample can reflect it accurately (*12*). Representativeness concerns how well the sample statistics (means, medians, variances) mirror the population parameters they are intended to estimate – not how bell-shaped the histogram appears. A perfectly normal sample can be completely unrepresentative, while a skewed sample can be highly representative. The shape of a sample distribution tells us nothing about how faithfully it reflects the underlying population (*11*). Normality testing, therefore, provides no information about what really matters: whether our inferences from sample to population are justified – a distinction that is often blurred in routine analytical practice.

How does the Central Limit Theorem deal with non-normal data?

To illustrate this principle, consider a simulated population of 10,000 observations of a discrete variable, "Clinical Score" (range 0–50), designed to follow a normal distribution (mean "Clinical Score" = 25.08, standard deviation (SD) = 5.01) (**Figure 1**). Supplementary dataset is available in Open Science Framework (*13*). From this population, 50 random samples of 30 observations each were drawn. All failed the Shapiro–Wilk and Kolmogorov–Smirnov normality tests (here shown side by side to illustrate the potential for inconsistent signals from commonly used normality tests), even though the underlying population was normal. However, the distributions of sample means and standard deviations were

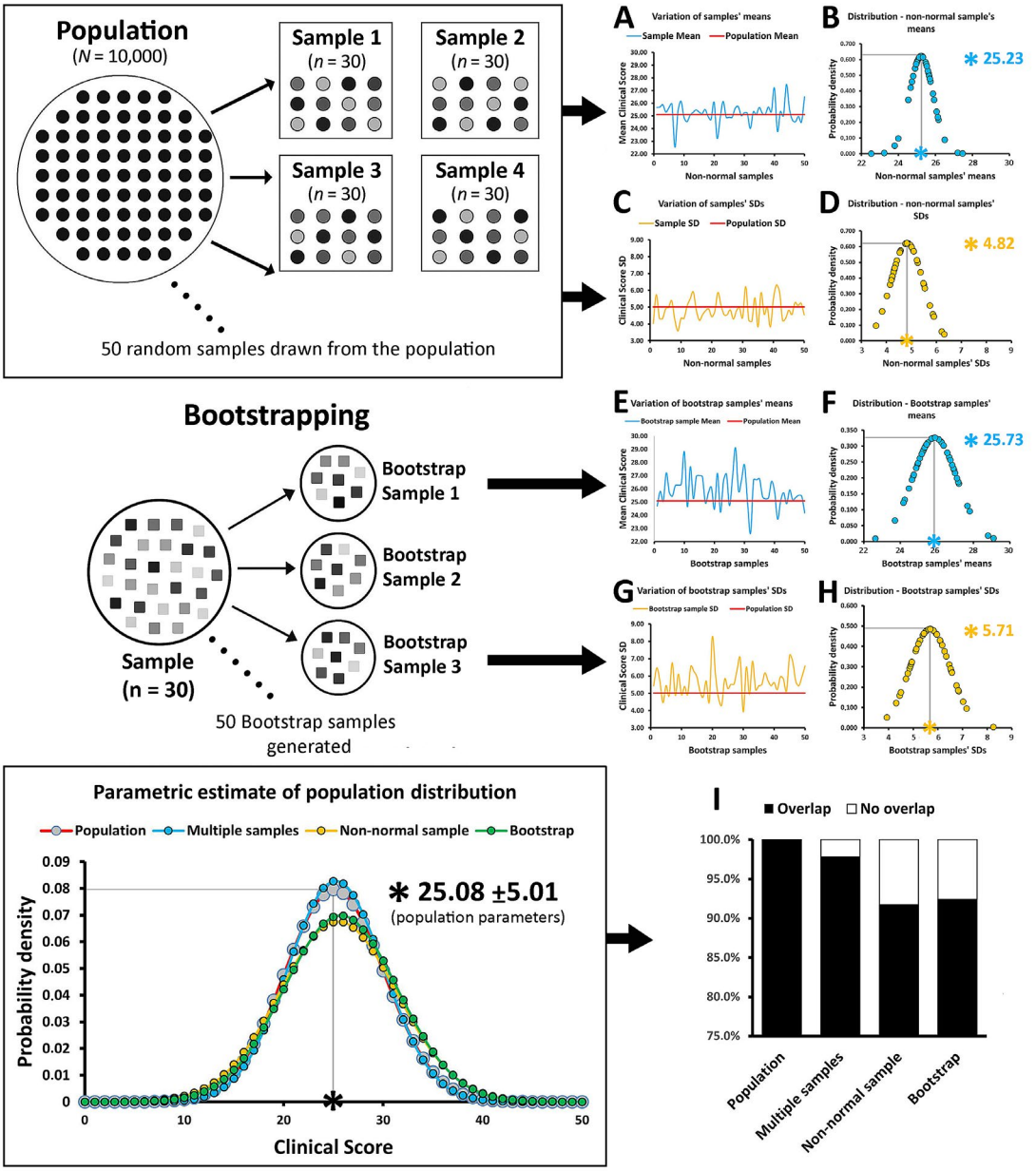

Figure 1. Simulated "Clinical Score" population and sample-derived estimates demonstrating robustness of parametric inference despite non-normal samples. **Panels A–D.** Random sampling from the population: Panels A and C show variability of sample means and SDs, while Panels B and D show their normal distributions. **Panels E–H.** Parametric bootstrap from one non-normal sample: Panels E and G show variability of means and SDs, while Panels F and H show their normal distributions. **Panel I.** Overlap (>90%) between inferred and true population distributions obtained by three inferential approaches.

approximately normal, consistent with the Central Limit Theorem. Population parameters were inferred in three complementary ways: first, from multiple random samples whose data were not normally distributed (mean "Clinical Score" = 25.23, SD = 4.82); second, by applying a parametric bootstrap to one of those non-normal samples (mean "Clinical Score" = 25.47, SD = 5.90, Kolmogorov–Smirnov $P = 0.048$, Shapiro–Wilk $P = 0.004$); and third, directly from that same single sample using only its mean and standard deviation (mean "Clinical Score" = 25.47, SD = 5.90). The bootstrap example is shown in greater detail to emphasize that, even when both the original sample and the resampled distribution fail normality tests, parametric inference remains valid. All three approaches yielded near-

ly identical estimates of the true population parameters, with the inferred distributions overlapping the original population by more than 90%. This simple exercise demonstrates that parametric inference remains robust even when individual samples deviate from normality (*1*-*3*). What matters for inference is not the shape of the data, but the validity of sampling and estimation.

When normality misleads representativeness

Imagine drawing a random sample of blood pressures from a population of patients. Your histogram might appear skewed simply because hypertensive individuals are more common – yet your sample could accurately represent the true population structure. Conversely, if you sample only healthy volunteers, your data might look textbook-normal, but be useless for clinical generalization. In both cases, a test of normality would mislead rather than inform. Representativeness is an empirical property of how data are collected, not how they appear. It depends on study design, randomness, and measurement, not on statistical appearance. Confusing normality with representativeness (which commonly arises in routine analytical practice) is to mistake aesthetics for science – and, in doing so, to take the first step toward another statistical refuge: the comforting, but often costly embrace of non-parametric tests.

The non-parametric refuge

Faith in normality tests often leads researchers to use non-parametric methods "just to be safe." However, this "safety" is an illusion. Non-parametric tests, such as Mann–Whitney or Kruskal–Wallis, are not stricter or more robust in any meaningful inferential sense – they are simply less powerful when parametric tests are inferentially appropriate, even in the presence of moderate deviations from normality (*14*, *15*). Treating them as inherently "safer" is like saying a thermometer without numbers is safer because it never overheats. The apparent simplicity of non-parametric methods conceals their cost. They trade information for convenience. By ranking or dichotomizing data, they suppress the very features that may matter most. A small deviation from normality becomes an excuse to abandon decades of mathematical insight in favor of mechanical ritual.

When "safety" costs information – a tale of two skewed distributions

To illustrate how misplaced caution can obscure genuine effects, consider two highly skewed populations of a discrete variable, "Clinical Score" (range 0–50). Although both have the same median value (*10*), their means differ markedly (14.85 *vs*. 17.87) – a difference of about three points that, in a clinical context, would alter patient management. From each population, 1,000 random pairs of small samples (n = 30) were drawn and compared using Welch's t-test and the Mann–Whitney U-test (**Figure 2**). Supplementary dataset is available in Open Science Framework (*13*). Despite over 99% of samples failing the Shapiro–Wilk normality test and 22% failing the Kolmogorov–Smirnov test, both

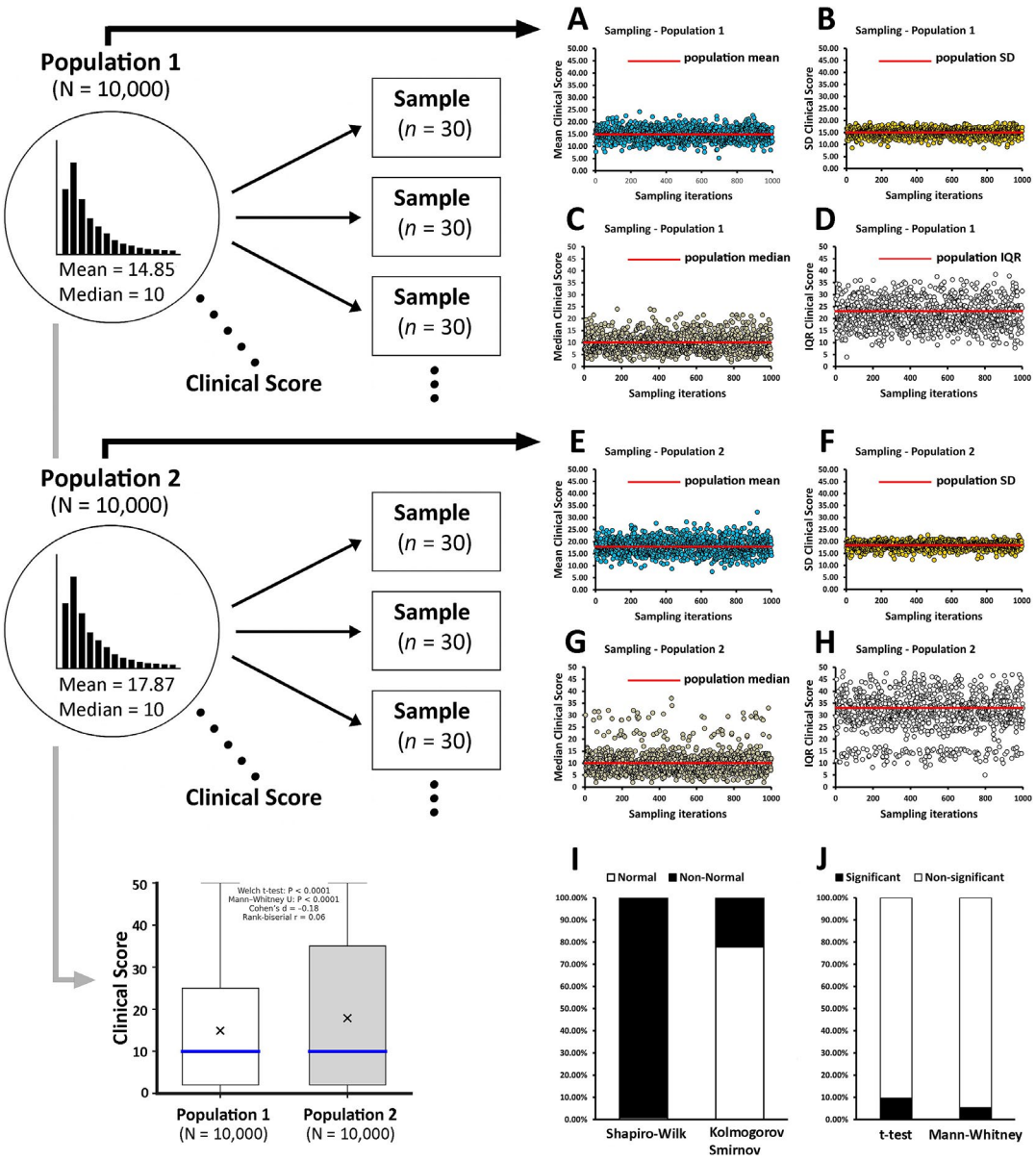

Figure 2. Two highly skewed simulated populations of the discrete variable Clinical Score (range 0–50), both exhibiting pronounced positive skewness. **Panels A–D.** The variability of sample means, medians, SDs, and IQRs from 1,000 random samples drawn from Population 1. **Panels E–H.** The corresponding variability from samples drawn from Population 2. The overall difference between populations is illustrated by the box-and-whisker plot (bottom left). **Panel I.** The results of the Shapiro–Wilk and Kolmogorov–Smirnov normality tests. **Panel J.** The comparative statistical power of Welch's t-test and the Mann–Whitney U-test for detecting the true difference between populations.

procedures performed similarly. In fact, the t-test was slightly more sensitive in detecting the true population difference in means, even under severe non-normality. This finding is hardly accidental. While the Mann–Whitney test merely notes that ranks differ, the t-test engages with magnitude – the very quantity that gives biomedical results their meaning. In practice, the pursuit of "distributional safety" through rank-based methods can erase clinically relevant signals, leaving researchers statistically correct, but scientifically blind. The exercise demonstrates a simple truth: robustness lies not in denying assumptions, but in reasoning through them. Parametric inference, grounded in the Central Limit Theorem, remains robust even when non-parametric alternatives are often selected by convention rather than by the inferential target.

The logic of absurdity

When researchers turn to non-parametric statistics due to misplaced caution, they sacrifice interpretive power. Parametric analyses provide effect sizes, confidence intervals, and model-based insights that are directly interpretable on the measurement scale, whereas non-parametric tests typically offer rank-based differences, often at the cost of precision and interpretability. Biomedical science seeks to quantify effects, not just detect them. This leads to a logical absurdity: if non-parametric tests require no assumption of normality, and if, for the sake of argument, they are indeed a "safer" option than parametric tests, then what is the purpose of testing for normality at all? If one plans to use a method that ignores the distribution's shape regardless of the test result, the normality test adds no information – though reporting that it was performed can still serve as a comforting opening line in the "Statistical analysis" section of the "Methods".

The outlier paradox

Another issue underlying this statistical ritual is the obsession with outliers, which are often seen as threats to parametric purity – anomalies to be controlled or removed. However, outliers are not always errors; in fact, they can be the most interesting observations in the dataset. In clinical and biological contexts, extreme values may reflect genuine biological variation – the "super-responders", resistant phenotypes, or clinical exceptions that drive discovery (*16*).

Non-parametric approaches handle outliers by converting all data to ranks. Once data are ranked, an outlier is placed just one step above its neighbor, regardless of the actual distance between their values. Although this method is mathematically convenient, it commits an epistemic error: it eliminates information about magnitude. Rank transformation is irreversible; after applying it, the original quantitative meaning of the data cannot be recovered. It is like replacing a high-resolution image with a blurred sketch – we can still distinguish between two shapes, but we no longer know by how much they differ. Spearman's correlation exemplifies this issue. It can show that two variables move together, but not how much change in one predicts change in the other. It substitutes measurement with monotony, reducing science to detecting rhythm without melody. Pearson's correlation, though susceptible to outliers, at least retains the essential scientific property of quantitative prediction (*17*).

Flattening the curve – how rank transformation masks structure

Outliers are often blamed for distorting correlations, but removing them – or reducing their influence through rank transformation – can conceal more than it reveals. In a simulated dataset of 30 participants linking "Age" (18–65 years) with "Clinical Score" (0–50) (**Figure 3**). Supplementary dataset is available in Open Science Framework (*13*). Several outliers made the relationship appear irregular. A cautious analyst might replace raw values with ranks and apply Spearman's correlation (nonparametric). The result ($\rho = 0.60$)

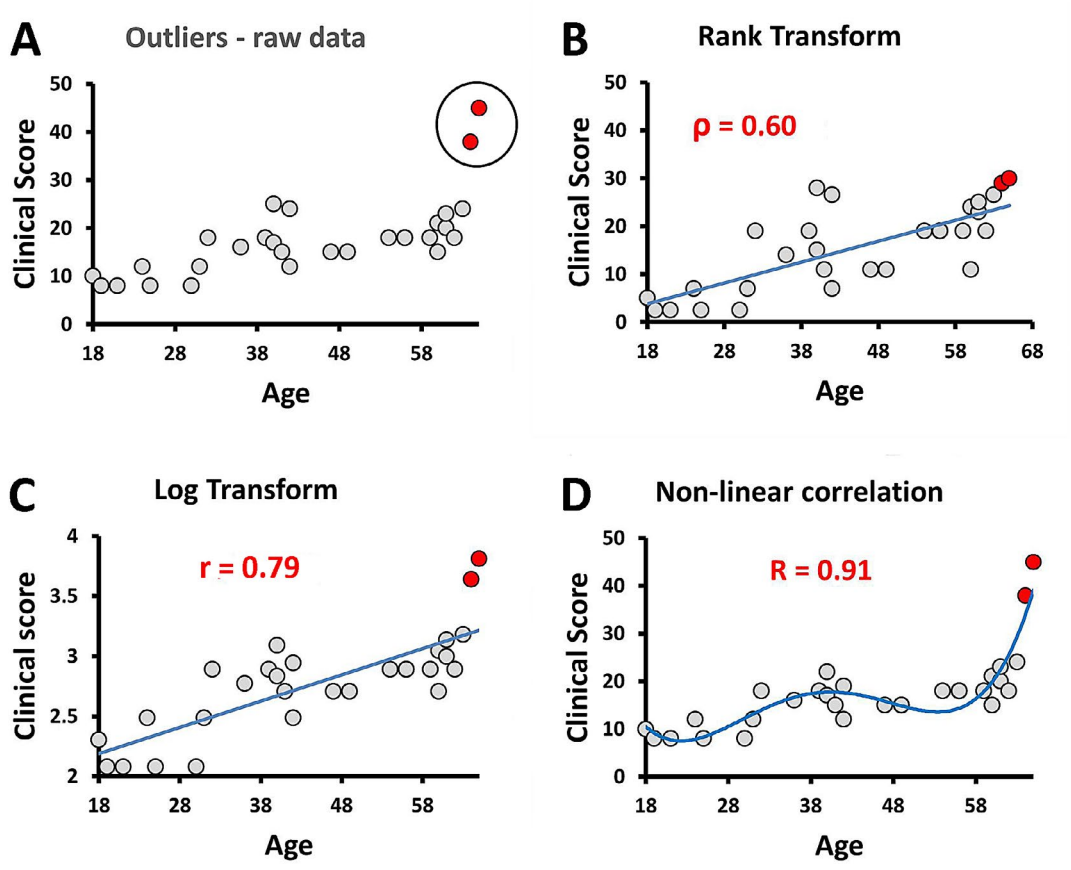

Figure 3. The effects of different data transformations on the assessment of relationship between Age and Clinical Score (n = 30). **Panel A.** Raw data showing outliers (red dots). **Panel B.** Rank transformation (Spearman's correlation) linearizes but flattens the true curvature. **Panel C.** Log transformation (Pearson's correlation) reduces outlier influence, while preserving scales. Nonlinear regression best captures the accelerating rise in "Clinical Score" with increasing "Age".

appeared reassuringly linear, but only because ranking flattened curvature and erased scale. A reversible transformation tells a different story. Logarithmic scaling reduces the visual weight of outliers, while preserving quantitative meaning. Pearson's correlation (parametric) on log-transformed data yielded a stronger link (r = 0.79), suggesting a nonlinear trend. Fitting a third-order polynomial to the untransformed data (parametric nonlinear model) achieved the best fit (R = 0.91), capturing the accelerating rise in "Clinical Score" with "Age". Relying on ranks offers safety at the cost of understanding – flattening complexity into comforting simplicity. This example shows how rank transformation, often used after failed normality checks, can create an illusion of order, while concealing structure. In seeking statistical safety, it trades complexity for comfort – and in doing so, obscures precisely what a researcher aims to understand.

Outliers should be understood, not simply removed. If an observation truly reflects measurement error, the issue is methodological, not mathematical. However, if it reflects genuine variability, it belongs in the model, not relegated to a rank table. Robust parametric regression models offer ways to accommodate such values without distorting the results of the analysis (*15*, *18*).

What robustness really means

Robustness, broadly defined, refers to the ability of statistical procedures to provide reliable and stable results, even when the underlying assumptions are moderately violated (*18*). Robust statistics were never intended to eliminate parametric reasoning, but to safeguard it (*18*). As Box famously said, "All models are wrong, but some are useful" (*6*). Robustness involves designing models that withstand modest violations of assumptions without failing, not abandoning quantitative inference entirely. Modern computational methods, such as bootstrapping (*18*) and Bayesian modeling (*19*), already handle non-normality and outliers effectively, without resorting to the intellectual self-harm of rank transformation. These methods preserve the measurement scale, while adjusting inferential uncertainty. They represent a mature response to imperfection, not an escape from it.

The preoccupation with testing normality reflects a desire for control, rather than understanding. It gives researchers the illusion of certainty: if the test passes, all is well; if it fails, they have a procedural excuse. However, control without deep understanding is bureaucracy, not science.

The way forward

The continued use of normality testing stems from habit, rather than necessity. It is a relic from the era of manual calculations and fragile assumptions. Modern diagnostics and simulation studies have repeatedly demonstrated the robustness of parametric inference (*7*, *14*, *15*). If your design is sound, your sample is reasonable, and your residuals behave appropriately (meaning when they are as as they should be in well-conducted studies), stop bothering Shapiro and Wilk and let them enjoy their drinks with Kolmogorov and Smirnov in peace.

When the design is sound, sampling is unbiased, and residuals are reasonable, the analysis is valid (**Figure 4**). Representativeness depends on how the data are collected, not on their distribution (*9*). Invest effort in better experiments, cleaner measurements, and transparent modeling. Handle outliers thoughtfully, not fearfully.

Embrace parametric reasoning for what it truly is – not a fragile mathematical fantasy, but a robust, information-rich framework for inference. Teach researchers that assumptions are guides, not constraints, and that robustness lies in reasoning, not rituals. However, non-parametric approaches, when applied with understanding rather than caution, remain valuable tools. They serve a legitimate purpose when measurement is imprecise, when variables are ordinal or subjective, or when the median and interquartile range better represent population characteristics than the mean and standard deviation (*20*).

The goal is balance: use parametric reasoning when measurement justifies it and non-parametric reasoning when meaning requires it. The aim is not to prove or disprove that the data are normal, but to show that the reasoning based on the data is sound. Statistical significance without understanding is merely noise labeled with *P*-values. Biomedical research can and must do better.

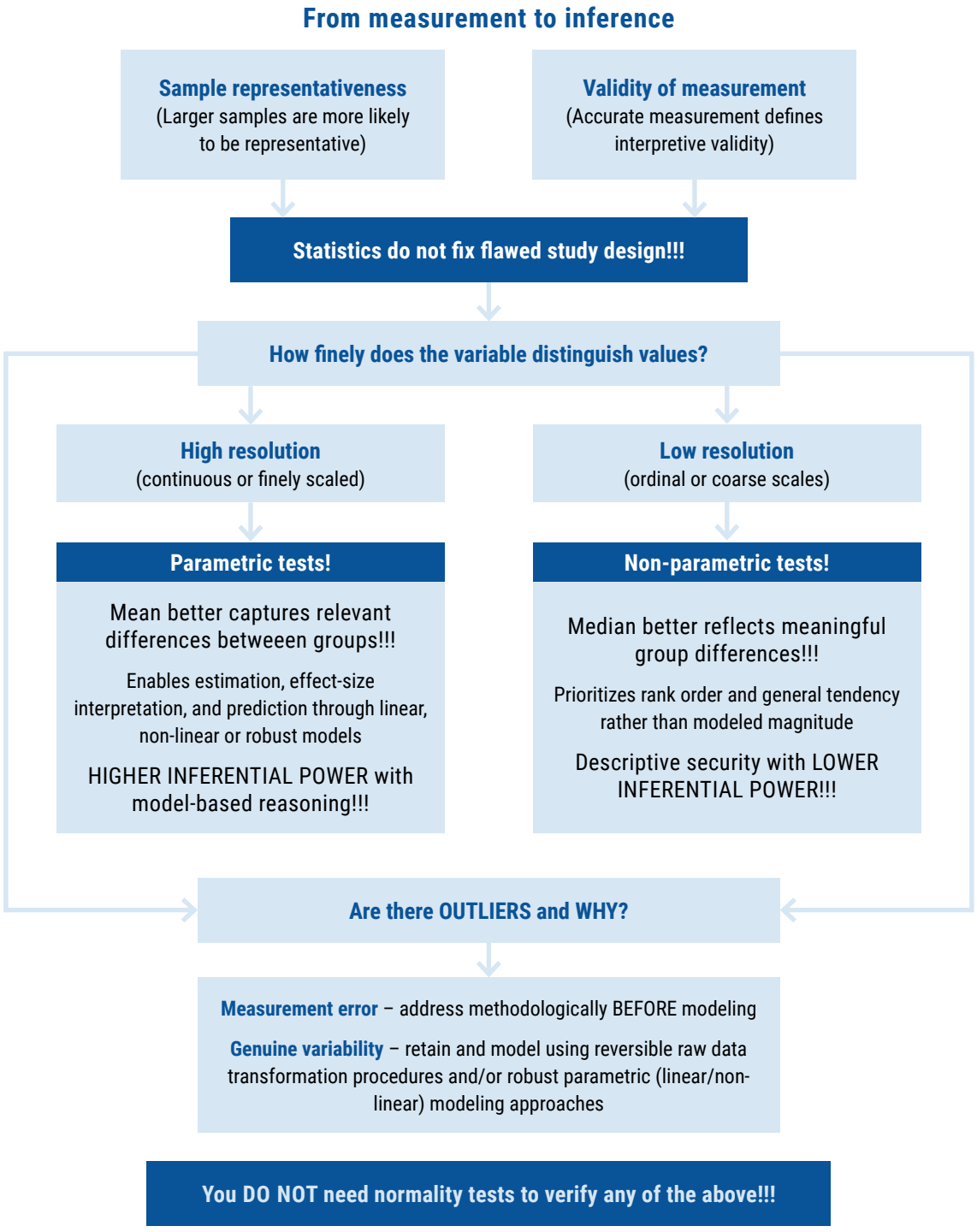

Figure 4. Conceptual flowchart outlining key principles for selecting analytical approaches – parametric and non-parametric approaches form a continuum guided by measurement resolution, the sensitivity of the central tendency measures (mean *vs.* median), and the treatment of outliers.

Provenance: Internally commissioned.

Peer review: Externally peer reviewed.

Received: 10 November 2025 / **Accepted:** 5 February 2026 / **Published:** 18 February 2026.

Availability of data: The raw data for this study are available in Open Science Framework (*13*).

Funding: No funding was received for this study.

Authorship declaration: DK is the sole author of this study.

Disclosure of interest: The author has completed the ICMJE disclosure of interest form (available upon request) and declares the following activities and/or relationships: DK volunteers as Co-editor in Chief of ST-OPEN. To ensure the integrity of the review process, the article has been reviewed according to the guidelines and procedures recommended by the Committee on Publication Ethics.

ORCID

Darko Kero https://orcid.org/0000-0002-8091-6347

References

1. Kim TK, Park JH. More about the basic assumptions of t-test: normality and sample size. Korean J Anesthesiol. 2019;72:331. PubMed
2. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology (Gött). 2010;6(4):147–51. https://doi.org/10.1027/1614-2241/a000016
3. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10(2):486–9. PubMed https://doi.org/10.5812/ijem.3505
4. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol. 2017;70(2):144–56. PubMed https://doi.org/10.4097/kjae.2017.70.2.144
5. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591
6. Student. The probable error of a mean. Biometrika. 1908;6(1):1–25. https://doi.org/10.2307/2331554
7. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151–69. PubMed https://doi.org/10.1146/annurev.publhealth.23.100901.140546
8. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, editors. Robustness in Statistics. New York: Academic Press; 1979. p. 201–236.
9. Osborne JW, Waters E. Four assumptions of multiple regression that researchers should always test. Pract Assess, Res Eval. 2002;8(2):2.
10. Cohen J. Multiple regression as a general data-analytic system. Psychol Bull. 1968;70(6):426–43. https://doi.org/10.1037/h0026714
11. Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ. 1995;310(6975):298. PubMed https://doi.org/10.1136/bmj.310.6975.298
12. Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley & Sons; 1977.
13. Kero D. Supplementary dataset. 2025 Nov 10 [cited 2026 Feb 9]. Available from: https://doi.org/10.17605/OSF.IO/4ZP2N
14. Norman G. Likert scales, levels of measurement and the “laws” of statistics. Adv Health Sci Educ Theory Pract. 2010;15(5):625–32. PubMed https://doi.org/10.1007/s10459-010-9222-y

15. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. 3rd ed. San Diego: Academic Press; 2012.

16. Tukey JW. Exploratory Data Analysis. Reading (MA): Addison-Wesley; 1977.

17. Field A. Discovering Statistics Using IBM SPSS Statistics. 5th ed. London (UK): SAGE Publications; 2018.

18. Huber PJ, Ronchetti EM. Robust Statistics. 2nd ed. Hoboken (NJ): Wiley; 2009.

19. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. Boca Raton (FL): CRC Press; 2013.

20. Yang HJ. Common statistical methods used in medical research. Kosin Med J. 2025;40(1):1–10. https://doi.org/10.7180/kmj.24.160