

A Distribution-Based Framework for Homogeneity Assessment of Engineering Data: Comparison with Conventional Statistical Inference

Yu Yun

Kunming Medical University

Weigang Duan

deardwg@126.com

Yunnan University of Traditional Chinese Medicine

Article

Keywords: engineering data analysis, interval-shift density estimation, overlap probability, homogeneity judgment, distribution recovery

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11099264/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

Tables are available in the Supplementary Files section.

1 **A Distribution-Based Framework for**
2 **Homogeneity Assessment of Engineering Data:**
3 **Comparison with Conventional Statistical**
4 **Inference**

5

6 Yu Yun¹, Weigang Duan²

7 1 Kunming Medical University

8 2 Yunnan University of Chinese Medicine

9

10 Corresponding Author:

11 Weigang Duan

12 1076, Yuhua Rd., Kunming, Yunnan Province 650500, China

13 Email address:deardwg@126.com

14

Abstract

Data generated in experimental science and engineering practice typically originate from measurement, sensing, and production systems rather than from random sampling. Such data possess a clear production attribute: they are physically bounded, may contain discrete and repeated values, and their distributional shapes are not determined by any presupposed model. As data volume increases, statistical inference methods based on random sampling assumptions may exhibit non-convergent behavior. This paper proposes an analytical framework for engineering data that treats data as outputs of a production system, directly recovers their distribution, and assesses homogeneity between datasets on this basis. The method employs Interval-Shift Density Estimation (ISDE) to recover probability densities, and uses distribution overlap probability and median distinguishability as decision criteria, without relying on distributional assumptions, kernel functions, or data transformations. Sample size is used only to assess the reliability of results and does not alter the direction of convergence. Controlled experiments using randomly generated data show that: for homogeneous data generated from identical parameters, the ISDE overlap probability converges stably to 1 as sample size increases; for heterogeneous data generated from different parameters, overlap probability converges to a non-zero value reflecting the degree of difference; and when a small number of outliers are introduced, the overlap probability and median-based judgment remain stable. The method is implemented in Excel 2019 and demonstrates good reproducibility and engineering applicability. We conclude that for engineering data originating from measurement, sensing, and production systems, distribution-based homogeneity assessment provides more stable analytical results than conventional statistical inference.

Keywords: engineering data analysis; interval-shift density estimation; overlap probability; homogeneity judgment; distribution recovery

Introduction

In experimental science and engineering practice, data typically originate from measurement, sensing, experimental records, or the output of production systems, rather than from random sampling of a “population.” Such data possess a clear production attribute: they are physically bounded (e.g., temperature has a lower limit, concentration cannot be negative), may contain discrete values and numerous repeated values due to limited measurement precision, and their distributional shapes are not determined by any presupposed model.

Conventional statistical inference methods are generally built on the premise of random sampling, inferring population parameters from samples and using P-values as the criterion for significance. This framework has a solid theoretical foundation in settings that satisfy the random sampling assumption. However, when data originate from measurement or production systems and the data volume is large

^{1,2}, the criterion may exhibit non-convergent behavior: two datasets generated from the same random process may still be judged as “statistically different.”

Sources of Bias in Traditional Statistical Analysis with Large Datasets

The Dependence of P-Value Drift on Distributional Assumptions Traditional statistical methods typically require the specification of a distributional model, with the normal distribution being the most commonly used. In normality testing, a P-value greater than 0.05 is taken as evidence that the data “follow a normal distribution”—with no distinction made between, say, $P=0.051$ and $P=0.999$. Clearly, a normal fit with $P=0.051$ carries a larger approximation error, yet this error is rarely re-evaluated in subsequent statistical calculations.

Moreover, when a dataset satisfies the conditions for multiple distributional models, different choices of model can lead to different P-values, complicating model interpretation and making the selection of a “correct” P-value arbitrary.

The Dependence of P-Value Drift on Sample Size Traditional significance testing depends on test statistics and their degrees of freedom, which are functions of sample size. In the case of the t-test, the t-statistic (Eq. 1) is determined by the mean difference, the standard deviation, and the sample size. When the two samples are of equal size, the t-statistic can be expressed in a simplified form (Eq. 2). It follows that even a very small mean difference can yield a large t-value—and thus a small P-value—provided the sample size is sufficiently large. This implies that, even when the underlying distribution remains unchanged, increasing sample size can alter both the P-value and the resulting conclusion. This property also constitutes the theoretical basis for P-value manipulation³⁻⁵.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (\text{Eq. 1})$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2 + S_2^2}{n}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2 + S_2^2}} \sqrt{n} \quad (\text{Eq. 2})$$

The Dependence of P-Value Drift on Outliers Traditional statistical measures involve variance or standard deviation (SD), which tend to amplify the influence of outliers⁶. A single outlier can significantly alter the test statistic and consequently change the P-value. In fact, the SD is a fully adequate measure of dispersion only under strict normality. Since most real-world data do not conform exactly to a normal distribution, the error introduced by this mismatch is further magnified.

Materials and Methods

Engineering Data Analysis: Concepts and Computational Principles

EDA treats data as products of a production system, rather than as samples drawn from a population. Accordingly, this framework does not employ concepts

such as "population", "sample", or "sampling". It examines the quantity of data (n), central tendency, dispersion, and the recovered distribution shape, and on this basis compares the similarity between two datasets. Similarity is assessed primarily through the overlap area between distributions and the distinguishability of their shapes.

Basic Parameters in Engineering Data Analysis Data quantity (n): the number of data points produced by the system, which reflects the stability of the distribution.

Central tendency: equivalent in meaning to its statistical counterpart, but restricted to the median—i.e., the value at cumulative probability 0.5 after sorting (the middle value for odd n , and the average of the two middle values for even n).

Dispersion: equivalent in meaning to its statistical counterpart, measured by the interquartile range (IQR).

Based on these parameters, the probability density distribution of the data can be recovered. This study recommends the interval-shift density estimation (ISDE) method for density estimation. ISDE is essentially an adaptive histogram: the window width is determined by local data density—narrowing in dense regions to preserve structural detail, and widening in sparse regions to avoid spurious fluctuations. The window size k is set to the largest odd integer not exceeding $n'/4$ (where n is the total data count and n' is the number of distinct values). Since k is approximately half the IQR and close to the median absolute deviation, it ensures good stability; being odd, it also provides inherent symmetry without requiring additional mathematical extrapolation. Compared with the difference method, the global-domain (n') ISDE method, and the half-domain ($n'/2$) ISDE method, Quarter-domain ($n'/4$) ISDE yields density curves that most closely resemble the theoretical normal distribution (Fig. 1).

Figure 1. Comparison of density estimates obtained by different methods ($n=10,000$). A, Difference method: the curve shows notable jaggedness and fluctuation. B, global-domain (n') ISDE: the curve is overly flattened. C, Half-domain ($n'/2$) ISDE: the curve is slightly flattened, but approximates the reference distribution. D, Quarter-domain ($n'/4$) ISDE: the curve is smooth and faithfully recovers the original distribution. E, Theoretical density curve of the standard normal distribution, shown as a reference.

Principles of Homogeneity Judgment The homogeneity of two datasets is assessed by the following two criteria:

1) Overlap area ($P_{overlap}$)

The overlap area⁷ is the total of the smaller values at each point where the two density curves intersect (Fig. 2A). Based on a common empirical standard in engineering practice, if the $P_{overlap} \geq 0.95$, the two datasets are considered to have a high degree of homogeneity.

2) Distinguishability

The two distributions are placed in the same coordinate system to determine whether the density at each median is masked by the other distribution. The criterion is as follows: for two datasets A and B, if both $f_A(\text{median}A) - f_B(\text{median}A) > 0$ and $f_B(\text{median}B) - f_A(\text{median}B) > 0$ hold, the two medians are distinguishable, indicating a difference between the distributions (Fig. 2E). Conversely, if either median density

is masked by the other distribution (i.e., equal to or lower than the other's density at that point), the two distributions are considered not yet distinguishable (Fig. 2B–D).

Figure 2. Schematic illustration of distribution difference assessment between two datasets. A, Overlap area (P_{overlap}) between two distributions (shaded region). B–E, Four typical scenarios of distribution overlap: B, small overlap area with close medians—the two datasets are not homogeneous but are difficult to distinguish; C, moderate overlap area—the median density of distribution B is masked by distribution A; D, moderate overlap area—the median density of distribution B falls on the boundary of distribution A; E, moderate overlap area—the median densities of both distributions are identifiable, indicating a distinguishable difference.

This paper proposes an analytical framework for engineering data. Its core premise is to treat data as outputs of a production system, directly recover the distributional shape of the data, and assess homogeneity between datasets on this basis—rather than inferring population parameters. The framework does not rely on random sampling assumptions, distributional assumptions, or P-value criteria. Its core criteria are: how large is the overlap between two distributions? Are the distributions distinguishable? How stable are the results?

To evaluate the framework, we generate data using random functions to simulate homogeneous, heterogeneous, and outlier-contaminated datasets, and compare the performance of Engineering Data Analysis with conventional statistical inference. The goal of this paper is not to deny the value of statistical inference, but to offer a more stable, transparent, and data-driven analytical pathway for engineering data.

Materials

Data were generated using the random function NORMINV(RAND(),mean,standard_dev) in Microsoft Office Excel 2019. Two-sample t-tests were performed using the T.TEST() function (two-tailed, equal variance assumed). Mean (AVERAGE), standard deviation (STDEV.S), and median (MEDIAN) were computed using built-in Excel functions. Theoretical normal probability densities were calculated using NORM.DIST(x,mean,standard_dev,FALSE). The IQR was calculated as $IQR = \text{QUARTILE.INC}(\text{range}, 3) - \text{QUARTILE.INC}(\text{range}, 1)$.

Data Generation

Sample size n was increased progressively by generating additional random numbers. To introduce outliers, existing values were manually replaced with integers greater than 4.

Difference Method for Density Estimation

After sorting the raw data, a sequence of distinct values $X'_1, X'_2, \dots, X'_n$ and their corresponding frequencies were obtained. For an interior value X'_i , the difference-method density was calculated as Eq.3⁸.

$$p_i = \frac{\text{freq}_i / n}{(X'_{i+1} - X'_{i-1}) / 2} \quad (\text{Eq.3})$$

For the minimum and maximum values, the denominators were $X'_2 - X'_1$ and $X'_n - X'_{n-1}$, respectively.

Interval-Shift Density Estimation (ISDE)

Density estimation was performed using an order-statistics-based adaptive interval method, termed ISDE. For engineering data that may contain repeated values, the distinct values are arranged on the real line. Let n be the total number of data points and n' the number of distinct values ($n=n'$ if there are no repeats). The window parameter k is the largest odd integer not exceeding $n'/4$, so that the window covers approximately $n'/4$ distinct values and is symmetric. Boundary cases follow symmetric stepping rules. The method uses no kernel functions, bandwidth parameters, or smoothing parameters; it is entirely data-driven and suitable for nonparametric density estimation of engineering data.

The procedure is as follows (Fig. 3):

- 1) Sort the raw data X_n to obtain the distinct-value sequence $X'_{n'}$ and their frequencies;
- 2) Compute n (total data count) and n' (number of distinct values);
- 3) Compute the window parameter k =largest odd integer $\leq n'/4$;
- 4) Compute the probability density for each distinct value from left to right:
 - 4.1) For the leftmost value X'_1 : frequency/ n , divided by $X'_2 - X'_1$ (window covers 1 value);
 - 4.2) For X'_2 : cumulative frequency from X'_1 to X'_2 , divided by $X'_3 - X'_1$ (window covers 3 values);
 - 4.3) For X'_3 : cumulative frequency from X'_1 to X'_4 , divided by $X'_5 - X'_1$ (window covers 5 values);
 - 4.4) Continue expanding by odd increments until the number of distinct values in the window reaches k ;
 - 4.5) Once the window size reaches k , slide the window rightward in steps of 1, each window covering k values, and assign the computed density to the middle value;
 - 4.6) When the right edge of the window reaches $X'_{n'}$, the window begins to contract from the right (in steps of 2) until the rightmost value is covered. The density for the maximum value is computed symmetrically to that for the minimum.
- 5) Plot the density curve using the computed densities.

Figure 3. Schematic workflow of the quarter-domain ISDE density estimation. 1) Obtain a set of raw data (n); 2) Sort the data in ascending order and remove duplicates to obtain the sequence of distinct values (n'); 3) Compute the window parameter k =largest odd integer $\leq n'/4$; 4) Iteratively compute the probability density for each distinct value according to the windowing rules.

Analogous ISDE implementations can be constructed using the half-domain method (k =largest odd integer $\leq n'/2$) and the global-domain method (k =largest odd integer $\leq n'$).

Calculation of Distribution Overlap Area (P_{overlap})

The $P_{overlap}$ between two distributions is defined as the total area of the smaller of the two density curves at each point. Since density estimates are obtained at discrete points, the $P_{overlap}$ is approximated by the trapezoidal rule (Eq.4).

$$P_{overlap} \approx \sum \min(f_A(x_i), f_B(x_i)) \cdot \Delta x_i \quad (\text{Eq.4})$$

Where x_i are the discrete sampling points and Δx_i is the interval width between adjacent points. When $P_{overlap} \geq 0.95$, the two datasets are considered to have a high degree of homogeneity and are not judged as distinguishable.

Calculation Probability Densities of Median and Judgment of Distribution

Distinguishability

For datasets A and B, first compute the medians medianA and medianB, and obtain the corresponding probability densities $f_A(\text{medianA})$ and $f_B(\text{medianB})$ either directly or by interpolation. Then evaluate $f_B(\text{medianA})$ and $f_A(\text{medianB})$. Compute the differences of Δp_A (Eq.5) and Δp_B (Eq.6).

$$\Delta p_A = f_A(\text{medianA}) - f_B(\text{medianA}) \quad (\text{Eq.5})$$

$$\Delta p_B = f_B(\text{medianB}) - f_A(\text{medianB}) \quad (\text{Eq.6})$$

The judgment criterion is as follows: if both $\Delta p_A > 0$ and $\Delta p_B > 0$, the medians of the two datasets are distinguishable, indicating a detectable difference between the distributions. If either condition fails, the medians are masked by the other distribution, and the distributions are not considered distinguishable.

Results

Random Data with Identical Parameters: Traditional Inference Diverges, Engineering Data Analysis Converges

Two datasets were generated using a random function with parameters (0,1), retaining 13 decimal places. As sample size increased incrementally, the cumulative means of both datasets converged toward 0 (Fig. 4A), and the cumulative SD converged toward 1 (Fig. 4B), indicating that larger samples yield more stable distributions. To assess the effect of measurement precision, the same data were rounded to 1 decimal place; the trends in means (Fig. 4A) and SD (Fig. 4B) remained nearly unchanged. This suggests that observational precision has a relatively minor effect on distributional description—when the data-generating mechanism is stable, classical descriptive statistics converge with increasing sample size. However, cumulative t-tests revealed that the P-values oscillated wildly, showing no stable trend, and differed between the 13-decimal and 1-decimal versions (Fig. 4C). Notably, near $n \approx 1,000$, P-values frequently fell below 0.05. This indicates that for two datasets generated by the same mechanism, classical statistical inference does not converge.

To confirm the replicability of this observation, we generated two new datasets of 10,000 values each using the same random function and performed the same cumulative analysis. The trends in means (Fig. 4D) and SD (Fig. 4E) were consistent

with the first run, but the cumulative t-test P-values exhibited another pattern of severe oscillation (Fig. 4F), with frequent $P < 0.05$ occurrences in the 5,000–9,000 range. This further confirms that classical statistical inference does not converge even when the data-generating mechanism remains unchanged.

Figure 4. Statistical behavior of two datasets generated from the same random function (0,1) with increasing sample size. A, D: Changes in means; B, E: Changes in standard deviations (SD); C, F: Changes in P-values from sequential t-tests (independent-samples t-test, two-tailed, equal variance assumed). A13, B13: raw values retained to 13 decimal places; A1, B1: raw values retained to 1 decimal place; A13vsB13: P-values for the 13-decimal data; A1vsB1: P-values for the 1-decimal data. A–C: first data generation from (0,1); D–F: second independent data generation from (0,1).

Applying the EDA framework to the same data produced different results. The medians converged rapidly toward 0 with increasing sample size, with little sensitivity to decimal precision (Fig. 5A); The IQR also converged quickly to stable values (Fig. 5B). The $P_{overlap}$ between the two distributions increased toward 1 (Fig. 5C). Once the sample size exceeded 3,000, the $P_{overlap}$ exceeded 0.95, indicating a high degree of homogeneity between the two datasets (Fig. 5C). Furthermore, Δp converged toward 0, and within the observed range, the median density of either dataset was consistently masked by the other distribution (i.e., Δp values alternated in sign), so the two distributions were not distinguishable (Fig. 5D).

The same analyses performed on the second set of randomly generated data yielded similar results (Fig. 5E–H): both central tendency (median) (Fig. 5E) and dispersion (IQR) (Fig. 5F) converged, and both $P_{overlap}$ (Fig. 5G) and Δp (Fig. 5H) showed convergent trends, all indicating that the two datasets were not distinguishable.

Figure 5. Re-analysis of the two datasets from Figure 4 using the Engineering Data Analysis (EDA) framework. With increasing sample size, the median (median) (A, E), interquartile range (IQR) (B, F), overlap probability ($P_{overlap}$) (C, G), and Δp (D, H) all show convergent trends. $\Delta p = f_A(\text{medianA}) - f_B(\text{medianA})$ or $f_B(\text{medianB}) - f_A(\text{medianB})$. A13, B13: raw values retained to 13 decimal places; A1, B1: raw values retained to 1 decimal place. A–D: first data generation; E–H: second data generation. Probability densities were computed using the quarter-domain ISDE method when $n > 100$ and $n \geq 12$; when $n < 100$ and/or $n < 12$, half-domain ISDE, global-domain ISDE, or the difference method was used.

Sensitivity to Outliers: Statistical Inference is Sensitive, Engineering Data Analysis is Robust

Two datasets were generated using the same random function with parameters (0,1) and retained to 1 decimal place. To examine the effect of outliers, values in dataset A were progressively replaced by outliers, while dataset B remained unchanged. During this process, the mean (Fig. 6A) and SD (Fig. 6B) of dataset A increased only slightly. When outliers were introduced sequentially from 4 to 22 (a total of 19 outliers), the change in SD still satisfied the assumption of homogeneity of variance, yet the t-test P-value had already dropped below 0.05 (Fig. 6C). This

indicates that introducing roughly one outlier per 500 data points can turn a supposedly non-significant t-test into a “statistically significant” one. Traditional statistical inference is thus highly sensitive to outliers, even though, from a distributional perspective, one outlier in 500 values should have only a minimal impact on the overall distribution.

Re-analyzing the same data using the EDA framework revealed that, during the introduction of the 19 outliers, the median of dataset A remained unchanged (Fig. 6D), and the IQR also remained stable (Fig. 6E). The overlap probability between the two datasets showed little decline (Fig. 6F), and Δp consistently alternated in sign (Fig. 6G), indicating that the two distributions were not clearly distinguishable. The probability density plots further confirmed that, both before (Fig. 6H) and after (Fig. 6I) the introduction of outliers, the two distributions remained highly overlapping, with the median density of one distribution consistently masked by the other. These results demonstrate that, within the EDA framework, the presence of a small number of outliers has little influence on the homogeneity assessment between two datasets.

Figure 6. Engineering Data Analysis (EDA) results when one dataset remains unchanged and the other is progressively contaminated with outliers ($n=10,000$). Two datasets A and B were generated from random function (0,1) and retained to 1 decimal place. Values in dataset A were progressively replaced by outliers starting from 4, up to a total of 19 replacements (4–22). A, Changes in mean; B, changes in SD; C, changes in t-test P-value (independent-samples t-test, two-tailed, equal variance assumed); the P-value fell below 0.05 after 19 outliers. D, median; E, interquartile range (IQR); F, overlap probability ($P_{overlap}$); G, Δp (maintaining alternating signs); H, Probability density (p) distributions without outliers; I, probability density distributions after 19 outliers. A1: dataset A retained to 1 decimal place; B1: dataset B retained to 1 decimal place. $\Delta p = f_A(\text{medianA}) - f_B(\text{medianA})$ or $f_B(\text{medianB}) - f_A(\text{medianB})$. Probability densities were computed using the quarter-domain ISDE method.

Comparison with Original Data: Statistical Inference is Sensitive to Outliers, Engineering Data Analysis Remains Stable

Two datasets (A and B) were generated using the same random function with parameters (0,1) and retained to 1 decimal place ($n=10,000$). Values were sequentially replaced by outliers ranging from 4 to 25 (a total of 22 outliers) in both datasets.

During the replacement process, the mean (Fig. 7A) and SD (Fig. 7B) of dataset A changed only slightly. However, compared with the original data, the t-test P-value dropped rapidly after the introduction of outliers. When 22 out of 10,000 values had been replaced by outliers, the P-value fell below 0.05 (Fig. 7C), leading to a judgment of “statistically significant difference” to the original dataset.

Re-analyzing the same data using the EDA framework revealed that neither the median (Fig. 7D) nor the IQR (Fig. 7E) changed. Even after 22 values were replaced by outliers, the overlap probability between the original and replaced data remained above 0.95 (Fig. 7F), and Δp values alternated in sign (Fig. 7G), indicating that the replaced datasets still exhibited a high degree of homogeneity to the original. The

probability density plots (Fig. 7H, Fig. 7I) further supported this conclusion: the introduction of a small number of outliers did not alter the overall comparability of the distributions.

Figure 7. Re-analysis of two datasets generated from random function (0,1), retained to 1 decimal place ($n=10,000$), with outliers introduced, using the Engineering Data Analysis (EDA) framework. Outliers from 4 to 25 were progressively introduced to replace existing values (22 outliers in total). A, Changes in mean; B, changes in SD; C, changes in t-test P-value (independent-samples t-test, two-tailed, equal variance assumed), which dropped rapidly below 0.05; D, median; E, interquartile range (IQR); F, overlap probability (P_{overlap}); G, Δp (alternating in sign). H, Probability density (p) distributions without outliers; I, probability density distributions after 22 outliers. A1' in H and I: dataset A with 22 outliers; B1' in H and I: dataset B with 22 outliers. $\Delta p = f_A(\text{medianA}) - f_B(\text{medianA})$ or $f_B(\text{medianB}) - f_A(\text{medianB})$. Probability densities were computed using the quarter-domain ISDE method.

Comparison of Large Datasets with Different Parameters: Engineering Data Analysis is More Reliable and Stable

Two datasets of 10,000 points each were generated using random functions with parameters (0,1) and (0.5,1), retaining either 13 or 1 decimal places. As sample size increased, both the means (Fig. 8A) and SD (Fig. 8B) showed convergent trends. In contrast, the t-test P-values oscillated wildly when the sample size was below 100, and then rapidly approached zero (Fig. 8C).

When the same data were analyzed using the EDA framework, the medians (Fig. 8D) and IQR (Fig. 8E) also converged. The P_{overlap} converged to around 0.8, consistently remaining below 0.9 (Fig. 8F). The Δp values showed a single sign change at $n \approx 30$, and eventually converged to positive values (Fig. 8G), indicating that the two datasets were not homogeneous.

The density curves showed that at $n=100$, the distribution profile was already discernible but still exhibited noticeable roughness (Fig. 8H). As the sample size increased to $n=300$ (Fig. 8I), 1,000 (Fig. 8J), 3,000 (Fig. 8K), and 10,000 (Fig. 8L), the curves became progressively smoother, demonstrating that the stability of the EDA method improves with increasing data volume.

Figure 8. Analysis of differences between two non-homogeneous datasets using the Engineering Data Analysis (EDA) framework. Two datasets were cumulatively generated from random functions (0,1) and (0.5,1) with 10,000 values each, retained to either 13 or 1 decimal places. With increasing sample size, both means (A) and SD (B) showed convergent trends; medians (D) and interquartile ranges (IQR) (E) also converged. The t-test P-values (independent-samples t-test, two-tailed, equal variance assumed) fluctuated wildly when $n < 100$ but eventually converged toward 0 (C). EDA revealed that the overlap probability (P_{overlap}) (F) converged to a value below 0.85; Δp (G) also showed convergence overall, though with sign fluctuations at small sample sizes. The probability density plots (H, $n=100$; I, $n=300$; J, $n=1,000$; K, $n=3,000$; L, $n=10,000$) indicate that the two distributions had some overlap but remained distinguishable, suggesting a discernible difference. $\Delta p = f_A(\text{medianA}) - f_B(\text{medianA})$ or $f_B(\text{medianB}) - f_A(\text{medianB})$. A13, B13: raw values retained to 13 decimal places; A1, B1: raw values retained to 1 decimal place. When $n < 100$ and/or $n' < 12$, probability densities were computed using half-domain ISDE, global-domain ISDE, or the difference method.

395

396 Discussion

397 In this study, data generated by random functions were used to compare the
398 performance of traditional statistical inference and the proposed EDA method. For
399 two datasets generated from the same random function, the central tendency (mean,
400 median) and dispersion (SD, IQR) both converged with increasing sample size,
401 while the P-values from conventional t-tests exhibited oscillatory behavior⁹,
402 indicating that the P-value does not converge as sample size increases^{10,11}. When
403 datasets were generated from different parameter settings (0,1) and (0.5,1), the
404 P-values again fluctuated wildly when $n \leq 100$; only when the sample size became
405 sufficiently large did the P-values approach 0, rather than stabilizing at some
406 non-zero value that reflected the actual difference between the distributions. This
407 suggests that the convergence of P-values is extreme (toward 0) rather than stable
408 (toward a fixed value).

409 By contrast, the EDA method showed a clear convergence. For datasets
410 generated from the same random function, the $P_{overlap}$ steadily approached 1 as
411 sample size increased. For datasets generated from different parameters (0,1) and
412 (0.5,1), the $P_{overlap}$ converged to a stable value between 0 and 1 that reflected the
413 actual degree of difference. Moreover, the number of decimal places retained had
414 little effect on the $P_{overlap}$, whereas the t-test P-values fluctuated considerably on the
415 same dataset. This indicates that the EDA method is more robust to both
416 measurement precision and sample size.

417 The classical law of large numbers concerns the convergence of descriptive
418 statistics such as the mean and SD. The logical extension of this study is that when
419 the data description converges, the judgments based on that description should also
420 converge: homogeneous data should yield an $P_{overlap}$ near 1; non-homogeneous data
421 should yield a stable value between 0 and 1 that reflects the actual difference; only
422 extreme differences should yield a $P_{overlap}$ tending to 0. This behavior extends the
423 scope of "large-number stability"¹² from the descriptive level to the judgment level,
424 achieving consistent convergence between data description and difference judgment,
425 and directly linking the judgment to the actual difference between distributions.

426

427 Distinction Between Engineering Data and Statistical Data

428 Before discussing the applicability of EDA, it is necessary to clarify a
429 fundamental issue: engineering data and statistical data differ fundamentally in their
430 origin, nature, and purpose. Confusing these two types of data is a major source of
431 misunderstanding in current data analysis practice.

432 The validity of statistical methods rests on "random sampling"—the assumption
433 that samples are representative of a population and that the sampling process is
434 random and independent. Engineering data, by contrast, are "produced": they are not
435 "drawn" from a "population", but rather "output" from a "production system". Each
436 value in engineering data is a product of the production process and should not be
437 treated as a "sample" (Table 1).

The numerical characteristics of engineering data are often constrained by physical conditions: temperature has a lower bound, concentration cannot be negative, counts are discrete integers, and repeated measurements at finite precision produce ties. These features differ markedly from the data forms assumed in statistical textbooks (Table 1).

The goal of statistics is "inference"—inferring population parameters from samples, estimating parameters from sampling, and discovering patterns from observations. The conclusions are probabilistic: "at the 95% confidence level...". The goal of EDA is "discrimination"—determining whether two datasets are homogeneous, whether distributions are distinguishable, and whether a system is stable. The conclusions are judgmental: "how big is the overlap area between the two distributions, below the threshold, therefore distinguishable" (Table 1).

Table 1. Comparison of characteristics between engineering data and statistical data

Performance of the Proposed Method

This section evaluates the proposed EDA method from three perspectives: intuitiveness, stability, and fidelity.

Intuitiveness: Distribution Plots Replace Summary Statistics Traditional statistical methods typically report results as "mean $\pm$ SD" or "P-values." While concise, these statistics lack direct visual correspondence to the actual distribution shape of the data. For asymmetric, multimodal, or boundary-affected distributions, a single summary statistic often obscures critical features of the data.

In contrast, ISDE directly outputs probability density estimates for each data point, allowing the construction of a complete density curve. Analysts can directly observe distribution features—peak locations, peak widths, tail directions, and the presence of multiple modes. This "plot-over-statistic" presentation makes the results more accessible to engineering practitioners and facilitates communication within teams.

More importantly, the ISDE-generated density curve does not depend on any distributional assumptions—it shows the shape of the data as it is. This intuitiveness not only improves analytical efficiency but also reduces the risk of misjudgment caused by misuse of summary statistics.

Stability: Adaptive Window, Limited Outlier Impact Conventional kernel density estimation (KDE) requires preselection of a kernel function and bandwidth¹³, both of which critically affect the density estimate. Too small a bandwidth yields excessive spurious fluctuations; too large a bandwidth smooths over genuine structures. In practice, bandwidth selection often relies on subjective judgment or complex optimization algorithms, lacking a stable and reproducible engineering standard.

ISDE adopts an adaptive strategy with a window covering approximately $n/4$ data points. The window width is determined automatically by local data density: contracting in dense regions to preserve structural details, and expanding in sparse regions to avoid spurious fluctuations. This adaptivity enables ISDE to perform consistently across different datasets without requiring repeated parameter tuning.

Furthermore, ISDE windows are positioned based on data point order rather than numerical range. As a result, the impact of extreme outliers on window width is naturally confined to the local region, unlike the mean or SD, which can be pulled by outliers. This robustness is particularly important for engineering data, which often contain measurement anomalies, boundary values, or extreme values from system faults—conditions to which conventional statistical methods are overly sensitive.

Fidelity: Data Determine the Distribution, Not the Reverse Conventional density estimation typically starts with an assumption—most commonly that the data follow a normal distribution—and then "fits" a density curve through parameter estimation. In practice, this means imposing a shape on the data. When the data deviate from the assumption, the fitted results may mask or even distort the true distribution.

ISDE reverses this logic: it imposes no distributional shape, instead allowing the data to "speak" its own density structure. The ISDE window is determined entirely by the rank order and local density of the data, with no external parameters. The curve rises where data are concentrated and flattens where data are sparse. This "data-driven" approach ensures that the density estimate remains faithful to the actual structure of the data, free from subjective assumptions.

In assessing homogeneity between two datasets, this fidelity is especially important. If data are first "reshaped" by normality assumptions, genuine distributional differences may be artificially eliminated or amplified. The ISDE-recovered density curves reflect the original data structure directly, making overlap-area judgments more reliable and more aligned with the actual engineering state.

Avoidance of Fit-Induced Artifacts In addition, many conventional density estimation methods involve complex numerical transformations or parameter optimization processes—essentially "fitting" the estimate to a theoretical distribution¹⁴. These fitting processes introduce additional sources of uncertainty: optimization convergence, initial value selection, and step-size decisions can all influence the final result.

ISDE involves no optimization and requires no numerical transformation. Its computation is direct and deterministic: sort, determine window positions, count points within the window, divide by window width. No iteration, no convergence criteria, no parameter tuning. This ensures that ISDE results are highly reproducible—the same data, run at any time on any platform with the same ISDE implementation, will yield identical results.

A Modest Guard Against Data Tampering This framework serves not only as an analytical tool but also as an engineering safeguard for data integrity. Under this framework, localized data tampering cannot alter the overall distribution shape, while wholesale fabrication of a distribution is prohibitively costly to achieve convincingly. Thus, the framework naturally encourages truthful data production and provides a practical, physically grounded mechanism for promoting research integrity.

Potential Value of Engineering Data Analysis for Experimental and Engineering Research

Based on the distinctions discussed above, EDA offers an analytical pathway distinct from statistical inference, with its value manifested in three main aspects.

Avoiding the False Significance Problem When data volume is sufficiently large, conventional statistical tests almost invariably yield "statistically significant" results—even when both datasets originate from the same random process. This phenomenon is not a flaw of statistical theory per se, but rather a misapplication of the inference framework to non-sampling data. EDA circumvents this issue by shifting the criterion from "whether $P < 0.05$ " to "whether the distribution overlap area is sufficiently small." Instead of asking "is the difference significant?", it asks "are the two distributions distinguishable?"

Revealing the True Structure of Data Statistical inference typically begins with assumptions (e.g., normality) and then uses data to "test" those assumptions. In effect, this forces the data to fit the model. EDA takes the opposite approach: it uses adaptive methods such as ISDE to recover the distribution directly from the data, letting the analysis fit the data. This orientation is closer to the "let the data speak" principle common in engineering practice, and it helps identify true features—such as multimodality, boundary effects, and tied-value structures—that are often masked by assumption-driven testing.

Directness in Supporting Engineering Decisions Statistical inference ultimately produces a probabilistic conclusion (e.g., " $P = 0.03$ "), which requires secondary interpretation before it can inform engineering decisions. EDA, by contrast, directly provides visual comparisons of distributions and overlap areas, allowing engineers to directly assess questions such as "are these two batches of products consistent?" or "is this measurement system stable?" The path from analysis result to decision is shorter, and the communication cost is lower.

Recommended Workflow for Engineering Data Analysis

Based on the findings of this study, we recommend the following workflow for EDA.

Data Preparation and Parameter Fixation The raw data should be prepared in the following sequence:

- 1) Sort and deduplicate the raw data to obtain a list of distinct values;

- 2) Count the frequency of each distinct value;

- 3) Fix three key parameters (n , n' , k):

n : total number of data points;

n' : number of distinct values;

k : window parameter, set to the largest odd integer $\leq n'/4$ (quarter-domain method), $\leq n'/2$ (half-domain method), or $\leq n'$ (global-domain method), depending on the chosen method.

Purpose: To ensure that each subsequent step is verifiable and to avoid result drift caused by redundant computation.

Density Estimation Select an appropriate density estimation method based on data characteristics (Table 2).

Table 2. Selection of density estimation methods

Single-Dataset Inspection Plot the probability density curve and inspect the following features:

- 1) Peak position and number: unimodal, multimodal, or flat regions;
- 2) Tail direction: presence of skewness;
- 3) Boundary accumulation: whether data are unusually concentrated near boundaries.

Purpose: To assess data quality and identify potential anomalies or batch effects.

Two-Dataset Comparison Compute the $P_{overlap}$ using the trapezoidal rule; Assess distinguishability (whether the median density of one distribution is masked by the other).

Reference criteria:

- $P_{overlap} \geq 0.95$: high homogeneity, not judged as different;
- $P_{overlap} < 0.95$ and medians distinguishable: distributions differ;
- $P_{overlap} < 0.95$ but medians not distinguishable: distribution shapes may differ; further inspection of density curves is recommended;

Very low $P_{overlap}$ tends to indicate clear heterogeneity; specific thresholds may be adjusted based on engineering context.

Judgment and Output When the $P_{overlap}$ and distinguishability criteria are consistent, a conclusion of "homogeneous" or "distinguishable" can be drawn directly. If they conflict, further judgment is recommended based on the density curves and engineering context; no forced conclusion should be made.

Core principle: Sample size affects the reliability of the result, but does not change the direction of convergence of the criteria.

Limitations of Engineering Data Analysis

EDA is not intended to replace statistical methods in all domains. For scenarios that genuinely satisfy the assumptions of random sampling—such as opinion surveys, market research, and epidemiological sampling surveys—statistical inference remains appropriate. The applicability of EDA lies in data originating from measurement, sensing, experimentation, and production pipelines, where the analytical objectives are to judge system status, compare batches, and assess stability—rather than to infer population parameters. Statistics is a toolbox for "mathematical inference", while Engineering Data Analysis is a toolbox for "engineering discrimination". Each has its own applicable contexts; conflating them does not bring benefits.

EDA also has certain data requirements. We recommend that the total data size exceed 100 and that the number of distinct values be at least 12 (i.e., $n'/4 \geq 3$) to enable the use of the quarter-domain ISDE method for probability density estimation. This also implies that measurement precision should not be too low. For data that do not meet these conditions, half-domain, global-domain, or even difference methods may still be used to compute probability density and overlap area, but with larger analytical errors.

We acknowledge that the framework proposed in this paper is still at a preliminary stage. The recommended criteria— $P_{overlap}$ greater than 0.95 and distribution distinguishability—are offered as practical guidelines. Further research is needed on threshold determination, comparison methods for multiple datasets, and integration of time-series data. However, our core argument remains unchanged: engineering-oriented big data should not be analyzed using traditional statistical inference.

Conclusions

This study proposes EDA as an analytical framework for engineering data. Its core premise is that experimental and engineering data should be treated as outputs of a production system rather than as randomly sampled observations. The framework does not rely on distributional assumptions, kernel functions, or data transformations. Instead, it recovers the underlying distribution directly through ISDE, and assesses homogeneity between two datasets using distribution overlap and distinguishability as criteria. In contrast, the proposed framework converges stably: homogeneous data yield high overlap and are not distinguishable, while heterogeneous data yield identifiable differences, with limited sensitivity to a small number of outliers. In this framework, sample size affects only the reliability of the results, not the direction of convergence of the criteria. We conclude that for engineering data with a large data size, EDA provides a more stable, transparent, and practically grounded basis for judgment.

Author's Note

This study received no funding from any institution or project. The logical framework was independently developed by a researcher in basic medical sciences, with the VBA code for Excel 2019 written with the assistance of DeepSeek. All code has been tested by the author and yielded reasonable and reliable results. Due to the memory and computational limitations of the Excel platform, the validation data in this study were limited to 10,000 rows, but the method itself is scalable to larger datasets. All code will also be provided as supplementary material. A Chinese version of this paper (slightly different) has been published on Zhihu web (link: <https://zhuanlan.zhihu.com/p/2080635336934405131>) to facilitate peer discussion and exchange. Readers are welcome to identify issues or suggest improvements; all feedback will contribute to further refinement of the method.

Acknowledgements

I thank the countless pathological slides I have encountered over the years. They taught me that the truest form of data often lies in its own folds—not in the mold of some mathematical formula.

References

653

654 1 Zhang, Z. & Zhao, L. An Analytical Study on the Logic of Data as a Factor of Production
655 (Chinese). *Journal of Hebei Agricultural University (Social Sciences Edition)* **27**, 13-21,
656 doi:10.13320/j.cnki.jauhe.2025.0060 (2025).

657 2 COURSEKATA. *3.10 The Data Generating Process*,
658 <[https://coursekata.org/preview/book/2333513e-d74b-4029-8592-e571d545f5ce/les](https://coursekata.org/preview/book/2333513e-d74b-4029-8592-e571d545f5ce/lesson/5/9)
659 [son/5/9](https://coursekata.org/preview/book/2333513e-d74b-4029-8592-e571d545f5ce/lesson/5/9)> (2025).

660 3 Ioannidis, J. P. Why most published research findings are false. *PLoS Med* **2**, e124,
661 doi:10.1371/journal.pmed.0020124 (2005).

662 4 Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed
663 flexibility in data collection and analysis allows presenting anything as significant.
664 *Psychol Sci* **22**, 1359-1366, doi:10.1177/0956797611417632 (2011).

665 5 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The extent and
666 consequences of p-hacking in science. *PLoS Biol* **13**, e1002106,
667 doi:10.1371/journal.pbio.1002106 (2015).

668 6 Zanin, M., Lóczy, R. & Zanin, A. "Outlier detection: Underused in sport science despite
669 outliers' impact on inference and prediction". *J Sports Sci* **42**, 2495-2505,
670 doi:10.1080/02640414.2024.2443313 (2024).

671 7 Yan, H., Duan, W. & Yun, Y. A New Strategy of Statistical Analysis Based on Graphic
672 Analysis. *IEEE Computer Society*, 174-177, doi:10.1109/ICISEM.2013.55 (2013).

673 8 J., V. R. A Histogram method of density estimation. *Communications in Statistics* **2**,
674 doi:10.1080/03610927308827093 (1973).

675 9 A, G. & H, S. The difference between "significant" and "not significant" is not itself
676 statistically significant. *The American Statistician* **60**, 328-331,
677 doi:10.1198/000313006X152649 (2006).

678 10 Xu, J. *et al.* A global P-Process map to avoid P-value abuse. *bioRxiv*,
679 doi:10.1101/2025.04.06.646254 (2025).

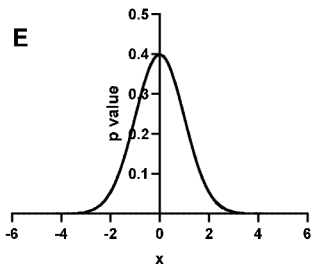
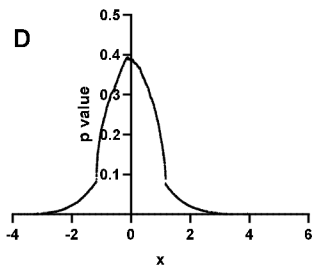
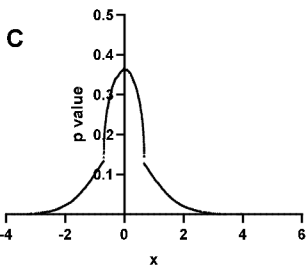
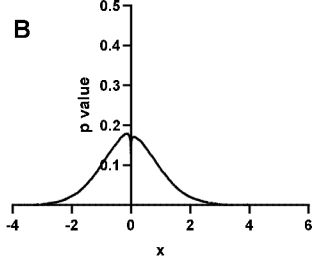
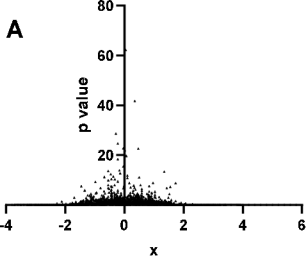
680 11 Brimacombe, M. P-values and laws of large numbers: Practical interpretation of
681 significance in very large samples. *Population Medicine*,
682 doi:<https://doi.org/10.18332/popmed/206958> (2025).

683 12 I, I. S. The Law of Large Numbers: Some Issues of Interpretation and Application for
684 Beginners. *Journal of the Royal Statistical Society: Series D (The Statistician)* **28**, 209-217,
685 doi:10.2307/2987870 (1979).

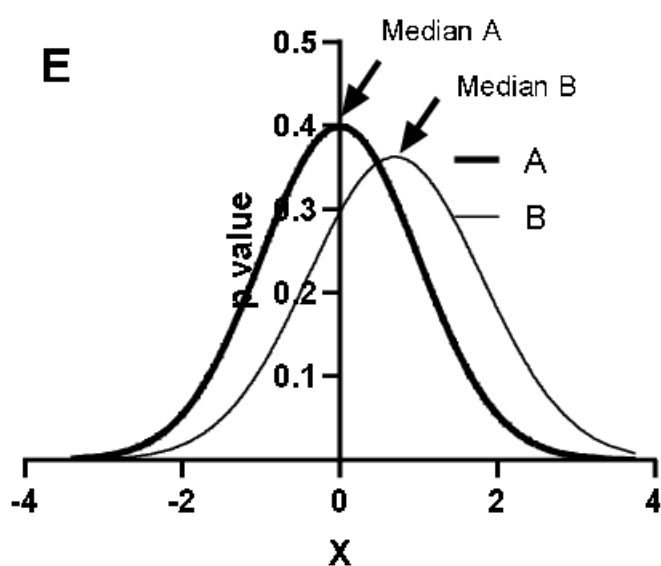
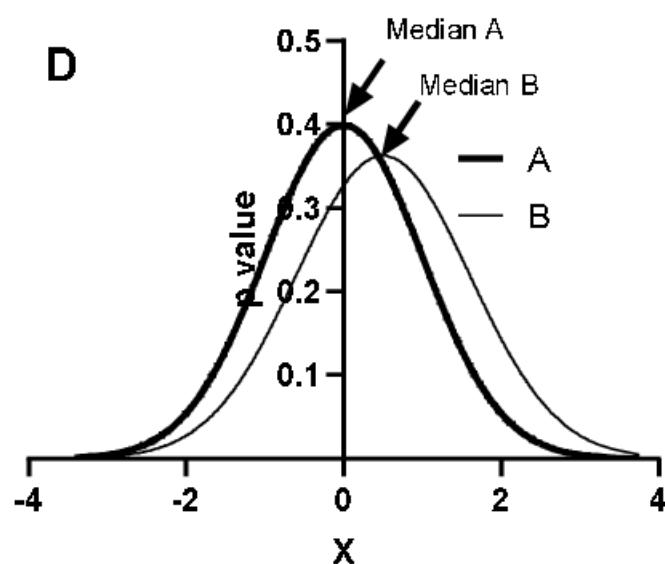
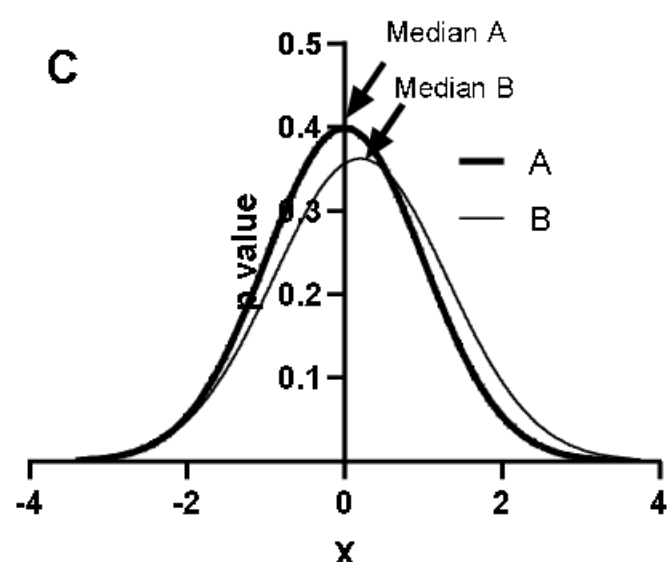
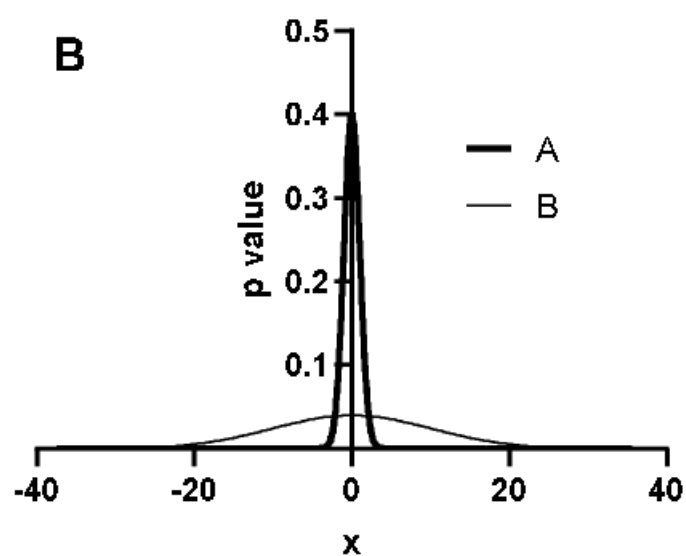
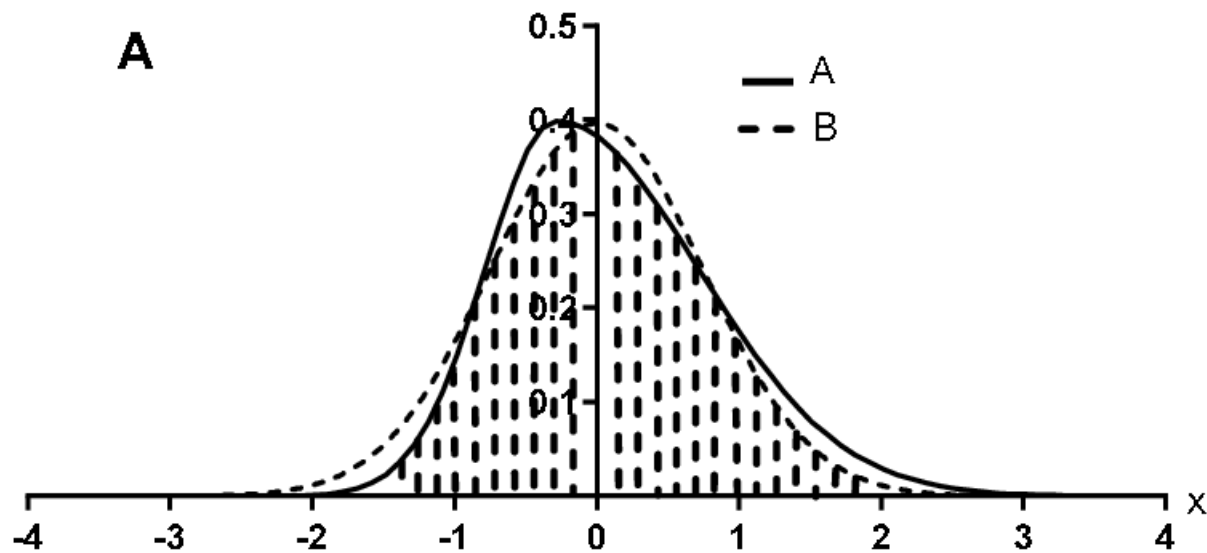
686 13 C, J. M., S, M. J. & J, S. S. A Brief Survey of Bandwidth Selection for Density Estimation.
687 *Journal of the American Statistical Association* **91**, 401-407,
688 doi:10.1080/01621459.1996.10476701 (1996).

689 14 M, D. J. & G, D. M. Invariance, nonlinear models, and asymptotic tests. *Econometrica* **59**,
690 1601-1620, doi:10.2307/2938281 (1991).

691



p value



1)

$$X_1, X_2, X_3, \dots, X_n$$



2)

$$X'_1, X'_2, X'_3, \dots, X'_{n'}$$

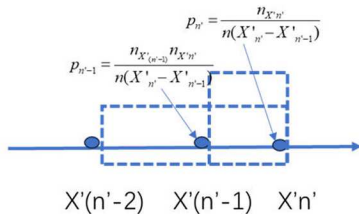
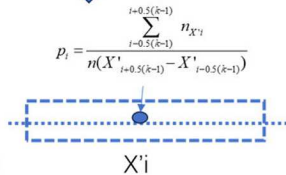
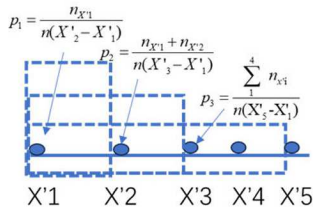


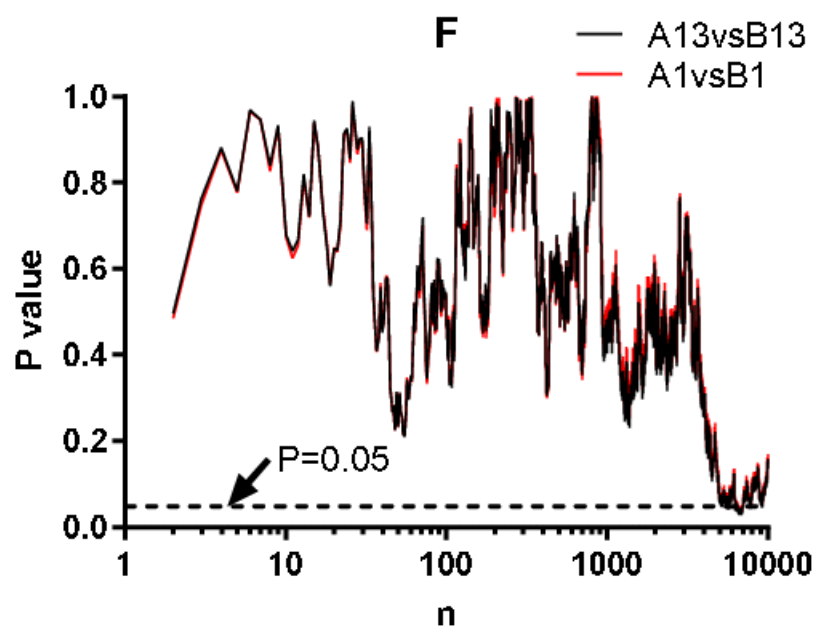
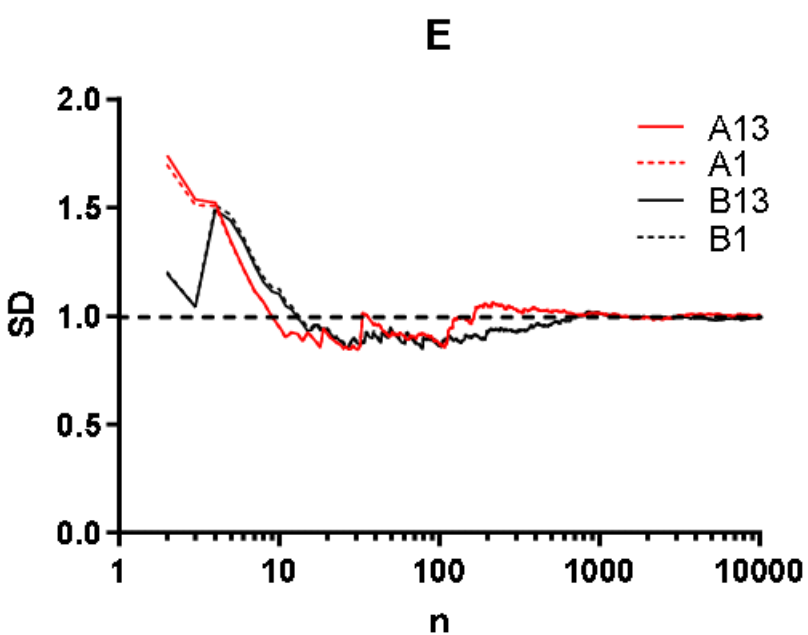
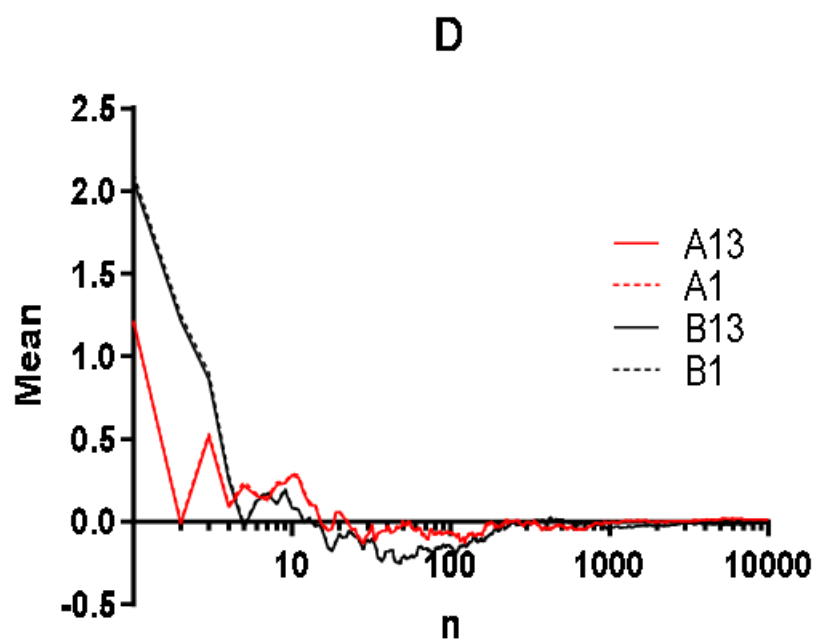
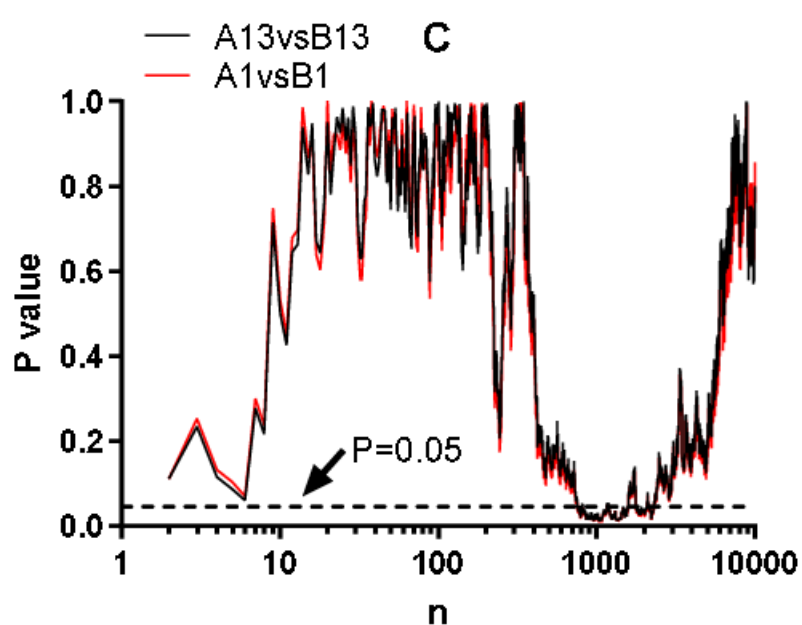
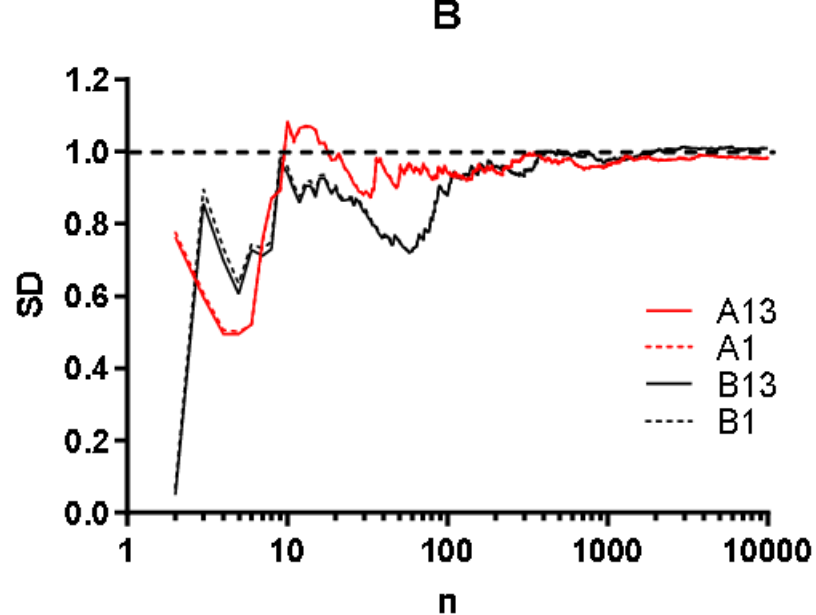
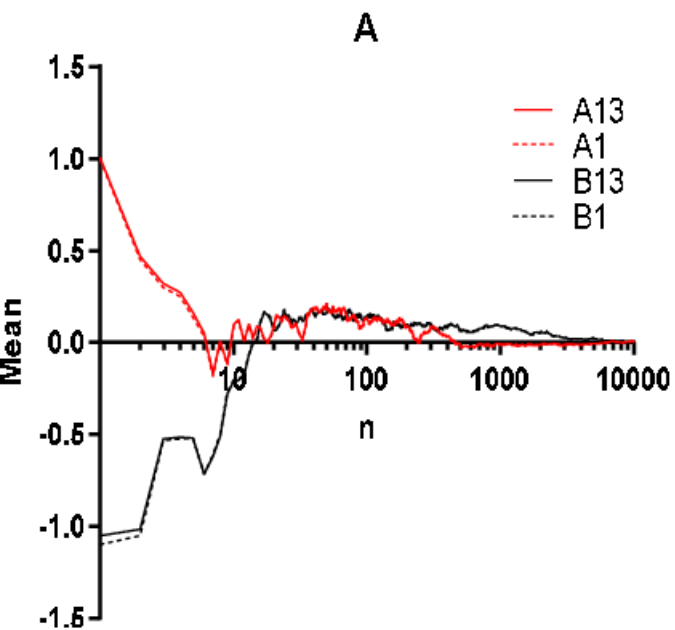
3)

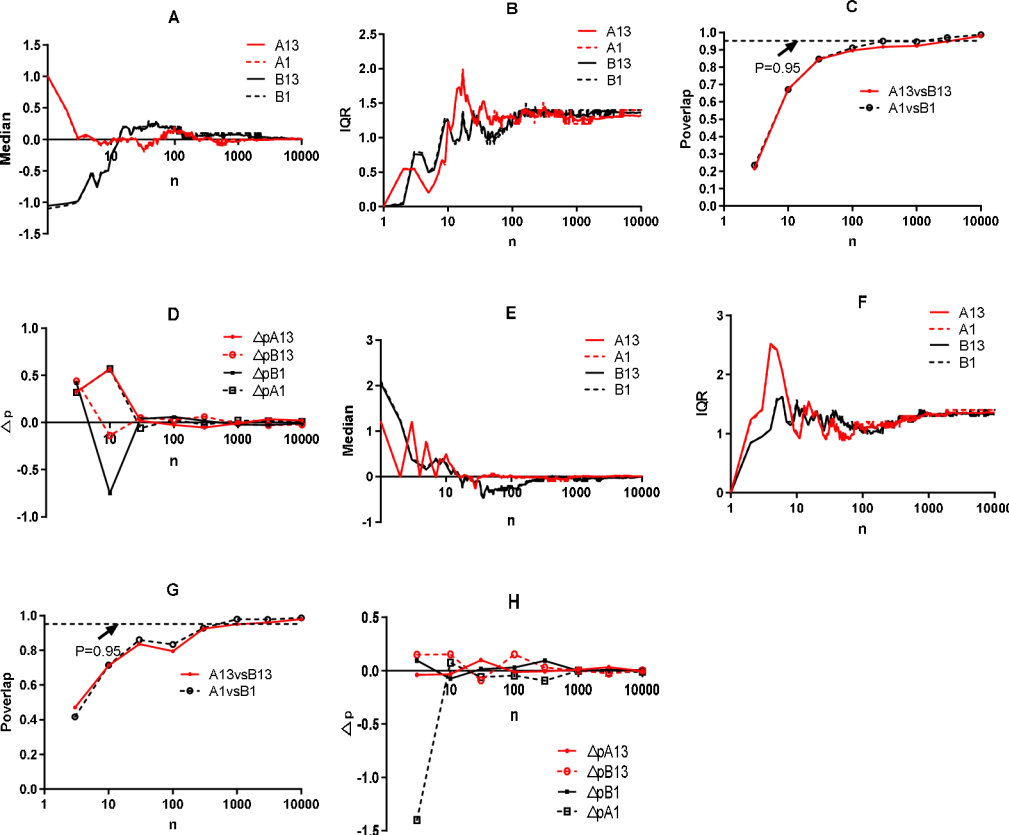
$$K = \text{Max_odd}(n'/4)$$

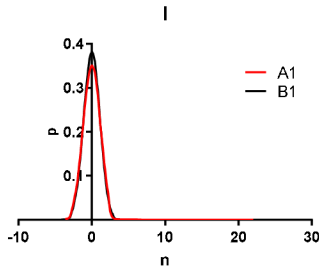
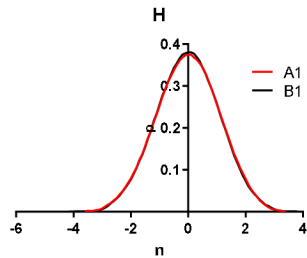
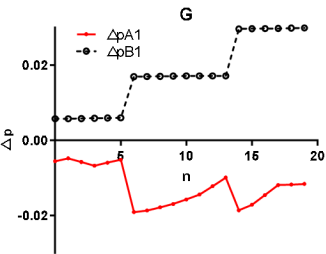
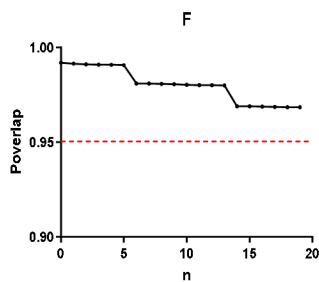
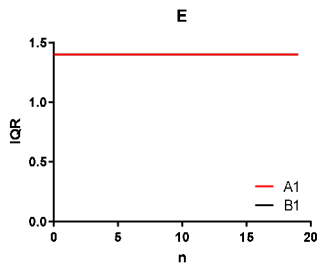
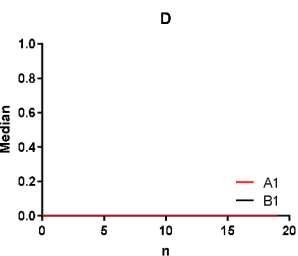
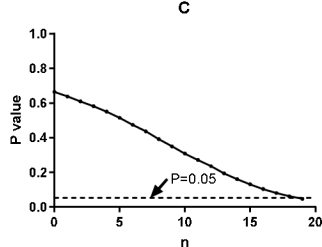
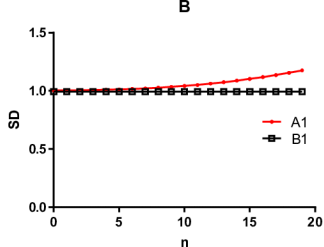
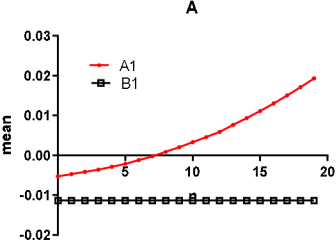


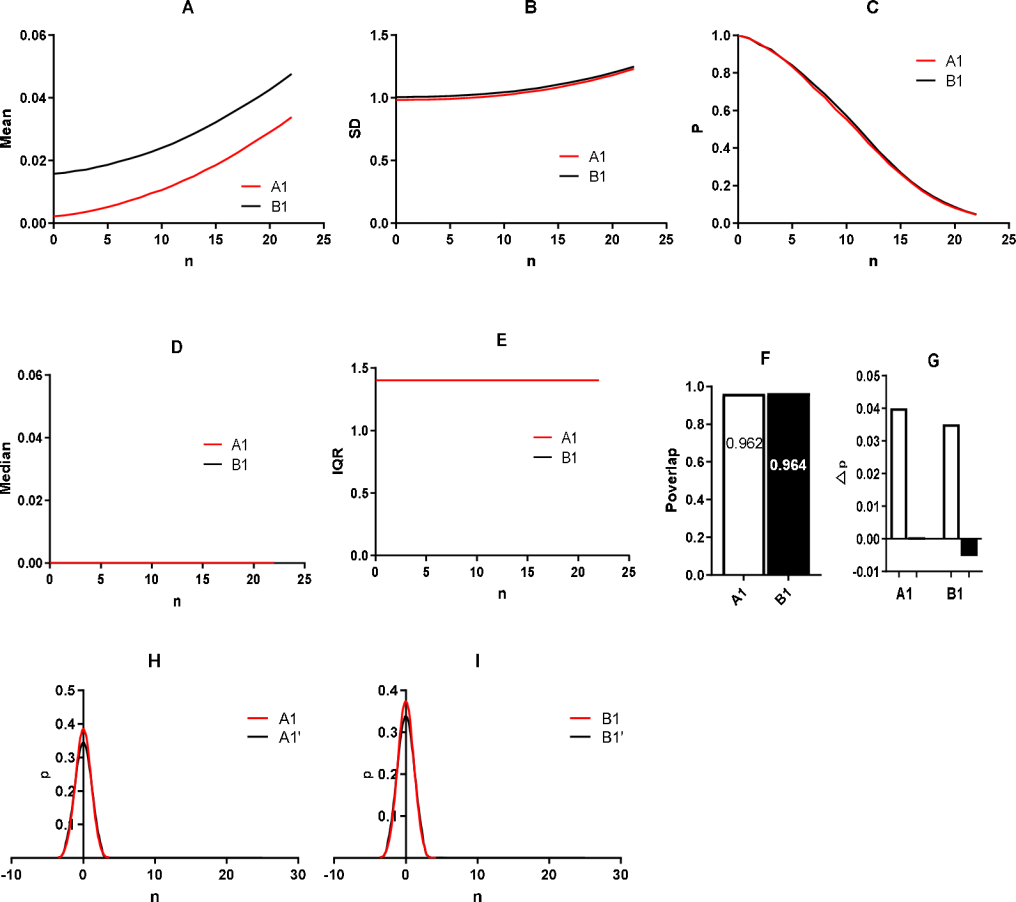
4)

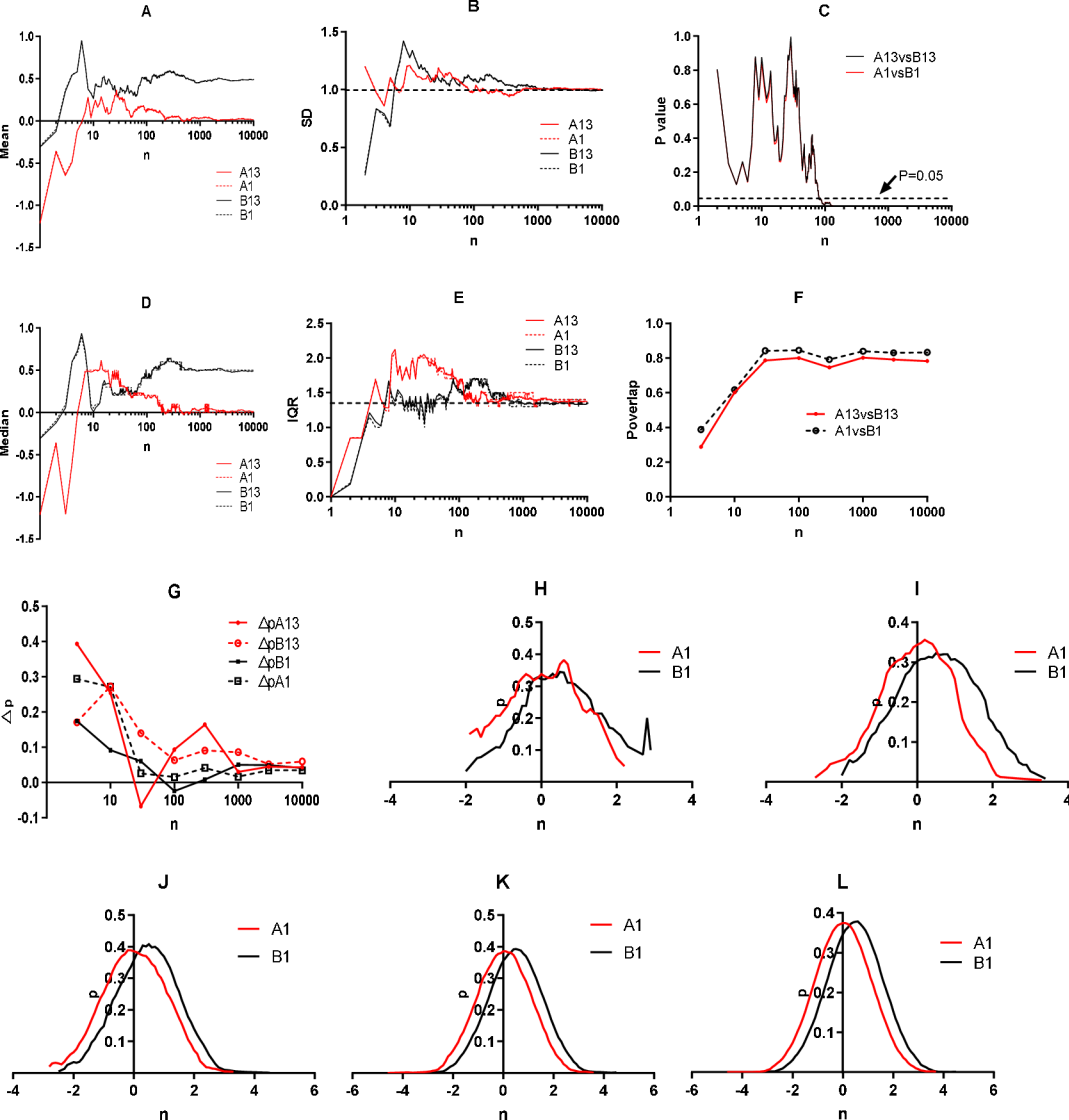












Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table2.docx](#)
- [Table1.docx](#)
- [Figure7FGHI.xlsx](#)
- [Figure1.xlsx](#)
- [Figure7ABCDE.xlsx](#)
- [Figure5ABCD.xlsx](#)
- [Figure6ABC.xlsx](#)
- [Figure6DEFGHI.xlsx](#)
- [Figure8FGHIJKL.xlsx](#)
- [Figure8ABCDE.xlsx](#)
- [Figure5EFGH.xlsx](#)
- [Figure4.xlsx](#)