Statistical Tests of Heteroscedasticity as a Fairness Metric for Bounded Regression Problems

M.A. Douma & E.M.A.L. Beauxis-Aussalet

VU Amsterdam

Abstract. Fair regression systems shall produce homogeneous residuals across protected groups, i.e., both mean and variance of residuals shall be equivalent across protected groups. Should residuals have unequal variance across protected groups, i.e., exhibit heteroscedasticity, some groups would be disadvantaged by such increased error in regression results. We thus evaluate the suitability of statistical tests of heteroscedasticity applied to identifying fairness issues in regression problems. We compare the performance of 6 standard variance tests (Levene, Brown-Forsythe, Fligner-Killeen, Bartlett, F-ratio and one-way ANOVA) applied on residuals with varying sample sizes and patterns of heteroscedasticity in 2 protected groups. We use synthetic data that represent residual values constrained within an [0,1] interval to simulate bounded regression problems (e.g., logistic regressions predicting class membership probability). We created 13 synthetic distributions of residuals that differ in their mean, variance, and distribution types (i.e. normal and binomial). We experimented with with 25 pairings of residual distributions, and 25 sample size combinations. Results show that Bartlett and F-ratio tests perform most reliably overall, even with binomial residual distributions. Notably, all tests fail to identify heteroscedasticity with mirrored binomial distributions, as standard tests are designed to consider these as statistically equivalent. Yet critical fairness issues remain, as one protected group has higher residuals than the other. We thus highlight the criteria to consider when choosing a suitable test of heteroscedasticity for fairness assessments, and the limitations that remain when dealing with symmetric distributions or small sample sizes.

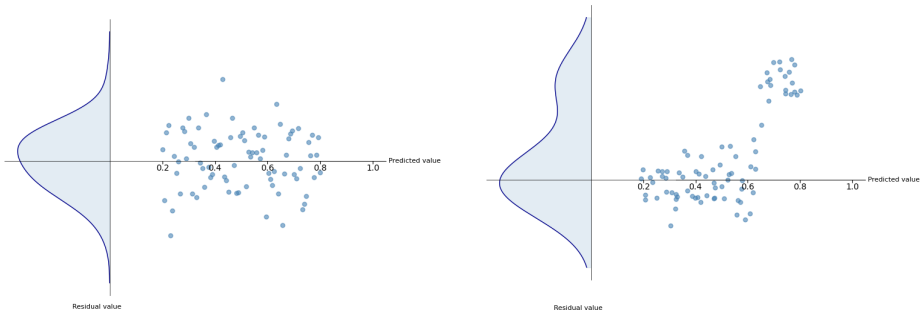
1 Introduction

Statistical tests of heteroscedasticity have been designed to assess the equality of variance. They can be used to determine whether differences in residual variance between groups of data points are significant or the result of random variability. The reliability of these tests depends on their underlying assumptions, and violations of these assumptions can result in incorrect conclusions. Although these tests are well understood theoretically, there is limited research on how well they perform for fairness assessment. For such purpose, some underlying assumptions may be violated, such as the normality of residual distributions. We investigate such fairness assessment purposes, where identifying heteroscedasticity is important for identifying unequal treatment of protected social groups. In such use

cases, heteroscedasticity of residuals across protected groups (e.g., of different genders) means that certain social groups are impacted by increased error and uncertainty, and more unreliable predictions. This is a form of discrimination (e.g., a sexist bias).

For this study, our use case concerns bounded regressions, e.g., logistic regressions used for binary classification. The protected groups can be minorities, and have small sample sizes, class imbalance, or non-normal residual distributions (even if regression models are fitted with normal residuals and null mean for the entire training data). We investigate the impact of these characteristics, which can violate the assumptions of certain statistical tests of heteroscedasticity.

We systematically compare the performance of 6 state-of-the-art statistical tests for heteroscedasticity under conditions that are relevant for fairness problems: small sample sizes, class imbalance, and non-normal residual distributions. Our leading research question is: *How is the performance of statistical tests for heteroscedasticity influenced by sample sizes and differences in the distribution of residuals?* To answer it, we use of synthetic data where we control the sample sizes of 2 protected groups and the residuals' distribution (normal and binomial, e.g., Fig. 1). The experimental setup compares 25 distinct distribution pairings, including normal-to-normal, normal-to-binomial, and binomial-to-binomial comparisons. Each pairing is evaluated across all 25 combinations of 5x5 sample sizes for each protected group (i.e., 20, 50, 100, 500, and 1000 data points). We evaluated 6 state-of-the-art statistical tests: Levene, one-way ANOVA, Brown-Forsythe, F-ratio, Fligner-Killeen and Bartlett. For each experimental condition, we generated 20 000 random samples and use 2 criteria to determine the performance of the tests: the percentages of significant p-values (< 0.5), and the median p-value.



(a) Homoscedastic Normal (b) Heteroscedastic Binomial

Fig. 1: Scatterplot of residuals versus predicted values with a smoothed distribution curve of the residual distribution shown on the left.

2 Related Work

The statistical tests covered in this study (i.e. Levene, one-way ANOVA on the squared residuals, Brown-Forsythe, F-ratio, Fligner-Killeen and Bartlett) have been covered by many comparative studies. These studies have repeatedly shown that no single test is uniformly reliable under all data conditions, particularly when assumptions such as normality and balanced sample sizes are violated [2,4,11,17].

Bartlett's test, F-ratio test, and methods based on ANOVA embody traditional methods which are quite powerful in the presence of balanced group sizes and normality of residuals. However, these tests tend to break down in the presence of unbalanced designs. More robust methods such as Levene's test, the Brown-Forsythe test and the Fligner-Killeen test have been shown to better control Type I error when used on skewed or heavy-tailed distributions, particularly with unbalanced designs [3,4]. Sometimes however, especially when sample sizes are small or slight variance differences are present, advances in Type I error control come at the expense of higher Type II error. [11].

In the context of fairness, the equilibrium of Type I and Type II error rates is very important. An excessive Type I error rate (false positives) could lead to statistically unsupported conclusions, which may reduce trust in the fairness assessment. On the other hand, a high Type II error rate (false negatives) may not detect significant differences that require attention. Most comparative studies report power $(1 - \beta)$ as a measure of sensitivity (where β represents the Type II error rate), but few analyze how both error rates change simultaneously under numerous interrelated and interacting conditions, including sample size imbalance, skewness, and distribution shape [5,14].

Several works have examined the effect of sample size on variance testing, showing that common rules of thumb for the Central Limit Theorem, such as $n \geq 30$ being "large enough" for normal approximation, are not universally valid [9]. These approximations can fail when data is skewed, leading to unstable p-values and inflated error rates even with moderate sample sizes. Although nonparametric and median-based methods have been proposed to improve robustness under such conditions, they have mainly been tested on continuous and unbounded outputs [12], overlooking the question of how well they work with bounded regression outputs.

Our research contributes to the literature in the following areas. To begin, we assess both Type I and Type II error rates for six common tests while systematically differing sample sizes, residual distributions, and group imbalances. In addition, we consider bounded regression outputs, which are targets constrained to the interval [0, 1]. This bounded case has not received significant direct attention in previous work. Lastly, instead of focusing solely on maximizing the power of tests, we examine the joint stability of p-values and detection rates under suboptimal conditions, highlighting these features as essential to trustworthy and fair inference in practical applications.

3 Methodology

In this section, we provide an overview of the synthetic residual distributions used in the experiments, followed by a description of the experimental set-up and the statistical tests evaluated. We first introduce the different distribution types and their parameter variations, after which we explain how sample sizes and pairwise comparisons were constructed. Finally, we describe the statistical tests and performance metrics used to assess heteroscedasticity under these conditions.

Note that in this study we focus on comparisons between two sensitive groups (e.g., male versus female), reflecting the most common setting in fairness assessments of regression models. All experiments therefore evaluate heteroscedasticity between two groups at a time. Extending this framework to more than two groups (e.g., multiple gender identities) would require additional considerations, as performing multiple pairwise comparisons can inflate Type I error rates and complicate interpretation.

3.1 Data Overview

To investigate the behavior of six statistical tests under varying conditions, we generated synthetic datasets simulating continuous distances to ground truth values. All values were constrained to lie within the interval of these ground truth values [0,1]. We constructed a total of 13 synthetic residual distributions: six normal and seven binomial. For the normal distributions, six cases were created by varying the mean and/or standard deviation of a baseline normal distribution. These include a baseline case, two variance-shift cases, two mean-shift cases, and one combined variance-and-mean-shift case.

The binomial data was generated as a mix of two normal distributions. In each binomial case, 80% of the data points came from one normal distribution (with mean = 0.4 and std. dev = 0.1), while 20% came from the the other normal distribution with varying mean and variance. This setup creates a dominant "large bump" and a smaller secondary bump, resulting in right- or left-skewed distributions depending on the location of the smaller bump. Seven binomial cases were created, including baseline, variance-shift, mean-shift, and combined shift variants. For both distribution types, each case consisted of 50,000 samples generated using NumPy's np.random.normal(loc = mean, scale = std) function. Values falling outside the [0,1] interval were resampled. An overview of all distribution settings, including approximate variance and distribution shape, is provided in Table 1.

As shown in 1, the left-skewed base distribution is the mirror image of the right-skewed base. Because values outside [0,1] are resampled, their total variances differ slightly (about $\approx 2.8\%$), and this difference is detectable. From a fairness standpoint, mass asymmetry (i.e., more points on the left vs. right) matters; therefore, when comparing right- and left-skewed cases, we treat them as having different variances. For the binomial mean shift cases we see that moving the small bump in turn also increases the variance (see Table 1). This results

Table 1: Overview of distribution settings. Note: standard deviation refers to the large bump for normal distributions and to the small bump for binomial distributions.

Name (Description)	**Mean (big)**	**Mean (small)**	**Std. Dev.**	**Approximate Distribution Shape**	**Variance ($\times 10^{-2}$)**
Normal Distributions					
norm_base (baseline)	0.40	–	0.10		1.00
norm_std+.05 (std ↑)	0.40	–	0.15		2.20
norm_std+0.1 (std ↑↑)	0.40	–	0.20		3.53
norm_m+.05 (mean →)	0.45	–	0.10		1.00
norm_m+.1 (mean →→)	0.50	–	0.10		0.99
norm_std&m+.05 (mean →, std ↑)	0.45	–	0.15		2.25
Binomial Distributions					
bin_RSk_base (small bump right baseline)	0.40	0.60	0.10		1.64
bin_RSk_stdS+.05 (small bump std ↑)	0.40	0.60	0.15		1.86
bin_RSk_stdS+.1 (small bump std ↑↑)	0.40	0.60	0.20		2.08
bin_RSk_mS+.05 (small bump mean →)	0.40	0.65	0.10		1.99
bin_RSk_mS+.1 (small bump mean →→)	0.40	0.70	0.10		2.41
bin_RSk_stdS&mS+.05 (small bump mean →, small std ↑)	0.40	0.65	0.15		2.21
bin_LSk_base (small bump left baseline)	0.40	0.20	0.10		1.59

in the mean shift cases for the binomial distributions having a higher variance than the standard deviation shift cases.

3.2 Experimental Set-up

To evaluate test performance across different data conditions, we considered five sample sizes: 20, 50, 100, 500, and 1000. For each distribution case and sample size, we performed 20,000 independent sampling runs using unique random seeds. All pairwise combinations of the five sample sizes were evaluated, resulting in 25 distinct sample-size configurations (e.g., group sizes of 50 and 1000). In total, we defined 25 distinct distribution comparisons, covering:

- Six comparisons between normal cases (including the base case against itself)
- Seven comparisons between normal and binomial cases
- Six comparisons between right skewed binomial cases
- Six comparisons between right skewed binomial and left skewed binomial cases

Each comparison was evaluated across all 25 sample-size combinations, resulting in a comprehensive assessment of test behaviour under both balanced and unbalanced conditions.

3.3 Statistical Tests

We used the following six statistical tests, each with a distinct set of assumptions, to determine if two samples originate from populations with equal variances:

- **F-ratio test** : A classical test for comparing two variances. It computes the ratio of the two sample variances and evaluates it using the F-distribution. *Assumptions*: Both samples are drawn from normal distributions and are independent. This test is sensitive to violations of normality, making it less robust under skewed or heavy-tailed data . [8,10]
- **One-way ANOVA**: A one-way ANOVA is performed on the squared residuals from the mean of the group. Unequal variances will lead to significant differences in the average squared deviation between groups. *Assumptions*: Normality and homogeneity of variances across groups. [6,10]
- **Levene's test**: Assesses heteroscedasticity by transforming the data into absolute deviations from the mean and then applying a standard one-way ANOVA to those deviations. *Assumptions*: Less strict than the F-test; it is more robust to non-normality, especially when distributions are symmetric. [1,10]
- **Brown–Forsythe test**: A modification of Levene's test that replaces the mean with the median when computing deviations. This makes it more robust to skewness and outliers, improving performance in non-normal conditions. *Assumptions*: Does not assume normality and handles heavy-tailed or asymmetric distributions better than Levene's test. [1,16]

- **Fligner–Killeen test**: A non-parametric test based on ranked data. It transforms the data using a rank-based method and evaluates variance homogeneity using a chi-squared statistic. *Assumptions*: Fully non-parametric, does not require normality or equal sample sizes. [10,13]
- **Bartlett's test**: A classical parametric test for the homogeneity of variances. It compares sample variances across groups using a likelihood-ratio test. *Assumptions*: Strictly assumes normal distributions. It is powerful under normality but very sensitive to deviations from it (e.g., skewness or kurtosis), which can inflate the Type I error rate. [10,16]

Per test, comparison type and experiment the following metrics were noted: median p-value, first and third quantiles (Q1 and Q3), interquantile range (IQR), minimum and maximum p-values and the percentage of tests that are significant. The median p-value represents the most common outcome of a test under a given condition, while Q1, Q3 and the IQR give an overview of the variability and stability of outcomes. Minimum and maximum p-values give an overview of extreme values that may arise due to sampling variability. The percentage of significant tests directly quantifies how often heteroscedasticity is detected, allowing comparison of sensitivity and error behaviour across tests and conditions. These measures were combined to provide a comprehensive and robust overview of test performance.

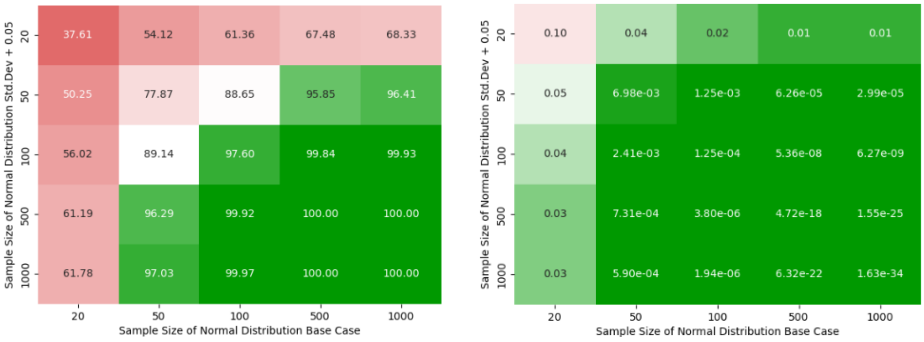
3.4 Criteria

In order to make an easy distinction between values and their performance, we defined three performance categories: Optimal (O), Acceptable (A) and Poor (P). Significance percentages and median p-values for each experiment and case are assigned these categories depending on whether there is a significant difference in variance, see 2 for the classification rules.

	Optimal (O)	Acceptable (A)	Poor (P)
% of significant p-value - in presence of heretoscedasticity	≥ 99.00	$98.00 \leq x < 99.00$	< 98.00
Significance % (no significantly different variance)	≤ 5.00	$5.00 < x \leq 7.00$	> 7.00
Median p-value (significantly different variance)	≤ 0.01	$0.01 < x \leq 0.05$	> 0.05
Median p-value (no significantly different variance)	≥ 0.10	$0.06 \leq x < 0.10$	< 0.06

Table 2: Thresholds for Optimal (O), Acceptable (A), and Poor (P) by metric and variance condition.

To illustrate the classification process, we consider the case where a normal base distribution is compared to a normal distribution with its standard deviation increased by 0.05. Figure 2 shows the results of the Bartlett test for the significance percentage and median p-value. We classified both measures separately using the thresholds in Table 2. Table 3 shows the resulting classifications for each criterion.



(a) Significance Percentage (b) Median p-value

Fig. 2: Bartlett Test Results (norm_base VS norm_std+05) Across Experiments

(a) Classified Significance Percentage

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	O	O
500	P	P	O	O	O
1000	P	P	O	O	O

(b) Classified Median p-value

	20	50	100	500	1000
20	P	A	A	A	A
50	A	O	O	O	O
100	A	O	O	O	O
500	A	O	O	O	O
1000	A	O	O	O	O

Table 3: Classification Example for Bartlett norm_base VS norm_std+.05 Results

To reach an overall evaluation method, we merge the two criteria following the combination rules in Table 4a. We decided on a conservative approach (i.e., displaying the lower of the two classes) with the number of "+" signs indicating the distance to the higher class. For example, if both classes are P, the result is P; if they are (P, A) the result is P+, and if (P, O) it is P++. Applying these rules to the median p-value and significance percentage classification tables (3) results in Table 4b. For instance, at $n = 20$ the significance percentage was classified as Poor and the median p-value as Poor, resulting in P, whereas at $n = 50$ the significance percentage was Acceptable and the median p-value Optimal, resulting in $P++$.

To further quantify these results we also gave each merged classification table a numerical score, using the labels present. These values are presented in Table 5 and the score of a table is calculated by taking the sum of these values. Resulting in score values ranging between 0 and 125 with 125 being optimal.

(a) Merging Classification Rules

	P	A	O
P	P	P^+	P^{++}
A	P^+	A	A^+
O	P^{++}	A^+	O

(b) Merged Classification Results Bartlett

	20	50	100	500	1000
20	P	P^+	P^+	P^+	P^+
50	P^+	P^{++}	P^{++}	P^{++}	P^{++}
100	P^+	P^{++}	P^{++}	O	O
500	P^+	P^{++}	O	O	O
1000	P^+	P^{++}	O	O	O

Table 4: Merging Classification Rules and Application Example on Bartlett (norm_base VS norm_std+.05) Results

Label	Value
P	0
P^+	1
P^{++}	2
A	3
A^+	4
O	5

Table 5: Overview of Classification Labels and their Assigned Values

4 Results

This section presents the results across all distribution pairs, sample-size combinations and heteroscedasticity tests. Results are reported using the merged classification scores, introduced in Section 3.3. This section is organised in four main distribution pairing categories: normal vs normal, normal vs binomial right-skewed, binomial right-skewed vs binomial right-skewed and binomial left-skewed vs binomial right-skewed. In each of these sections the general behaviour is described first, followed by an analysis of behaviour per test. Score comparison figures, presented at the start of each section, summarise overall performance per test. Expected test behaviour is interpreted according to the expected heteroscedasticity matrix (see Table 6).

4.1 Normal VS Normal

General behaviour across tests:

Expected behaviour overall, with comparable values across tests for the baseline and mean-shift cases. For variance-shift and variance&mean-shift cases, all tests show degradation as sample sizes decrease. Median and IQR follow similar patterns, while Min–Max values diverge slightly between tests under certain conditions, but without a consistent trend. All tests behave optimally in the normal–normal baseline and degrade mainly under variance shifts.

For all mean-shift cases, results show only A+ and O classifications, with differing patterns where they shift from A+ to O. One-way ANOVA achieves

	norm_base	norm_std+.05	norm_std+.1	norm_m+.05	norm_m+.1	norm_std&m+.05	bin_RSk_base	bin_RSk_stdS+.05	bin_RSk_stdS+.1	bin_RSk_mS+.05	bin_RSk_mS+.1	bin_RSk_stdS&mS+.05	bin_LSk_base
norm_base	×	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓
norm_std+.05	✓	×	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
norm_std+.1	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
norm_m+.05	×	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓
norm_m+.1	×	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓
norm_std&m+.05	✓	×	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
bin_RSk_base	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×
bin_RSk_stdS+.05	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
bin_RSk_stdS+.1	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
bin_RSk_mS+.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
bin_RSk_mS+.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
bin_RSk_stdS&mS+.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
bin_LSk_base	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×

Table 6: Matrix of Expected Heteroscedasticity for all Pairwise Distribution Comparisons. Expected Heteroscedasticity is Indicated with a Green Checkmark ✓.

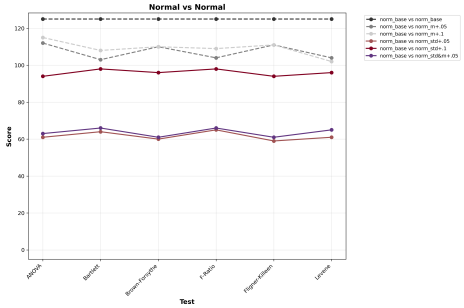


Fig. 3: Score Comparisons per Test: Normal vs Normal Cases. Grey Lines Indicate no Heteroscedasticity Expected.

the highest scores in both +.05 (112) and +.1 (115) mean-shift cases, followed by Fligner–Killeen (111 for both). The lowest scores are for Bartlett (+.05: 103, +.1: 108) and Levene (+.05: 104, +.1: 102). Levene's score decreases as the mean-shift increases. F-ratio is similar to Levene in the +.05 case (104) but higher for +.1 (109).

For all tests, applied on the +.05 variance-shift case, results are optimal for all sample size combinations of 100 or higher, except when both groups are 100. In the +.1 case, ANOVA, Bartlett and F-ratio are optimal (O) for any sample size combination of 50 or higher, while Brown–Forsythe, Fligner–Killeen and Levene are optimal for the same combinations except when both distributions have a sample size of 50. For the combined variance&mean-shift case, all tests show optimal results for 100 or higher sample size combinations, except when both distributions have a sample size of 100.

F-ratio and Bartlett obtain the highest scores across all variance(&mean)-shift cases. For the +.1 variance-shift and combined variance&mean-shift cases, they have equal values (98 and 66 respectively). For the +.05 variance-shift case, F-ratio performs slightly better (65) than Bartlett (64). The lowest-performing test is Fligner–Killeen, with scores of 59, 94, and 61 for +.05, +.1 and variance&mean-shift cases respectively.

One-way ANOVA: One-way ANOVA performed as expected in the baseline case and achieved a maximum classification score of 125. The highest seen significance percentage across sample sizes was 3.65%, which is the lowest across tests. For the mean-shift cases (+.05 and +.1), it performs better when the sample sizes of two groups are not equal, especially when the base distribution is at the lower end of the sample size range and the mean-shift distribution at the higher end of the sample size range. In variance-shift cases, ANOVA gives the best results of all tests when the base distribution has size 500 or 1000 and the variance-shift distribution 20. However, it gave the worst results of all tests when the base distribution has sample size 20 and the variance-shift distribution 50 or higher.

Bartlett: Bartlett behaves as expected for the baseline (classification score: 125) and mean-shift cases. For the mean-shift cases results are largely unaffected by differences in sample size, however Bartlett has the least optimal (O) values for the +.05 mean-shift case. For variance-shift cases, Bartlett has the best performance together with F-ratio, but better performance when the base distribution has sample size 20 and the variance-shift distribution 50 or higher.

Brown-Forsythe: Brown-Forsythe behaves as expected for the baseline (classification score: 125) and mean-shift cases. For the mean-shift cases, performance is highest when both groups are at the lower end of the sample-size range (20), based on significance percentages. For variance-shift and variance&mean-shift cases performance drops sharply and is the worst after Fligner–Killeen, especially when both groups are at the lower end of the range.

F-ratio: Behaves as expected for baseline (classification score: 125) and mean-shift cases. For the mean-shift cases performance is not influenced a lot by differences in sample sizes. In variance-shift cases it has the highest overall performance together with Bartlett and slightly better performance than Bartlett when the variance-shift distribution is 20 and the base at the mid-to-higher end of the sample size range (50–1000). It is also slightly better than Bartlett in significance percentages and median p-values for the +.05 variance-shift case, but slightly worse for the +.1 case.

Fligner-Killeen: Fligner-Killeen behaves as expected for the baseline (classification score: 125) but has the highest maximum significance percentage (4.65%) across tests. The test performs strongly for mean-shift cases when both groups are at the lower end of the sample size range with many optimal (O) values. However, it gives the lowest performance for variance&mean-shift cases when both distributions are at the lower end of the sample size range.

Levene: Levene baseline behaves as expected with a maximum classification score of 125. For the mean-shift cases it has a slightly higher significance percent-

age values than other tests. This behaviour is not reflected in the other metrics (e.g. median p-values or IQR). Levene also has the least optimal (O) values for mean-shift case +.05.

4.2 Normal VS Binomial

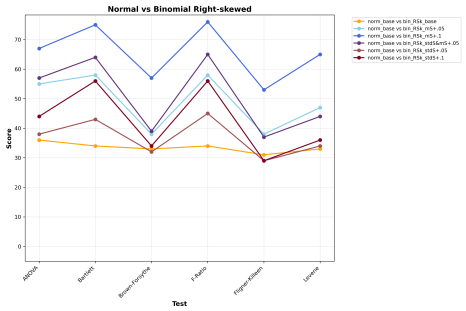


Fig. 4: Score Comparisons per Test: Normal vs Binomial Right-Skewed Cases

General behaviour across tests: When binomial variance is low and group sizes are at the higher end of the sample-size range, performance across tests is high. As binomial variance increases or group sizes decrease, performance degrades. Median and IQR follow similar patterns, while Min–Max values diverge slightly between tests, though without a consistent trend. Overall, behaviour follows the same trend as in the normal–normal cases but degrades under increased binomial variance. For both normal base case VS left-skewed base case and normal base case VS right-skewed base case comparisons, optimal values are present for any combination of sample sizes of 500 or higher. Scores across all tests are slightly lower for the left-skewed comparison, with one-way ANOVA performing the best for both right-skewed (score: 36) and left-skewed (score: 34). Fligner–Killeen performs the worst in both comparison cases, with scores of 31 (right-skewed) and 29 (left-skewed).

For mean-shift cases, Bartlett and F-ratio perform best, with equal scores for the mean-shift +.05 case and F-ratio performing slightly better for +.1. The lowest scores are for Fligner–Killeen and Brown–Forsythe, both 38 for the mean-shift +.05 case, while for +.1 Fligner–Killeen performs worse (53) than Brown–Forsythe (57).

For variance-shift cases, Bartlett and F-ratio again perform best, both scoring 56 for the variance-shift +.1 case. For the +.05 case, F-ratio (45) performs slightly better than Bartlett (43). The lowest-performing test is Fligner–Killeen, with 29 for both cases. For variance&mean-shift cases, F-ratio again performs best (65) and Fligner–Killeen the worst (37).

Optimal (O) values are present for mean-shift +.05 when:

- ANOVA and F-ratio: for any combination of 100 or higher sample sizes, except when both groups have size 100
- Brown–Forsythe, Fligner–Killeen and Levene: for any combination of sample sizes of 500 or higher

– Bartlett: for any combination of 100 or higher sample sizes, except when both groups have size 100 and when (norm base, Rsk binomial +.05 mean) = (500, 100)

Optimal (O) values are present for mean-shift +.1 when:

– ANOVA, Levene and Brown–Forsythe: for any combination of 100 or higher sample sizes, except when both groups have size 100
– Fligner–Killeen: for any combination of sample sizes of 500 or higher, and for (norm base, bin +.1) = (1000, 100) or (100, 1000)
– F-ratio: for any combination of 100 or higher sample sizes
– Bartlett: for any combination of 100 or higher sample sizes, and for (norm base, bin +.1) = (50, 1000)

For all variance-shift +.05 cases, optimal values occur for any combination of sample sizes of 500 or higher. The same pattern holds for +.1 cases for ANOVA, Levene, Fligner–Killeen and Brown–Forsythe. For Bartlett and F-ratio, optimal values occur for any combination of sample sizes of 500 or higher and for (norm base, bin +.1) = (100, 500) or (100, 1000).

For variance&mean-shift cases, Brown–Forsythe, Fligner–Killeen and Levene show optimal values for any combination of sample sizes of 500 or higher. ANOVA, Bartlett and F-ratio have optimal values for any combination of 100 or higher sample sizes, except when both groups have size 100.

One-way ANOVA: One-way ANOVA is stable for large equal groups, with performance often close to Bartlett and F-ratio. For the baseline (binomial variance fixed), the significance rate ranges from 16% (when both 20) to 3–13% otherwise. For mean-shift cases, performance is highest when the binomial group has size 20 and the base distribution is at the higher end of the sample-size range, which is most evident in the median and IQR values. For variance&mean-shift cases, performance decreases when both groups are at the lower end of the range: when norm = 20 and binomial = 20, the significance rate rises up to 12.6% as the binomial variance increases.

Bartlett: Bartlett behaves as expected across all baseline and mean-shift cases and performs consistently well together with F-ratio when variances differ. For variance&mean-shift cases, Bartlett is highly stable and performs better than F-ratio when norm = 20 and binomial = 50–1000. Min–Max values are very comparable, with small differences between Bartlett and F-ratio depending on the metric (median or %).

Brown–Forsythe: Brown–Forsythe behaves as expected for the baseline and mean-shift cases.For variance&mean-shift cases, performance is weak, ranking just above Fligner–Killeen, especially when both groups are at the lower end of the sample-size range.

F-ratio: F-ratio behaves as expected for the baseline and mean-shift cases. For variance differences, it performs among the best, especially when norm = 20 and binomial = 50–1000, where median and IQR values are balanced. In certain variance&mean-shift cases, F-ratio performs slightly better than Bartlett in significance percentages, although Bartlett maintains a narrower IQR.

Fligner-Killeen: Fligner–Killeen consistently shows the lowest performance across all conditions. It performs particularly poorly for variance&mean-shift cases, with significance rates dropping substantially across sample-size combinations.

Levene: Levene behaves as expected for the baseline and mean-shift cases, performing slightly below Bartlett and F-ratio but still within reasonable limits. For variance&mean-shift cases, it performs worse than both Bartlett and F-ratio, especially when the base group is at the higher end and the binomial group at the lower end of the sample-size range. Overall, performance is poor as variance differences increase and sample sizes decrease.

4.3 Right-skewed Binomial VS Right-skewed Binomial

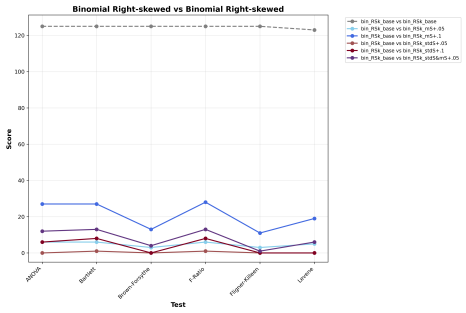


Fig. 5: Score Comparisons per Test: Binomial Right-Skewed vs Binomial Right-Skewed Cases. Grey Lines Indicate no Heteroscedasticity Expected.

General behaviour across tests: Overall performance is lower than in the normal–normal and normal–binomial cases, particularly when both groups are small. Many conditions show significance values below 10%, even when sample sizes are large, though sample size combinations such as 500/500, 500/1000 and 1000/1000 often reach a significance percentage close to 100%. Performance is generally higher when the right-skewed base sample size is large and the variance-shift or mean-shift has a smaller sample size, rather than the reverse. Median and IQR shrink substantially when both groups are small, while Min–Max values sometimes diverge under extreme conditions, such as very small sample sizes or large variance differences, but without a consistent direction.

All right-skewed baseline comparisons show optimal classifications (O) and maximum scores of 125, except Levene, which shows A+ when one group is 20 and the other 1000 (in either direction), resulting in a score of 123. For the mean-shift +.05 case, overall performance is low across tests, with ANOVA, Bartlett and F-ratio obtaining the highest scores (6). For mean-shift +.1, ANOVA and Bartlett score 27 and F-ratio slightly higher (28). For the same conditions, Brown–Forsythe and Fligner–Killeen are the lowest, with scores of 3 for the +.05 case and 13 and 11 respectively for +.1. For variance-shift +.05, all tests except Bartlett and F-ratio have scores of 0, while Bartlett and F-ratio reach 1. For variance-shift +.1, Fligner–Killeen, Brown–Forsythe and Levene remain at 0, ANOVA reaches 6, and Bartlett and F-ratio again perform best with scores of 8.

For variance&mean-shift cases, Fligner–Killeen performs the worst (score = 1), while Bartlett and F-ratio again perform best (score = 13).

For the mean-shift +.05 case, no optimal values are present, with P++ as the highest classification label.

For the mean-shift +.1 case:

- Fligner–Killeen: no optimal values, but A+ when both groups have size 1000

- Brown–Forsythe and Levene: optimal when both groups have size 1000; Levene also A+ for combinations (500, 1000) and (1000, 500)
- Bartlett and F-ratio: optimal for any combination of sample sizes of 500 or higher
- One-way ANOVA: optimal for any combination of sample sizes of 500 or higher, except when both groups have size 500, where it shows A+

For variance-shift cases, no optimal values are present. The highest classification label is P+ for the +.05 case and P++ for the +.1 case. For variance&mean-shift cases, Brown–Forsythe, Fligner–Killeen and Levene show no optimal values either, with the highest labels being P+ for Fligner–Killeen and P++ for the others. One-way ANOVA, Bartlett and F-ratio show optimal values only when both groups have size 1000.

One-way ANOVA: One-way ANOVA behaves as expected in the baseline case, showing the lowest overall maximum significance rate (3.53%) across tests. For mean-shift cases it performs well when sample sizes are large, comparable to Bartlett and F-ratio, and performs best when the base distribution is at the lower end of the sample-size range (20) and the mean-shift distribution large. For variance-shift cases it is highly sensitive to size imbalance, performing very poorly when the base is 20 and the variance-shift distribution 1000 (as low as 0.16%), but reaching very high values when both groups are large (up to 99.07% when both 1000).

Bartlett: Bartlett behaves as expected for the baseline case. For mean-shift cases it performs best together with F-ratio, especially when the base distribution is small and the mean-shift distribution large. For variance-shift cases, Bartlett consistently ranks among the best performers, with slightly better median values than F-ratio when the base is 20 and the variance-shift group ranges between 50 and 1000.

Brown–Forsythe: Brown–Forsythe behaves as expected for the baseline case. For both mean-shift and variance-shift cases, performance ranks just above Fligner–Killeen and declines sharply as sample sizes decrease.

F-ratio: F-ratio behaves as expected for the baseline case. For mean-shift cases it performs best together with Bartlett, especially when the base distribution is small (20) and the mean-shift distribution large, with the strongest median values. For variance-shift cases it performs best together with Bartlett when the base distribution is 20 and the variance-shift distribution 50–1000, across significance percentages, IQR and median. Min–Max values show minor divergence under certain conditions, but not systematically.

Fligner–Killeen: Fligner–Killeen shows the lowest performance across all conditions. For mean-shift and variance-shift cases, it performs consistently poorly, with significance percentages ranging from 4–7% depending on the sample-size combination.

Levene: Levene behaves as expected for the baseline case but shows the highest significance rate (5.08%) across tests. For mean-shift cases it performs moderately when both groups are small (20), better than Fligner–Killeen but lower than Bartlett and F-ratio. For variance-shift cases it performs poorly overall, of-

ten below 15%, but improves when both groups are large (500 or higher). Median values remain weaker than those of Bartlett and F-ratio across all sample-size combinations.

4.4 Left-skewed Binomial VS Right-skewed Binomial

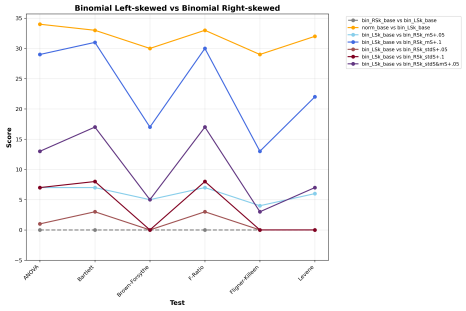


Fig. 6: Score Comparisons per Test: Binomial Left-Skewed vs Binomial Right-Skewed Cases. Grey Lines Indicate no Heteroscedasticity Expected.

General behaviour across tests: This comparison shows less consistent behaviour across tests compared to the normal–normal, normal–binomial, and right-skewed binomial–right-skewed binomial cases. Performance differences are minimal when both groups are large, but values degrade rapidly as sample sizes decrease or under increased variance. Performance tends to improve when the base group is large and the shifted group (mean or variance) small, rather than the reverse. Left-skewed base versus right-skewed base baseline comparisons show poor performance for all tests (label P) with a minimal score of 0.

The overall classification and score patterns reflect these trends: For the mean-shift +.05 case, overall performance is low across tests, with one-way ANOVA, Bartlett, and F-ratio obtaining the highest scores (7). For the mean-shift +.1 case, ANOVA scores 29, F-ratio 30, and Bartlett performs best with 31. The lowest performance is observed for Fligner–Killeen, with scores of 4 for the +.05 case and 13 for +.1.

For variance-shift cases, Brown–Forsythe, Fligner–Killeen and Levene have scores of 0 for both +.05 and +.1 cases. ANOVA performs moderately with scores of 1 for +.05 and 7 for +.1, while Bartlett and F-ratio perform best, scoring 3 and 8 respectively. For variance&mean-shift cases, Fligner–Killeen again performs worst (score = 3), while Bartlett and F-ratio perform best (score = 17).

For mean-shift cases, no optimal values are present for the +.05 condition, with P++ as the highest classification label. For the mean-shift +.1 condition, Bartlett, F-ratio and one-way ANOVA show optimal values for any combination of sample sizes of 500 or higher, while Brown–Forsythe and Fligner–Killeen are optimal only when both groups have size 1000. Levene performs optimally for any combination of sample sizes of 500 or higher, except when both groups have size 500.

For variance-shift cases, no optimal values are present. For the +.05 condition, P+ is the highest label for one-way ANOVA, Bartlett and F-ratio, and P

for all other tests. For the +.1 condition, the highest label is P++ for one-way ANOVA, Bartlett, Levene and F-ratio, and P for the remaining tests.

Variance&mean-shift cases:

- F-ratio and Bartlett: optimal when both groups have size 1000, and A+ when (Lsk base, Rsk variance&mean-shift) = (500, 1000)
- One-way ANOVA: optimal only when both groups have size 1000
- Levene and Brown–Forsythe: no optimal values (highest P++)
- Fligner–Killeen: no optimal values (highest P+)

One-way ANOVA: One-way ANOVA shows the lowest significance rate in the baseline case (3.05%), the smallest across all tests in this comparison. The significance percentage increases modestly (up to 6.9%) when both the left-skewed and right-skewed groups have sample sizes of 1000. For mean-shift cases, it performs stably when group sizes are large but poorly when the base group has size 20 (up to 12%). For variance-shift cases, performance decreases sharply as sample sizes decrease, performing very poorly when the base is 20 (as low as 0.36%) but very well when both groups have size 1000 (up to 99.75%).

Bartlett: Bartlett behaves reliably across baseline and mean-shift cases. It performs best together with F-ratio when the base group is small and the mean-shift group large. For variance-shift cases, Bartlett remains among the best performers, particularly when the base is 20 and the variance-shift group 50–1000. Min–Max values follow similar patterns, with negligible differences between Bartlett and F-ratio.

Brown–Forsythe: Brown–Forsythe shows slightly lower baseline values than other tests when the left-skewed group is 20 and the right-skewed group larger, with average significance around 5.69%. For mean-shift and variance-shift cases, performance is low, ranking above Fligner–Killeen but improving slightly when sample sizes are large.

F-ratio: F-ratio behaves as expected in the baseline case, with lower values when the left-skewed group is small and the right-skewed group large, but higher when the base is large and the comparison small. For mean-shift cases, it performs best together with Bartlett, particularly when the base group is small and the mean-shift group large, across significance, median and IQR values. For variance-shift cases, F-ratio performs best together with Bartlett when the base has size 20 and the variance-shift group ranges between 50 and 1000, showing similar median and Min–Max values.

Fligner–Killeen: Fligner–Killeen shows the lowest performance across all conditions. In the baseline case, values are lower than for other tests when the left-skewed group is small and the right-skewed group large (average around 5.82%). For mean-shift and variance-shift cases, it consistently performs worst, with significance percentages ranging from 0.5–9%.

Levene: Levene behaves as expected for the baseline case but is more sensitive to variance shifts than the other tests, with average significance around 6.57% (higher than ANOVA). For mean-shift cases, performance is generally low, slightly above Fligner–Killeen but below Bartlett and F-ratio, performing best when both groups are small (20), though still only reaching 8.29%. For

variance-shift cases, performance is moderate when both groups are large (e.g., 1000/1000 = 92.44%), but declines sharply as sample sizes decrease.

5 Discussion

5.1 Summary of Findings

Across varying experimental conditions, two tests consistently were the most robust: F-ratio and Bartlett. These tests had the most reliable performance, especially under non-normal conditions and variance shifts. They remain relatively stable even when group sizes are unbalanced. While Brown–Forsythe and Fligner–Killeen overall perform the worst, especially when both groups are at the lower end of the sample-size range, under variance shifts, or when distributions deviate from normality. Levene tends to sit in the middle: it performs reasonably well in baseline and mean-shift c, but its sensitivity to variance differences decreases as distributions become more skewed or sample sizes diverge. One-way ANOVA performs comparably to Bartlett and F-ratio when at least one group is at the larger end of the sample-size range or when both groups have high sample sizes.

In general, performance improves as sample sizes increase and decreases when distributions deviate from normality. This trend is visible in the binomial experiments, where performance decreases compared to the normal experiments. When both groups are at the lower end of the sample-size range, significance rates and classification scores decrease significantly and results become less consistent across tests. Moreover, performance tends to be higher when the wider distribution has the larger sample size and the narrower distribution the smaller one. In this configuration, the narrower distribution's data points are more likely to fall within the range of the wider distribution, leading to more stable test behaviour

5.2 Fairness Note: Left- VS Right-skewed Binomial Comparisons

Although the left-and right-skewed binomial distributions differ by only about 2.8% in variance, we consider them to have significantly different variance which in turn adjusts the way we classify these results. The small bump to the left or right affects which distribution occupies the upper or lower tail, and thus which sample is more likely to contain extreme values. From a fairness perspective this difference is meaningful, even a minor variance gap or location difference can produce unequal output (e.g. classification). This highlights that a purely statistical approach to measuring variance is not sufficient to evaluate fairness. Tests can fail to detect a difference in variance, while one group suffers from a systematic disadvantage due to location difference (i.e. skew direction).

5.3 Practical Advice: When to Use Which Test

Bartlett and F-ratio are the most reliable across all experimental conditions, particularly in cases of different variance. F-ratio performs slightly better when the

distribution with increased variance (i.e. the wider distribution) has a smaller sample size and the base distribution is at the larger end of the sample size range compared to the reverse. Conversely, Bartlett has better performance when the base distribution has a smaller sample size and the increased variance distribution is at the larger end of the sample size range.

One-way ANOVA is a good alternative if both distributions have large and approximately equal sample sizes. Levene performs decently for mean-shift cases but becomes less reliable as distributions deviate from normality or when sample sizes are unequal. Brown-Forsythe and Fligner-Killeen show the least robust performance across experiments and are not recommended in general. For applications with non-normal distributions, small or unequal sample sizes we suggest referencing our Appendix for case-specific overview tables.

5.4 Future Work

Future research could further test the boundaries of these tests under other non-normal distributions with heavier tails or multimodal shapes. Additional examples for left- and right-skewed cases would also be valuable, as only one left-skewed case was considered in this study. More varied binomial cases, for instance with a more defined bump, could help clarify how distributional shape affects test behaviour. Adding more sample sizes when testing different distributions would also allow for more informed test selection.

The use of the Earth Movers's Distance (EMD) should be considered as an addition to heteroscedasticity tests. EMD compares distributions by quantifying the minimal amount of effort needed to transform one distribution into another [15]. This is particularly relevant for fairness as small changes in distributional mass can determine which subgroup of the population is more likely to fall into extreme regions of the population distribution. Another direction for future work could be to model heteroscedasticity using polynomial regression. This method can capture complex error patterns that remain undetected by linear models or standard variance tests [7]. The integration of multiple such methods would create a more complete framework for detecting heroscedasticity in a fairness setting.

All tests in this study either consider absolute or squared residuals. However, considering the sign of the residuals can also provide useful information related to fairness, because it reveals whether a group is systematically overestimated or underestimated. The overall performance for small sample sizes, which can be present for a subgroup of the population, also highlights the need for new statistical- and evaluation methods for small sample sizes specifically.

In addition, future work should extend the current two-group comparison set-up to settings involving more than two sensitive groups (e.g., multiple gender identities). While this study focuses on pairwise comparisons, assessing heteroscedasticity across multiple groups simultaneously raises methodological challenges, particularly related to inflated Type I error rates when performing multiple pairwise tests. Developing approaches that control error rates while remaining interpretable in multi-group fairness settings would be a valuable addition.

6 Conclusion

This paper compared six statistical tests for heteroscedasticity under both optimal and sub-optimal conditions in bounded regression problems. We investigated how differences in sample sizes, distribution type (binomial, normal), variance and mean affect test reliability when residuals fall within the [0,1] interval.

How is the performance of statistical tests for heteroscedasticity influenced by sample sizes and differences in the distribution of residuals? Across all experiments Bartlett and F-ratio performed the most reliably and showed the most robustness under non-normal conditions. Levene and one-way ANOVA showed adequate performance when both groups are large and approximately similar in size. Brown-forsythe and Fligner-Killeen were the least robust overall, especially for small sample sizes and skewed distributions. Across all statistical tests, small sample sizes and deviations from normality decreased performance significantly.

From a fairness perspective, these findings show that state-of-the-art tests for heteroscedasticity are not sufficient when considering non-normal distributions or small sample sizes. It is important to combine multiple approaches in such cases to ensure a fair determination of variance inequality. Possible approaches include adding the Earth's Mover's Distance to capture subtle shifts in distributional mass, visual inspection of the distribution and using polynomial modeling of residuals to identify more complex forms of heteroscedasticity.

References

1. Abdullah, N.F., Muda, N.: An overview of homogeneity of variance tests based on type I error rate and power of a test. Journal of Quality Measurement and Analysis **18**(3), 111–130 (2022)
2. Abdullah, N.F., Muda, N.: An overview of homogeneity of variance tests on various conditions based on type 1 error rate and power of a test. Journal of Quality Measurement and Analysis **18**(3), 111–130 (2022)
3. Adeleke, I., Oladeji, T.F., Adesina, O.S.: Exploring robust methods for testing equality of variances. International Journal of Statistics and Economics **19**(3) (2018), page numbers were not available in the document
4. Beyene, K.M., Bekele, S.A.: Assessing univariate and multivariate homogeneity of variance: A guide for practitioners. Mathematical Theory and Modeling **6**(5) (2016)
5. Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., Bendayan, R.: Non-normal data: Is anova still a valid option? Psicothema **29**(4), 552–557 (2017). https://doi.org/10.7334/psicothema2016.383
6. Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., Bendayan, R.: Non-normal data: Is ANOVA still a valid option? Psicothema **29**(4), 552–557 (2017). https://doi.org/10.7334/psicothema2016.383
7. Douma, M., Beauxis-Aussalet, E.: Modelling heteroscedasticity for fair regression using polynomial models. In: Proceedings of the 36th Benelux Conference on Artificial Intelligence (BNAIC 2024) (2024)
8. Gorsuch, R.L., Lehmann, C.: Chi-square and F ratio: Which should be used when? Journal of Methods and Measurement in the Social Sciences **8**(2), 58–71 (2017)
9. Islam, M.R.: Sample size and its role in central limit theorem (clt). Computational and Applied Mathematics Journal **4**(1), 1–7 (2018)
10. Katsileros, A., Antonetsis, N., Mouzaidis, P., Tani, E., Bebeli, P.J., Karagrigoriou, A.: A comparison of tests for homoscedasticity using simulation and empirical data. Communications in Statistics – Simulation and Computation **53**(1), 1–35 (2024). https://doi.org/10.29220/CSAM.2024.31.1.001
11. Katsileros, A., Antonetsis, N., Mouzaidis, P., Tani, E., Bebeli, P.J., Karagrigoriou, A.: A comparison of tests for homoscedasticity using simulation and empirical data. Communications for Statistical Applications and Methods **31**(1), 1–35 (2024). https://doi.org/10.29220/CSAM.2024.31.1.001
12. Manly, B., Francis, R.I.C.C.: Testing for mean and variance differences with samples from distributions that may be non-normal with unequal variances. Journal of Statistical Computation and Simulation **72**(8), 633–646 (2002). https://doi.org/10.1080/00949650213745
13. Manly, B.F.J., Francis, R.I.C.C.: Testing for mean and variance differences with samples from distributions that may be non-normal with unequal variances. Journal of Statistical Computation and Simulation **72**(8), 633–646 (2002). https://doi.org/10.1080/00949650213745
14. Ogbonna, C.J., Okenwe, I., Ifeanyichukwu, O.S.: Effect of sample sizes on the empirical power of some tests of homogeneity of variances. International Journal of Mathematics Trends and Technology (IJMTT) **65**(6) (2019)
15. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**(2), 99–121 (2000)
16. Yonar, A., Yonar, H., Demirsöz, M., Tekindal, M.A.: A comparative analysis for homogeneity of variance tests. Journal of Science and Arts **24**(2), 305–328 (2024). https://doi.org/10.46939/J.Sci.Arts-24.2-a06

17. Yonar, A., Yonar, H., Demirsöz, M., Tekindal, M.A.: A comparative analysis for homogeneity of variance tests. Journal of Science and Arts **24**(2), 305–328 (2024). https://doi.org/10.46939/J.Sci.Arts-24.2-a06

A Merged Classification Reports Normal vs Normal

A.1 Normal Base Case vs Normal Base Case

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(a) ANOVA, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(b) Bartlett, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(c) Brown-Forsythe, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(d) F-Ratio, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(e) Fligner-Killeen, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(f) Levene, Score: 125

Fig. 7: Merged Classification Results for Normal Base vs Normal Base, Average Score: 125.00

	20	50	100	500	1000
20	A^+	A^+	O	O	O
50	O	A^+	A^+	O	O
100	O	A^+	A^+	A^+	A^+
500	O	O	A^+	A^+	A^+
1000	O	O	O	A^+	A^+

(a) ANOVA, Score: 112

	20	50	100	500	1000
20	A^+	A^+	A^+	A^+	A^+
50	O	A^+	A^+	O	A^+
100	O	A^+	A^+	A^+	A^+
500	A^+	A^+	A^+	A^+	A^+
1000	A^+	A^+	A^+	A^+	A^+

(b) Bartlett, Score: 103

	20	50	100	500	1000
20	O	O	A^+	A^+	A^+
50	O	O	A^+	A^+	A^+
100	O	O	A^+	A^+	A^+
500	O	O	A^+	A^+	A^+
1000	A^+	A^+	A^+	O	O

(c) Brown-Forsythe, Score: 110

	20	50	100	500	1000
20	A^+	A^+	A^+	A^+	A^+
50	A^+	A^+	A^+	O	A^+
100	O	A^+	A^+	A^+	A^+
500	O	A^+	A^+	A^+	A^+
1000	O	A^+	A^+	A^+	A^+

(d) F-Ratio, Score: 104

	20	50	100	500	1000
20	O	O	O	A^+	A^+
50	O	O	A^+	A^+	A^+
100	O	O	A^+	A^+	A^+
500	O	O	A^+	A^+	A^+
1000	A^+	A^+	A^+	O	O

(e) Fligner-Killeen, Score: 111

	20	50	100	500	1000
20	A^+	A^+	A^+	A^+	A^+
50	A^+	A^+	A^+	A^+	A^+
100	O	A^+	A^+	A^+	A^+
500	O	O	A^+	A^+	A^+
1000	A^+	A^+	A^+	A^+	O

(f) Levene, Score: 104

Fig. 8: Merged Classification Results for Normal Base (column) vs Normal Mean Shift of 0.05 (row), Average Score: 107.33

	20	50	100	500	1000
20	A^+	O	O	O	O
50	O	A^+	A^+	A^+	A^+
100	O	O	O	A^+	A^+
500	O	O	O	A^+	A^+
1000	O	O	O	O	A^+

(a) ANOVA, Score: 115

	20	50	100	500	1000
20	A^+	O	A^+	O	A^+
50	A^+	A^+	A^+	A^+	A^+
100	A^+	O	O	A^+	A^+
500	A^+	A^+	O	O	A^+
1000	A^+	A^+	O	O	A^+

(b) Bartlett, Score: 108

	20	50	100	500	1000
20	O	O	A^+	A^+	A^+
50	O	O	O	A^+	A^+
100	O	O	O	A^+	A^+
500	A^+	O	O	A^+	A^+
1000	A^+	A^+	A^+	A^+	A^+

(c) Brown-Forsythe, Score: 110

	20	50	100	500	1000
20	A^+	A^+	A^+	A^+	O
50	O	A^+	A^+	A^+	A^+
100	O	O	O	A^+	A^+
500	A^+	A^+	O	O	A^+
1000	A^+	A^+	O	O	A^+

(d) F-Ratio, Score: 109

	20	50	100	500	1000
20	O	O	O	A^+	A^+
50	O	O	O	A^+	A^+
100	O	O	O	A^+	A^+
500	O	O	A^+	A^+	A^+
1000	A^+	A^+	A^+	A^+	A^+

(e) Fligner-Killeen, Score: 111

	20	50	100	500	1000
20	A^+	A^+	A^+	A^+	A^+
50	A^+	A^+	A^+	A^+	A^+
100	A^+	O	A^+	A^+	A^+
500	A^+	O	A^+	A^+	A^+
1000	A^+	A^+	A^+	A^+	A^+

(f) Levene, Score: 102

Fig. 9: Merged Classification Results for Normal Base (column) vs Normal Mean Shift of 0.1 (row), Average Score: 109.17

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	O	O
500	P	P^{++}	O	O	O
1000	P	P^{++}	O	O	O

(a) ANOVA, Score: 61

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{+}	P^{+}
50	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(b) Bartlett, Score: 64

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(c) Brown-Forsythe, Score: 60

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(d) F-Ratio, Score: 65

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(e) Fligner-Killeen, Score: 59

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(f) Levene, Score: 61

Fig. 10: Merged Classification Results for Normal Base (column) vs Normal Standard Deviation Shift of 0.05 (row), Average Score: 61.67

	20	50	100	500	1000
20	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{+}	O	O	O	O
100	P^{+}	O	O	O	O
500	P^{+}	O	O	O	O
1000	P^{++}	O	O	O	O

(a) ANOVA, Score: 94

	20	50	100	500	1000
20	P^{++}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{++}	O	O	O	O
100	P^{++}	O	O	O	O
500	P^{++}	O	O	O	O
1000	P^{++}	O	O	O	O

(b) Bartlett, Score: 98

	20	50	100	500	1000
20	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{++}	A^{+}	O	O	O
100	P^{++}	O	O	O	O
500	P^{++}	O	O	O	O
1000	P^{++}	O	O	O	O

(c) Brown-Forsythe, Score: 96

	20	50	100	500	1000
20	P^{++}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{++}	O	O	O	O
100	P^{++}	O	O	O	O
500	P^{++}	O	O	O	O
1000	P^{++}	O	O	O	O

(d) F-Ratio, Score: 98

	20	50	100	500	1000
20	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{++}	P^{++}	O	O	O
100	P^{++}	O	O	O	O
500	P^{++}	O	O	O	O
1000	P^{++}	O	O	O	O

(e) Fligner-Killeen, Score: 94

	20	50	100	500	1000
20	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
50	P^{++}	A^{+}	O	O	O
100	P^{++}	O	O	O	O
500	P^{++}	O	O	O	O
1000	P^{++}	O	O	O	O

(f) Levene, Score: 96

Fig. 11: Merged Classification Results for Normal Base (column) vs Normal Standard Deviation Shift of 0.1 (row), Average Score: 96.00

	20	50	100	500	1000
20	P	P^{+}	P^{++}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	O	O
500	P	P^{++}	O	O	O
1000	P	P^{++}	O	O	O

(a) ANOVA, Score: 63

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(b) Bartlett, Score: 66

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(c) Brown-Forsythe, Score: 61

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(d) F-Ratio, Score: 66

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(e) Fligner-Killeen, Score: 61

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(f) Levene, Score: 65

Fig. 12: Merged Classification Results for Normal Base (column) vs Normal Standard Deviation and Mean Shift of 0.05 (row), Average Score: 63.67

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(a) ANOVA, Score: 36

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(b) Bartlett, Score: 34

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(c) Brown-Forsythe, Score: 33

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(d) F-Ratio, Score: 34

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P^{+}	P^{++}	O	O

(e) Fligner-Killeen, Score: 31

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(f) Levene, Score: 33

Fig. 13: Merged Classification Results for Normal Base Case (column) vs Right-skewed Binomial Base Case (row), Average Score: 33.50

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(a) ANOVA, Score: 34

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(b) Bartlett, Score: 33

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{++}	O	O

(c) Brown-Forsythe, Score: 30

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(d) F-Ratio, Score: 33

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(e) Fligner-Killeen, Score: 29

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(f) Levene, Score: 32

Fig. 14: Merged Classification Results for Normal Base Case (column) vs Left-skewed Binomial Base Case (row), Average Score: 31.83

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	O	O
500	P	P^{+}	O	O	O
1000	P	P^{+}	O	O	O

(a) ANOVA, Score: 55

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	A^{+}	O
500	P	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(b) Bartlett, Score: 58

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(c) Brown-Forsythe, Score: 38

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	O	O
500	P	P^{++}	O	O	O
1000	P	P^{++}	O	O	O

(d) F-Ratio, Score: 58

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(e) Fligner-Killeen, Score: 38

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P^{+}	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	A^{+}
500	P	P^{++}	P^{++}	O	O
1000	P	P^{++}	A^{+}	O	O

(f) Levene, Score: 47

Fig. 15: Merged Classification Results for Normal Base (column) vs Binomial Right-skewed Mean Shift of 0.05 (row), Average Score: 49.00

	20	50	100	500	1000
20	P	P^{+}	P^{++}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	A^{+}
100	P	P^{++}	A^{+}	O	O
500	P	P^{++}	O	O	O
1000	P	P^{++}	O	O	O

(a) ANOVA, Score: 67

	20	50	100	500	1000
20	P	P^{+}	P^{++}	P^{++}	P^{++}
50	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	O	O	O
500	P^{+}	A^{+}	O	O	O
1000	P^{+}	O	O	O	O

(b) Bartlett, Score: 75

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	O	O
500	P	P^{++}	O	O	O
1000	P	P^{++}	O	O	O

(c) Brown-Forsythe, Score: 57

	20	50	100	500	1000
20	P	P^{+}	P^{++}	P^{++}	P^{++}
50	P^{+}	P^{++}	P^{++}	P^{++}	A^{+}
100	P^{+}	P^{++}	O	O	O
500	P^{+}	A^{+}	O	O	O
1000	P^{+}	A^{+}	O	O	O

(d) F-Ratio, Score: 76

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P^{+}	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	A^{+}	O
500	P	P^{++}	A^{+}	O	O
1000	P	P^{++}	O	O	O

(e) Fligner-Killeen, Score: 53

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(f) Levene, Score: 65

Fig. 16: Merged Classification Results for Normal Base (column) vs Binomial Right-skewed Mean Shift of 0.1 (row), Average Score: 65.50

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(a) ANOVA, Score: 38

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P^{+}	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{++}	P^{++}	O	O
1000	P	P^{++}	A^{+}	O	O

(b) Bartlett, Score: 43

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(c) Brown-Forsythe, Score: 32

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P^{+}	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{++}	P^{++}	O	O
1000	P	P^{++}	A^{+}	O	O

(d) F-Ratio, Score: 45

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(e) Fligner-Killeen, Score: 29

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(f) Levene, Score: 34

Fig. 17: Merged Classification Results for Normal Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.05 (row), Average Score: 36.83

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{++}
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	A^{+}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(a) ANOVA, Score: 44

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	P^{++}	A^{+}
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(b) Bartlett, Score: 56

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(c) Brown-Forsythe, Score: 34

	20	50	100	500	1000
20	P	P	P^{+}	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{++}	P^{++}	P^{++}	A^{+}
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(d) F-Ratio, Score: 56

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(e) Fligner-Killeen, Score: 29

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(f) Levene, Score: 36

Fig. 18: Merged Classification Results for Normal Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.1 (row), Average Score: 42.50

	20	50	100	500	1000
20	P	P	P^{+}	P^{++}	P^{++}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	O	O
500	P	P^{+}	O	O	O
1000	P	P^{+}	O	O	O

(a) ANOVA, Score: 57

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{+}	P^{+}
50	P^{+}	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(b) Bartlett, Score: 64

	20	50	100	500	1000
20	P	P	P	P	P^{+}
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(c) Brown-Forsythe, Score: 39

	20	50	100	500	1000
20	P	P^{+}	P^{+}	P^{++}	P^{++}
50	P	P^{++}	P^{++}	P^{++}	P^{++}
100	P^{+}	P^{++}	P^{++}	O	O
500	P^{+}	P^{++}	O	O	O
1000	P^{+}	P^{++}	O	O	O

(d) F-Ratio, Score: 65

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P^{+}	P^{++}	P^{++}
100	P	P^{+}	P^{+}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	P^{++}	O	O

(e) Fligner-Killeen, Score: 37

	20	50	100	500	1000
20	P	P	P	P^{+}	P^{+}
50	P	P^{+}	P^{++}	P^{++}	P^{++}
100	P	P^{+}	P^{++}	P^{++}	P^{++}
500	P	P^{+}	P^{++}	O	O
1000	P	P^{+}	A^{+}	O	O

(f) Levene, Score: 44

Fig. 19: Merged Classification Results for Normal Base (column) vs Binomial Right-skewed Standard Deviation and Mean Shift of 0.05 (row), Average Score: 51.00

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(a) ANOVA, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(b) Bartlett, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(c) Brown-Forsythe, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(d) F-Ratio, Score: 125

	20	50	100	500	1000
20	O	O	O	O	O
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	O	O	O	O	O

(e) Fligner-Killeen, Score: 125

	20	50	100	500	1000
20	O	O	O	O	A^+
50	O	O	O	O	O
100	O	O	O	O	O
500	O	O	O	O	O
1000	A^+	O	O	O	O

(f) Levene, Score: 123

Fig. 20: Merged Classification Results for Binomial Right-skewed Base Case vs Binomial Right-skewed Base Case, Average Score: 124.67

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^+	P^{++}
1000	P	P	P	P^+	P^{++}

(a) ANOVA, Score: 6

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^+	P^{++}
1000	P	P	P	P^+	P^{++}

(b) Bartlett, Score: 6

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^+
1000	P	P	P	P^+	P^+

(c) Brown-Forsythe, Score: 3

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^+	P^{++}
1000	P	P	P	P^+	P^{++}

(d) F-Ratio, Score: 6

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^+
1000	P	P	P	P^+	P^+

(e) Fligner-Killeen, Score: 3

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^+	P^+
1000	P	P	P	P^+	P^{++}

(f) Levene, Score: 5

Fig. 21: Merged Classification Results for Binomial Right-skewed Base (column) vs Binomial Right-skewed Mean Shift of 0.05 (row), Average Score: 4.83

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	A^{+}	O
1000	P	P	P^{+}	O	O

(a) ANOVA, Score: 27

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P^{+}
100	P	P	P	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(b) Bartlett, Score: 27

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	O

(c) Brown-Forsythe, Score: 13

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(d) F-Ratio, Score: 28

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	A^{+}

(e) Fligner-Killeen, Score: 11

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P^{+}	P^{++}	A^{+}
1000	P	P	P^{+}	A^{+}	O

(f) Levene, Score: 19

Fig. 22: Merged Classification Results for Binomial Right-skewed Base (column) vs Binomial Right-skewed Mean Shift of 0.1 (row), Average Score: 20.83

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(a) ANOVA, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P^+

(b) Bartlett, Score: 1

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(c) Brown-Forsythe, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P^+

(d) F-Ratio, Score: 1

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(e) Fligner-Killeen, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(f) Levene, Score: 0

Fig. 23: Merged Classification Results for Binomial Right-skewed Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.05 (row), Average Score: 0.33

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{+}	P^{++}

(a) ANOVA, Score: 6

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(b) Bartlett, Score: 8

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(c) Brown-Forsythe, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(d) F-Ratio, Score: 8

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(e) Fligner-Killeen, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(f) Levene, Score: 0

Fig. 24: Merged Classification Results for Binomial Right-skewed Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.1 (row), Average Score: 3.67

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	O

(a) ANOVA, Score: 12

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P^{+}	P^{++}	O

(b) Bartlett, Score: 13

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^{+}
1000	P	P	P	P^{+}	P^{++}

(c) Brown-Forsythe, Score: 4

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	O

(d) F-Ratio, Score: 13

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P^{+}

(e) Fligner-Killeen, Score: 1

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{+}	P^{++}

(f) Levene, Score: 6

Fig. 25: Merged Classification Results for Binomial Right-skewed Base (column) vs Binomial Right-skewed Standard Deviation and Mean Shift of 0.1 (row), Average Score: 8.17

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(a) ANOVA, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(b) Bartlett, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(c) Brown-Forsythe, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(d) F-Ratio, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(e) Fligner-Killeen, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(f) Levene, Score: 0

Fig. 26: Merged Classification Results for Binomial Left-skewed Base Case (row) vs Binomial Right-skewed Base Case (column), Average Score: 0.00

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(a) ANOVA, Score: 7

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(b) Bartlett, Score: 7

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{+}
1000	P	P	P	P^{+}	P^{++}

(c) Brown-Forsythe, Score: 5

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(d) F-Ratio, Score: 7

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^{+}
1000	P	P	P	P^{+}	P^{++}

(e) Fligner-Killeen, Score: 4

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{+}	P^{++}

(f) Levene, Score: 6

Fig. 27: Merged Classification Results for Binomial Left-skewed Base (column) vs Binomial Right-skewed Mean Shift of 0.05 (row), Average Score: 6.00

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{+}	O	O

(a) ANOVA, Score: 29

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{++}	O	O
1000	P	P	P^{++}	O	O

(b) Bartlett, Score: 31

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P	P^{++}	A^{+}
1000	P	P	P	A^{+}	O

(c) Brown-Forsythe, Score: 17

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P^{+}	P^{+}
100	P	P	P^{+}	P^{++}	P^{++}
500	P	P	P^{+}	O	O
1000	P	P	P^{++}	O	O

(d) F-Ratio, Score: 30

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P^{++}	O	O

(e) Fligner-Killeen, Score: 13

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{++}
500	P	P	P^{+}	P^{++}	O
1000	P	P	P^{+}	O	O

(f) Levene, Score: 22

Fig. 28: Merged Classification Results for Binomial Left-skewed Base (column) vs Binomial Right-skewed Mean Shift of 0.1 (row), Average Score: 23.67

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P^+

(a) ANOVA, Score: 1

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^+
1000	P	P	P	P^+	P^+

(b) Bartlett, Score: 3

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(c) Brown-Forsythe, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^+
1000	P	P	P	P^+	P^+

(d) F-Ratio, Score: 3

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(e) Fligner-Killeen, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(f) Levene, Score: 0

Fig. 29: Merged Classification Results for Binomial Left-skewed Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.05 (row), Average Score: 1.17

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(a) ANOVA, Score: 7

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(b) Bartlett, Score: 8

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(c) Brown-Forsythe, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(d) F-Ratio, Score: 8

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P
1000	P	P	P	P	P

(e) Fligner-Killeen, Score: 0

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(f) Levene, Score: 0

Fig. 30: Merged Classification Results for Binomial Left-skewed Base (column) vs Binomial Right-skewed Standard Deviation Shift of 0.1 (row), Average Score: 3.83

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P	P^{++}	P^{++}
1000	P	P	P	P^{++}	O

(a) ANOVA, Score: 13

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P^{+}	P^{++}	P^{++}
1000	P	P	P^{+}	A^{+}	O

(b) Bartlett, Score: 17

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{+}
1000	P	P	P	P^{+}	P^{++}

(c) Brown-Forsythe, Score: 5

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P^{+}	P^{+}
500	P	P	P^{+}	P^{++}	P^{++}
1000	P	P	P^{+}	A^{+}	O

(d) F-Ratio, Score: 17

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P	P^{+}
1000	P	P	P	P^{+}	P^{+}

(e) Fligner-Killeen, Score: 3

	20	50	100	500	1000
20	P	P	P	P	P
50	P	P	P	P	P
100	P	P	P	P	P
500	P	P	P	P^{+}	P^{++}
1000	P	P	P	P^{++}	P^{++}

(f) Levene, Score: 7

Fig. 31: Merged Classification Results for Binomial Left-skewed Base (column) vs Binomial Right-skewed Mean and Standard Deviation Shift of 0.05 (row), Average Score: 10.33