Submitted to Statistical Science

Unifying Boxplots: A Multiple Testing Perspective

Bowen Gang, Hongmei Lin and Tiejun Tong

Abstract. Tukey's boxplot is a foundational tool for exploratory data analysis, but its classic outlier-flagging rule does not account for the sample size, and subsequent modifications have often been presented as separate, heuristic adjustments. In this paper, we propose a unifying framework that recasts the boxplot and its variants as graphical implementations of multiple testing procedures. We demonstrate that Tukey's original method is equivalent to an unadjusted procedure, while existing sample-size-aware modifications correspond to controlling the Family-Wise Error Rate (FWER) or the Per-Family Error Rate (PFER). This perspective not only systematizes existing methods but also naturally leads to new, more adaptive constructions. We introduce a boxplot motivated by the False Discovery Rate (FDR), and show how our framework provides a flexible pipeline for integrating state-of-the-art robust estimation techniques directly into the boxplot's graphical format. By connecting a classic graphical tool to the principles of multiple testing, our work provides a principled language for comparing, critiquing, and extending outlier detection rules for modern exploratory analysis.

Key words and phrases: Boxplot, Multiple Testing, Outlier detection, Exploratory Data Analysis, False Discovery Rate.

1. INTRODUCTION

Tukey's box-and-whisker plot remains a cornerstone of exploratory data analysis celebrated for its elegant distillation of a sample's key features [31]. Its outlier flagging rule, however, based on a fixed multiple of the interquartile range, does not account for the sample size.[1] This property, while simple, leads to a procedural artifact where the number of "outliers" detected in samples from a pure normal distribution grows with the number of observations, a behavior inconsistent with modern inferential standards.

Over the decades, numerous authors have proposed modifications to make the boxplots outlier-detection mechanism adaptive to the sample size [2–4, 14, 26]. While these proposals vary in their approach, with some replacing the fence rule entirely, many are introduced as heuristic adjustments. For this latter group, a closer examination reveals a notable, though typically implicit, relationship to the core principles of multiple hypothesis testing. For example, procedures that control the "some-outside rate per sample" the probability of incorrectly flagging at least one non-outlying observation as an outlier are conceptually equivalent to controlling the Family-Wise Error Rate (FWER) [13, 27, 28]. More recently, the "Chauvenet-type boxplot" introduced by [20] implicitly aligns with the control of the Per-Family Error Rate (PFER), as it aims to limit the expected number of false positives to a small constant.

In this paper, we argue that these are not isolated parallels but rather different manifestations of a single, unifying idea: the boxplot can be viewed as a graphical implementation of a multiple testing procedure. This perspective provides a powerful and coherent framework for understanding, comparing, and extending boxplot methodologies. The classic Tukey boxplot represents an unad-

Bowen Gang is Assistant Professor, Department of Statistics and Data Science, Fudan University, Shanghai, China (e-mail: bgang@fudan.edu.cn). Hongmei Lin is Associate Professor, School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai, China (e-mail: hmlin@suibe.edu.cn). Tiejun Tong is Professor, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China (e-mail: tongt@hkbu.edu.hk).

[1] It is worth noting that the interpretation of the boxplot's fences as a definitive rule for identifying "outliers" is a later development. Tukey's original proposal was more nuanced, distinguishing between "outside" observations (those between 1.5 and 3 interquartile ranges (IQRs) from the quartiles) and "far out" observations (more than 3 IQRs away), without explicitly labeling either as outliers. In its modern application, both "outside" and "far out" points are typically flagged simply as potential outliers.

justed procedure, while its successors correspond to different, more stringent modes of error control. The primary contribution of this work is the establishment of this unifying framework. Once the boxplot is seen through this lens, its potential is dramatically expanded. As a first natural consequence, we demonstrate how to construct a boxplot motivated by the False Discovery Rate (FDR), the canonical error metric for modern large-scale inference [5]. This yields fences that adapt not only to the sample size but also to the data themselves. Perhaps more importantly, this conceptual bridge connects a classic graphical tool to the rich and evolving literature on multiple testing. It transforms the boxplot from a rigid set of outlier-detection rules into a flexible and adaptive inferential framework. By incorporating recent advances in robust estimation and applying p-value-based multiple testing procedures, researchers can readily tailor the boxplot to their specific analytical needs, enhancing both its relevance and effectiveness in modern data analysis.

2. A MULTIPLE TESTING FRAMEWORK FOR BOXPLOTS: A CONCEPTUAL REVIEW

The core observation of this short paper is that the evolution of the boxplot's outlier-detection rules can be systematically understood as the application of different multiple testing error-control standards. To formalize this, let us consider a sample of observations $\{X_1, \ldots, X_n\}$. To motivate our framework, we assume these are independent and identically distributed (iid) draws from a continuous distribution. This allows us to conceptually model the sample as a potential mixture: the bulk of the data drawn from one distribution, with a small fraction of outliers arising from one or more contaminating distributions.

For each observation, we can define a null hypothesis, H_{0i}, that states "X_i is not an outlier", meaning it is drawn from the same underlying distribution as the bulk of the data. The act of flagging X_i as a potential outlier is then equivalent to rejecting H_{0i}. With n such tests being performed simultaneously, the central question becomes how to manage the rate of Type I errors. Let V be the number of non-outlying observations that are incorrectly flagged as outliers, i.e., the number of true null hypotheses that are rejected. Let also Q_1 and Q_3 be the 1st and 3rd quartiles of the samples,[2] respectively, and $\text{IQR} = Q_3 - Q_1$ be the interquartile range. We can now map existing boxplot variations to specific strategies for controlling V.

[2]There are multiple ways to define sample quartiles. Throughout this paper, we use the sample quartile definition corresponding to Definition 7 in [17], as it is the default in the R function `quantile` and widely adopted in practice.

2.1 The Unadjusted Procedure: Tukey's Classic Boxplot

Tukey's original boxplot defines its lower and upper inner fences at $\text{LF} = Q_1 - 1.5 \times \text{IQR}$ and $\text{UF} = Q_3 + 1.5 \times \text{IQR}$, respectively. For data from a normal distribution, the probability of an observation falling outside these two fences is approximately 0.7%. In the language of hypothesis testing, this corresponds to setting a Per-Comparison Error Rate (PCER)—the probability of a Type I error for a single test—at a fixed level of 0.007.

Critically, this procedure makes no adjustment for multiplicity. The threshold is constant regardless of the number of tests, n. Consequently, the expected number of falsely flagged outliers, $E(V)$, is approximately $0.007n$, which grows linearly with the sample size. This lack of adjustment is the defining feature of an unadjusted multiple testing procedure and explains the well-known tendency of the classic boxplot to flag an excessive number of points in large datasets.

2.2 Controlling the FWER

The most stringent form of multiple testing control is the management of the FWER, defined as the probability of making at least one Type I error:

$$\text{FWER} = P(V \geq 1).$$

This is precisely the conceptual goal of boxplot modifications that seek to control the "some-outside rate per sample" [13, 28]. These methods derive a sample-size-dependent fence coefficient, k_n, to define the outlier region $\mathcal{O}_n = (-\infty, \text{LF}_n) \cup (\text{UF}_n, \infty)$, where

$$\text{LF}_n = Q_1 - k_n \times \text{IQR} \quad \text{and} \quad \text{UF}_n = Q_3 + k_n \times \text{IQR}.$$

Considerable effort has been dedicated to numerically solving for the value of k_n that ensures

$$P(\text{at least one of } X_1, \ldots, X_n \in \mathcal{O}_n) \leq \alpha,$$

for some given α, assuming the bulk of the data are drawn from a normal distribution [12, 28].

2.3 Controlling the PFER

A less conservative Type I error notion is the PFER, which controls the expected number of Type I errors:

$$\text{PFER} = E(V).$$

The Chauvenet-type boxplot recently proposed by [20] is a direct implementation of this principle. The method's fence coefficient, k_n, is derived from Chauvenet's criterion, which historically identifies an observation as an outlier if its p-value is less than $0.5/n$. For n tests, this corresponds to controlling the PFER at a level of $n \times (0.5/n) = 0.5$. The fence coefficient is set to

$$k_n = \frac{\Phi^{-1}(1 - 0.25/n)}{1.35} - 0.5,$$

where Φ^{-1} is the quantile function of the standard normal distribution, and the constants 0.5 and 1.35 arise from using quartiles to estimate the mean and standard deviation. By construction, assuming the non-outlying data are normally distributed, this method aims to produce, on average, half a false positive per dataset, regardless of the sample size.

3. A UNIFIED p-VALUE PIPELINE FOR BOXPLOT CONSTRUCTION

The conceptual framework above, which connects boxplots to multiple testing, naturally leads to a unified and flexible methodology for the fence construction. Instead of deriving fence coefficients k_n for each error metric, we propose a general pipeline that transforms the outlier detection problem into a standard p-value-based multiple testing problem. This approach is not only conceptually simpler but also vastly more extensible. The pipeline consists of four steps as follows.

1. **Parameter Estimation.** First, we obtain robust estimates of the parameters governing the distribution of the non-outlying data. Assuming the bulk of the data follows a normal distribution, we can robustly estimate its location μ and scale σ using the sample quartiles as recommended by [11, 31][3]:
$$\hat{\mu} = \frac{Q_1 + Q_3}{2} \quad \text{and} \quad \hat{\sigma} = \frac{Q_3 - Q_1}{1.35}. \tag{1}$$
2. **p-value Calculation.** Next, we convert each observation X_i into a two-sided p-value based on the estimated null distribution:
$$p_i = 2\left(1 - \Phi\left(\left|\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right|\right)\right).$$
This transforms the raw data into a set of p-values $\{p_1, \ldots, p_n\}$, which serve as standardized evidence against the null hypothesis of not being an outlier.
3. **Multiple Testing Adjustment.** The set of n p-values is fed into any standard p-value-based multiple testing procedure to control a desired error rate (e.g., FWER, PFER, FDR) at a level α. The procedure returns a significance threshold, t_{adj}, which is potentially data-dependent. Any hypothesis H_{0i} with $p_i \leq t_{\text{adj}}$ is rejected by the chosen testing procedure.
4. **Fence Construction.** Finally, to translate this decision rule back into the graphical language of a boxplot, we determine the z-score corresponding to the p-value threshold, $z_{\text{adj}} = \Phi^{-1}(1 - t_{\text{adj}}/2)$. The fences are then defined directly on the original data scale:
$$\text{LF}_n = \hat{\mu} - z_{\text{adj}} \cdot \hat{\sigma} = Q_1 - (z_{\text{adj}}/1.35 - 0.5) \times \text{IQR},$$
$$\text{UF}_n = \hat{\mu} + z_{\text{adj}} \cdot \hat{\sigma} = Q_3 + (z_{\text{adj}}/1.35 - 0.5) \times \text{IQR}.$$

If prior knowledge suggests that outliers are likely to appear only on one side of the distribution, the procedure can be adapted by using one-sided p-values. For example, to detect large outliers, the p-value calculation in Step 2 would be modified to $p_i = 1 - \Phi((X_i - \hat{\mu})/\hat{\sigma})$. Consequently, the fence construction in Step 4 becomes asymmetric. The z-score threshold is now $z_{\text{adj}} = \Phi^{-1}(1 - t_{\text{adj}})$, which defines an upper fence at $\text{UF}_n = \hat{\mu} + z_{\text{adj}} \cdot \hat{\sigma}$. No lower fence is statistically defined by this test, so the lower whisker simply extends to the sample minimum.

This pipeline provides a principled way to construct boxplots. The choice of error metric and control procedure in Step 3 directly determines the behavior of the resulting outlier-flagging rule, allowing for transparent design and easy extension.

3.1 The FWER Boxplot

To create a boxplot that controls the FWER, we simply apply a standard FWER-controlling procedure in Step 3 of our pipeline. For instance, using the Bonferroni or the more powerful Holm method [15] on the calculated p-values at level α will yield a threshold t_{FWER}. The resulting fences will graphically represent the rejection region of the chosen FWER procedure. This approach elegantly circumvents the need for complex numerical simulations to find k_n and connects the boxplot directly to established inferential machinery.

3.2 The PFER Boxplot

The principle behind the Chauvenet-type boxplot can also be implemented through our pipeline. To control the PFER at a level γ (e.g., $\gamma = 0.5$), one would reject any p-value smaller than $t_{\text{PFER}} = \gamma/n$. This is equivalent to taking the z-value threshold z_{adj} in Step 4 of the general pipeline to be $\Phi^{-1}(1 - \gamma/(2n))$. If μ and σ are estimated using Eq. (1), we recover the Chauvenet-type boxplot as in [20].

3.3 A Natural Extension: The FDR Boxplot

Having established this flexible pipeline, a natural and powerful extension is to construct a boxplot that controls the False Discovery Rate (FDR), the canonical error metric for modern large-scale inference [5]. The FDR is the

[3] In the case where $Q_1 = Q_3$, which implies IQR $= 0$, our estimator $\hat{\sigma}$ becomes zero. This typically occurs with discrete or heavily rounded data. When this happens, we recommend using a scale estimator based on the Median Absolute Deviation (MAD) to ensure a non-zero denominator whenever the IQR is zero but the data are not all identical [23]. That is, $\hat{\sigma} = \text{MAD}/0.675$, where MAD $=$ median$(|X_i - \text{median}(X_1, \ldots, X_n)|)$.

expected proportion of false discoveries among all rejected hypotheses. Let R be the total number of observations flagged as outliers and V be the number of non-outlying observations that are incorrectly flagged as such. The FDR is defined as

$$\text{FDR} = E\left[\frac{V}{R}\right], \quad \text{with } \frac{V}{R} \equiv 0 \text{ if } R = 0.$$

To construct an FDR-controlling boxplot, we apply an FDR procedure, such as the Benjamini-Hochberg (BH) method, in Step 3 of our pipeline. This yields a data-dependent threshold t_{FDR}. The fences are then set based on this threshold:

$$\text{LF}_n = \hat{\mu} - z_{\text{FDR}} \cdot \hat{\sigma} = Q_1 - (z_{\text{FDR}}/1.35 - 0.5) \times \text{IQR}$$

$$\text{UF}_n = \hat{\mu} + z_{\text{FDR}} \cdot \hat{\sigma} = Q_3 + (z_{\text{FDR}}/1.35 - 0.5) \times \text{IQR},$$

where $z_{\text{FDR}} = \Phi^{-1}(1 - t_{\text{FDR}}/2)$.

The key advantage of the FDR boxplot is its **adaptivity**. If the data contain many true outliers with very small p-values, the BH procedure will yield a larger threshold t_{FDR}. This results in narrower fences and thus greater power to detect additional outliers. Conversely, if the data contain few or no outliers, the threshold will be small, resulting in wider, more conservative fences. This dynamic behavior makes the FDR boxplot a uniquely powerful tool for modern exploratory analysis, and its straightforward derivation showcases the practical utility of our unifying framework.

3.4 A Toy Example: From Data to Fences

To make this pipeline concrete, consider a small dataset of $n = 11$ observations as

$$X = \{9, 16, 18, 20, 20, 22, 22, 24, 26, 36, 50\}.$$

In this dataset, the observation 50 is well-separated from the bulk of the data, while 9 and 36 can be regarded as borderline cases. We will walk through the pipeline to see how the FWER, PFER, and FDR controls yield different flagged outliers and significance threshold t_{adj} with the control levels being $\alpha = 0.01$ for FWER/FDR and $\gamma = 0.5$ for PFER.

Steps 1-2: Parameter Estimation and p-value Calculation.

- The sample quartiles are $Q_1 = 19$ and $Q_3 = 25$, giving IQR $= 6$.
- From Eq. (1), we have $\hat{\mu} = (19 + 25)/2 = 22$ and $\hat{\sigma} = 6/1.35 \approx 4.44$.
- The eleven p-values, sorted in ascending order, are:

$$p_{(1)} \approx 2.98 \times 10^{-10} \text{ (for 50)},$$

$$p_{(2)} \approx 1.63 \times 10^{-3} \text{ (for 36)},$$

$$p_{(3)} \approx 3.44 \times 10^{-3} \text{ (for 9)},$$

$$p_{(4)} \approx 0.177 \text{ (for 16)}, \ldots$$

The rest of the p-values are all greater than 0.36.

Steps 3-4: Multiple Testing and Fence Construction. Now we apply different adjustment procedures to find the significance threshold t_{adj}.

- **PFER (Chauvenet):** The threshold is fixed: $t_{\text{PFER}} = \gamma/n = 0.5/11 = 0.045$. The first three p-values $(p_{(1)}, p_{(2)}, p_{(3)})$ are all well below this threshold. This flags $\{50, 36, 9\}$ as outliers.
- **FDR (BH) at $\alpha = 0.01$:** The BH procedure finds the largest i such that $p_{(i)} \le i\alpha/n = i(0.01)/11$. We have $p_{(1)} \le 1(0.01)/11$, $p_{(2)} \le 2(0.01)/11$, and $p_{(3)} \approx 0.00344 \not\le 3(0.01)/11$. That is, the largest i is 2. The procedure rejects the first two hypotheses, flagging $\{50, 36\}$, and the significance threshold is 1.63×10^{-3}.
- **FWER (Holm) at $\alpha = 0.01$:** The Holm procedure compares $p_{(i)}$ to $\alpha/(n-i+1) = 0.01/(12-i)$. We have $p_{(1)} < 0.01/11$ and $p_{(2)} \not\le 0.01/10$. It thus rejects only the first hypothesis, flagging just $\{50\}$ as outlier. The significance threshold is 2.98×10^{-10}.

Summary of Results. This toy example clearly illustrates the distinct behaviors of boxplots under different error-control philosophies. In this example with small n, the PFER approach is the most liberal. The FDR method, even at a strict $\alpha = 0.01$, is adaptive enough to detect two outliers due to the strong evidence from the first. The FWER method, being the most conservative, stops short and only flags 50 as an outlier at this strict α level. The resulting fences, shown in Table 1, are starkly different. For reference, Tukey's classic method, which relies on a fixed $1.5\times$IQR rule (corresponding to $t_{\text{adj}} = 0.007$) rather than a statistical adjustment, would have flagged $\{50, 36, 9\}$ in this instance.

4. NUMERICAL EXPERIMENTS

To visually demonstrate the distinct behaviors of different outlier detection rules, we conduct simulation studies that compare five types of boxplots. We compare the unadjusted Tukey boxplot, the three boxplots derived from our framework (controlling FWER, PFER, and FDR), and the method proposed by [4] (hereafter, BGL).

4.1 Implementation Details

The five methods are implemented as follows. The FWER, PFER, and FDR methods all rely on the same p-value generation pipeline, differing only in the final multiple testing adjustment.

- **PCER-type (Tukey):** The classic boxplot with fences defined at $Q_1 - 1.5 \times$ IQR and $Q_3 + 1.5 \times$ IQR.
- **FWER-type (Holm):** The Holm procedure [15] is used for testing at a target FWER level of $\alpha = 0.01$.

TABLE 1
Comparison of significance threshold and outliers flagged for the toy example.

Method	t_{adj}	Outliers Flagged	Lower/Upper Fences
PCER (Tukey)	0.007	$\{50, 36, 9\}$	$[10, 34]$
PFER (Chauvenet) at 0.5	0.05	$\{50, 36, 9\}$	$[13.3, 30.7]$
FDR (BH) at 0.01	$p_{(2)} \approx 0.00163$	$\{50, 36\}$	$[8, 36]$
FWER (Holm) at 0.01	$p_{(1)} \approx 2.98 \times 10^{-10}$	$\{50\}$	$[-6, 50]$

- **PFER-type (Chauvenet):** We use the Chauvenet-type boxplot from [20], which is constructed to control the PFER at a target level of 0.5.
- **FDR-type (BH):** The BH procedure [5] is used for testing at a target FDR level of $\alpha = 0.01$.
- **BGL-type:** The method from [4], which makes the fences sample-size-dependent. The two fences are defined at $Q_1 - 1.5 \times \text{IQR} \times [1 + 0.1 \log(n/10)]$ and $Q_3 + 1.5 \times \text{IQR} \times [1 + 0.1 \log(n/10)]$.

For the three p-value-based methods (Holm, Chauvenet, and BH), the p-value for each observation is calculated relative to an assumed normal distribution. The parameters of this distribution are defined by the robust quartile-based estimators from Eq. (1). The implementations of the FWER-type (Holm) and FDR-type (BH) boxplots are available in our newly developed R package `AdaptiveBoxplot`, which can be found on GitHub at https://github.com/bgang92/AdaptiveBoxplot.

4.2 Normally Distributed Majority

We generate data from a normal mixture model to simulate a common scenario: the bulk of the observations are drawn from a standard normal $N(0, 1)$ distribution, but 1% are contaminating "true outliers" from a $N(5, 1)$ distribution. The formal model is as follows:

$$\theta_i \overset{iid}{\sim} \text{Bernoulli}(0.01),$$

$$X_i | \theta_i \overset{ind}{\sim} (1 - \theta_i) N(0, 1) + \theta_i N(5, 1),$$

where $i = 1, \ldots, n$. We simulate datasets with sample sizes $n = 5 \times 10^k$ for $k = 1, 2, 3$.

Figure 1 provides a visual confirmation of our framework. The failure of the Tukey-type boxplot at large sample sizes is evident. As n increases from 50 to 5000, the number of observations flagged from the $N(0, 1)$ distribution explodes, flooding the plot with false positives and obscuring the true outliers. In contrast, all four adjusted methods adapt their fences to account for the sample size, preventing the flood of false discoveries seen with the fixed coefficient approach.

While the figure gives a qualitative overview, a deeper analysis of the fence coefficients in Table 2 reveals the core mechanistic differences between the methods. This table highlights two distinct levels of adaptivity. The

TABLE 2
Comparison of fence coefficients for data from a normal mixture model. The coefficients are averaged over 5000 simulation replicates.

Method	$k = 1$	$k = 2$	$k = 3$
PFER (Chauvenet) at 0.5	1.41	1.93	2.38
FDR (BH) at 0.01	2.08	2.82	2.46
FWER (Holm) at 0.01	2.11	3.02	3.06
BGL	1.60	1.75	1.90

Chauvenet (PFER) and BGL boxplots illustrate the first level: their fence coefficients are deterministic functions of the sample size n. Both become more conservative as n grows, a clear improvement over the fixed Tukey fences, but they remain blind to the actual content of the data.

In contrast, the FWER and FDR boxplots exhibit a more profound, data-driven adaptivity. Their fence coefficients are not fixed by n alone but are determined by the empirical distribution of the p-values from the sample. This is powerfully illustrated by the fence coefficient for FDR (BH) in Table 2. As n increases from 50 ($k = 1$) to 500 ($k = 2$), the coefficient increases to guard against false positives. However, as n grows to 5000 ($k = 3$), the coefficient surprisingly decreases. This is the signature of FDR's adaptivity: at $n = 5000$, there are approximately 50 strong outlier signals. The BH procedure detects this abundance of outliers, becomes more powerful, and sets a more liberal threshold (a smaller z-score), resulting in tighter fences to capture additional, less extreme outliers. The Holm-FWER method, bound by its goal of preventing even a single error, cannot leverage this information as efficiently as BH-FDR.

This analysis shows that the choice is not merely between fixed and adaptive fences, but between different philosophies of adaptation: one that is pre-determined by the sample size, and another that dynamically responds to the evidence within the data itself.

4.3 Performance on Skewed Data

We now consider a scenario where the data are drawn from a single, skewed distribution. We generate n observations directly from a chi-square distribution with 10 degrees of freedom, χ^2_{10}. This distribution is right-skewed,

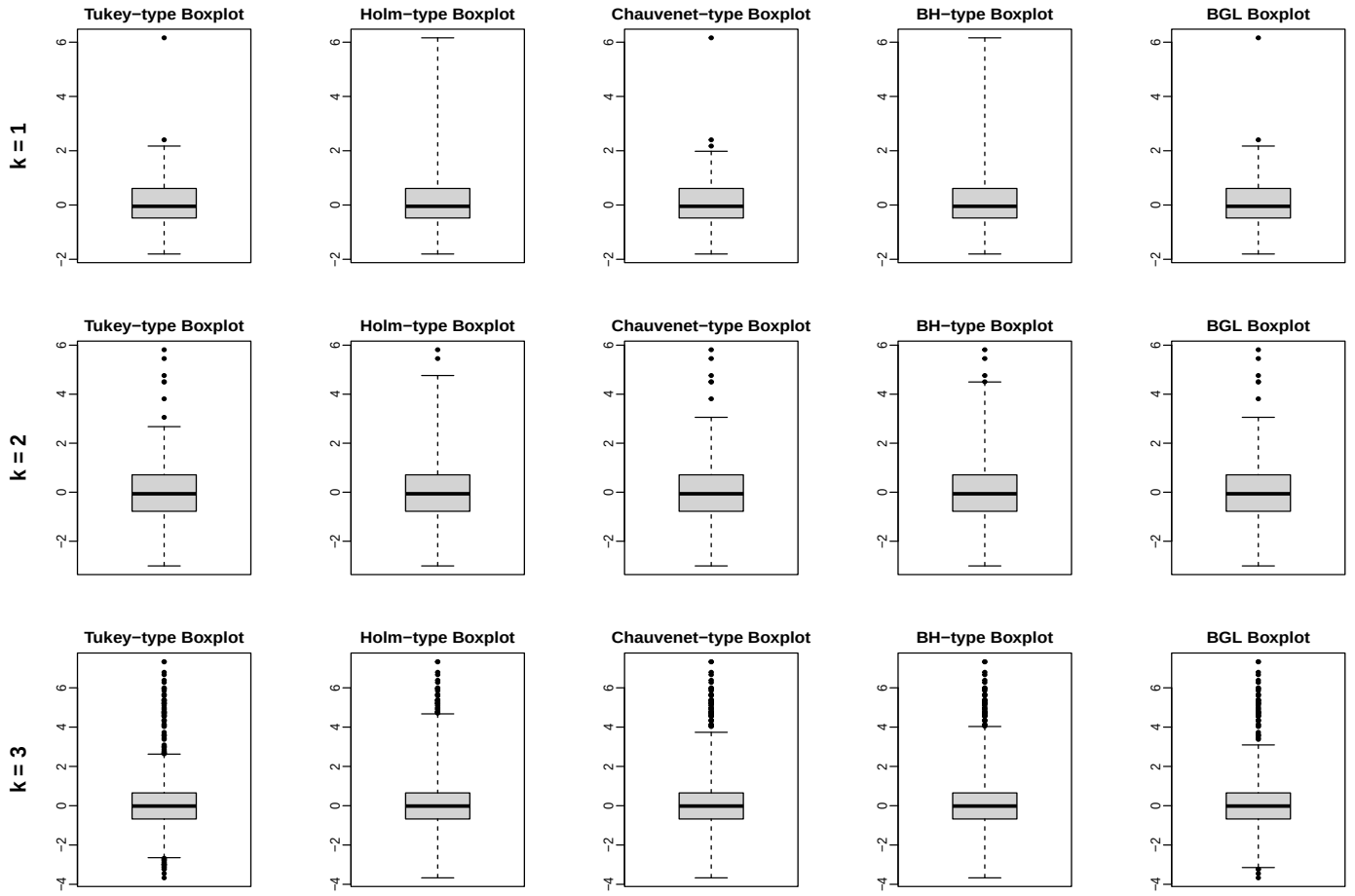

FIG 1. *A visual comparison of the five boxplot methodologies across different sample sizes. The data are from a normal mixture model with 1% true outliers. The behavior of each boxplot aligns perfectly with its underlying error control principle.*

meaning that even without true mixture-model outliers, there is a heavy upper tail that can be taken for outliers.

Crucially, for the FWER, PFER, and FDR-type boxplots, we deliberately maintain the simple, normal-based procedure from the previous section. That is, we continue to estimate the location and scale parameters using Eq. (1) and proceed as if the majority of the data were sampled from a normal distribution. This allows us to investigate the practical performance of the methods under model misspecification. We again simulate datasets with sample sizes $n = 5 \times 10^k$ for $k = 1, 2, 3$. The results are summarized in Figure 2 and Table 3.

As the sample size n increases, all five methods begin to flag a growing number of points in the upper tail as outliers. This is the expected outcome when a symmetric model is imposed on an asymmetric distribution. The right skewness ensures that many genuine, albeit extreme, observations from the χ^2_{10} distribution will fall far above the symmetrically-placed upper fence. We observe a similar hierarchy of conservatism as in the normal case. The Tukey-type boxplot, being unadjusted, flags the most points. The Holm-type boxplot is the most conservative, flagging the fewest points due to its stringent FWER control, which results in the widest fences. The Chauvenet-type, BH-type, and BGL boxplots lie in between.

Table 3 quantifies this and reveals deeper insights when compared to the results from the normal case (Table 2). The PFER and BGL fence coefficients are identical in both tables, this is not surprising given the two methods are blind to the data's distribution and depend only on the sample size. In contrast, the FDR and FWER methods are clearly responsive. The FDR boxplot, in particular, is more adaptive than the FWER boxplot. It interprets the flood of small p-values from the heavy tail as evidence of abundant "outliers", causing it to become more liberal. Consequently, its fence coefficient only increases slightly when the sample size increases from 500 to 5000. This demonstrates that the FDR procedure is working as designed.

This simulation reveals the diagnostic power of our framework. When a normal-based procedure is applied to skewed data, all methods are naturally affected. However, the problem is not a flaw in the multiple testing logic itself, but rather in the initial model of the data used to gen-

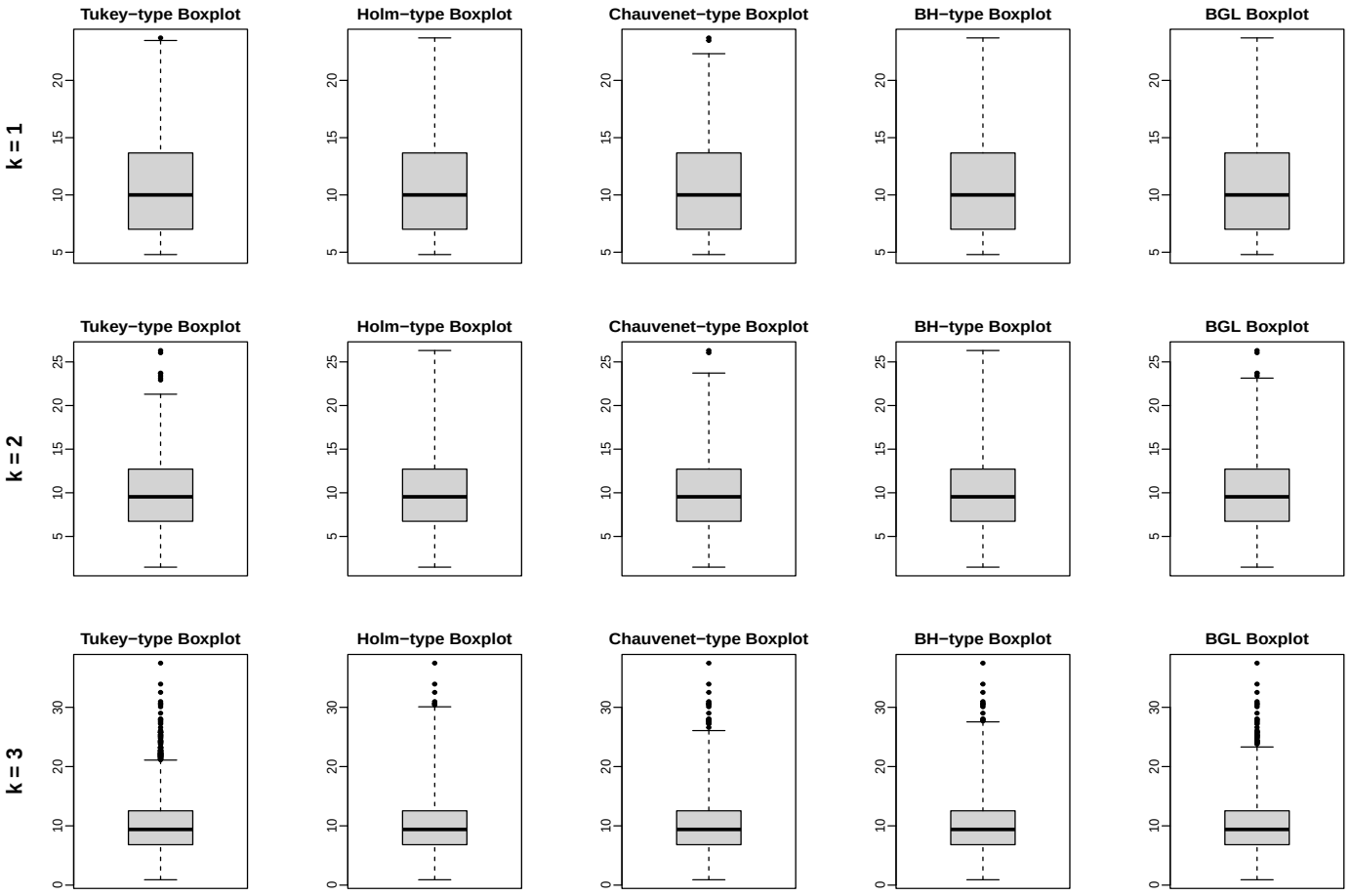

FIG 2. *Performance of the five boxplot methodologies on right-skewed data generated from a* χ^2_{10} *distribution.*

TABLE 3
Comparison of fence coefficients for χ^2_{10} *data, fence coefficients computed by averaging results over 5000 simulation replicates.*

Method	$k=1$	$k=2$	$k=3$
PFER (Chauvenet) at 0.5	1.41	1.93	2.38
FDR (BH) at 0.01	1.82	2.65	2.78
FWER (Holm) at 0.01	1.84	2.73	3.17
BGL	1.60	1.75	1.90

erate p-values. The challenge of improving the boxplot for non-normal data is thus cleanly reframed as a more familiar statistical task: selecting a more appropriate probability distribution for the main body of the data. The adjustment machinery is sound; its inputs must simply be more accurate.

5. GENERALIZING THE FRAMEWORK: BEYOND NORMALITY

So far our analysis has proceeded under the working assumption that the bulk of the data is drawn from a normal distribution. This is not an arbitrary choice, but rather a robust and pragmatic starting point for exploratory analysis. Its justification is rooted in Winsor's principle, which famously states that "*All distributions are normal in the middle*" [30]. However, the power of the p-value pipeline lies in its flexibility, not its adherence to a single distribution. If prior knowledge suggests that the bulk of the data is better described by a different parametric family, which we denote generally as $F(\cdot;\theta)$, the framework can and should be adapted. Here, F is the cumulative distribution function and θ is the vector of its parameters (e.g., location, scale, shape). The procedure is generalized as follows:

1. **Parameter Estimation:** We obtain robust estimates of the parameters of the chosen parametric family, denoted as $\hat{\theta}$.
2. **p-value Calculation:** We compute p-value for each observation using the estimated distribution function, $F(X_i;\hat{\theta})$.
3. **Multiple Testing Adjustment:** With the new set of p-values, we apply the chosen multiple testing procedure (e.g., FWER, PFER, FDR) to obtain an ad-

justed significance threshold, t_{adj}. Any hypothesis for which $p_i \leq t_{\text{adj}}$ is rejected.

4. **Fence Construction:** Finally, the decision rule is translated back into the graphical language of the boxplot. The fences are constructed directly from the quantiles of the fitted distribution $F(\cdot, \hat{\theta})$. When two-sided p-value is used, the rejection region is split between both tails:

$$\text{LF}_n = F^{-1}(t_{\text{adj}}/2; \hat{\theta}),$$
$$\text{UF}_n = F^{-1}(1 - t_{\text{adj}}/2; \hat{\theta}).$$

As noted in Section 3, this pipeline can be made more appropriate for skewed data by using one-sided p-values. In the general case, the p-value calculation in Step 2 is modified to test for outliers in a specific tail. For a right-sided test (to detect large outliers), the p-value is $p_i = 1 - F(X_i; \hat{\theta})$. Consequently, the fence construction in Step 4 is asymmetric, defining only an upper fence at $\text{UF}_n = F^{-1}(1 - t_{\text{adj}}; \hat{\theta})$. For a left-sided test (to detect small outliers), the p-value is $p_i = F(X_i; \hat{\theta})$, which defines only a lower fence at $\text{LF}_n = F^{-1}(t_{\text{adj}}; \hat{\theta})$. In either one-sided case, the whisker on the non-tested side simply extends to the sample minimum or maximum.

To demonstrate the practical utility of this generalized framework, we now conduct a simulation study. We use the same data generating process from Section 4.3, where datasets are drawn from a χ^2_{10} distribution and contain no contaminating outliers. An ideal outlier detection procedure should therefore flag very few, if any, observations. We compare the performance of the five methods mentioned in Section 4.1.

To estimate the degrees of freedom parameter, we use a robust estimator based on the highly accurate Wilson Hilferty approximation for the median of a χ^2 variable [32], which states that $\text{median}(\chi^2_k) \approx k(1 - 2/(9k))^3$. Accordingly, k is estimated by numerically solving the equation

$$\text{median}(X_1, \ldots, X_n) = \hat{k}\left(1 - \frac{2}{9\hat{k}}\right)^3.$$

To compute p-values, since the χ^2_k distribution is skewed to the right, we use the one-sided p-value

$$p_i = 1 - F(X_i; \hat{k}),$$

where $F(\cdot, \hat{k})$ is the distribution function of the $\chi^2_{\hat{k}}$ random variable. These p-values are then used to construct two boxplots: one using the Holm procedure to control the FWER and another using the BH procedure to control the FDR, both at a target level of 0.01. For the Chauvenet-type boxplot, we define the upper fence at $F^{-1}(1 - 0.5/n, \hat{k})$ and the lower fence at $\min\{X_1, \ldots, X_n\}$. The results are summarized in Figure 3.

The figure provides a striking and unambiguous confirmation of our generalized framework's utility. The Tukey and BGL boxplots, which are both based on an assumption of symmetry, consequently flag a substantial number of observations in the upper tail. This effect becomes more pronounced with increasing sample size, leading to a visualization containing numerous false positives that can obscure the true structure of the data. In stark contrast, the methods based on the correctly specified χ^2 model perform exceptionally well. The Holm-type and BH-type boxplots achieve near-perfect performance, flagging no observations as outliers. The Chauvenet-type boxplot is also highly effective, flagging only a small number of the most extreme observations. All three methods demonstrate a vast improvement over the normal-based approaches because they use an appropriate reference distribution, correctly recognizing that the heavy upper tail is an intrinsic feature of the data.

6. CONCLUDING REMARKS

This paper reframes the boxplot, transforming it from a collection of disparate, heuristic rules into a coherent and extensible statistical methodology. For decades, the evolution of the boxplot has been characterized by ad-hoc, sample-size-specific adjustments. We have shown that many seemingly isolated modifications can be understood and improved through the unified lens of multiple hypothesis testing. Our primary contribution is the development of a general p-value pipeline that operationalizes this insight. This pipeline reveals a powerful underlying unity: it shows that all major boxplot variations can be generated by a single fence formula $(z_{\text{adj}}/1.35 - 0.5) \times \text{IQR}$, whose sensitivity is governed solely by an effective z-score, z_{adj}. Remarkably, Tukey's classic $1.5 \times \text{IQR}$ rule can be regarded as a special case, corresponding to a fixed $z_{\text{adj}} = 2.7$. Our pipeline, therefore, not only systematizes existing methods by mapping them to explicit error control principles (PCER, FWER, PFER) but, more importantly, provides an engine for creating new, more powerful diagnostic tools. The power of this unified methodology is demonstrated through its immediate practical payoffs. We introduced the FDR boxplot, a novel construction that brings the canonical error metric of modern large-scale science to a classic exploratory data analysis tool.

The practical utility of this framework lies in its modularity and flexibility. As we demonstrated in Section 4.3, the pipeline is not limited to the assumption of normality but provides a flexible scaffold for building principled, context-aware outlier detection tools for any parametric family. This transforms the boxplot from a static summary graphic into a dynamic diagnostic tool, where its performance offers direct insight into the appropriateness of the underlying distributional assumptions.

Further refinements to our p-value pipeline are readily available by incorporating more advanced estimation

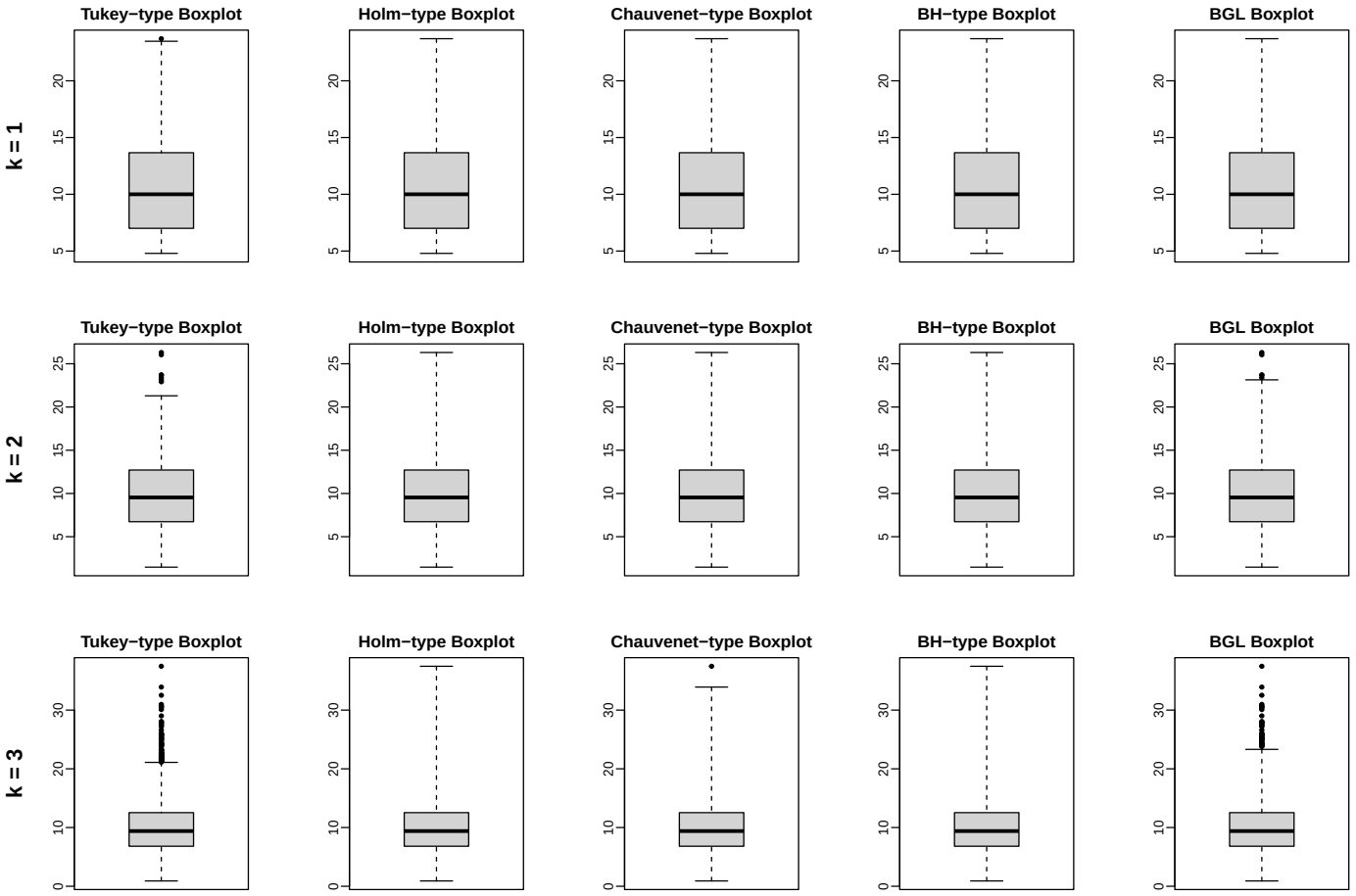

FIG 3. *Performance of the five boxplot methodologies on right-skewed data generated from a* χ^2_{10} *distribution, assuming the model is correctly specified.*

techniques in Step 1. While our analysis relied on simple, robust quartile-based estimators, the framework can be significantly enhanced by employing methods from the rich literature on robust statistics. This field provides a wealth of tools designed to yield reliable parameter estimates in the presence of data contamination. For comprehensive treatments of robust methods, from foundational theory to modern applications, see, for example, [10, 16, 21]. In the specific but common case where the bulk of the data is assumed to follow a normal distribution, a powerful alternative is the empirical null methodology. This data-adaptive approach, developed for large-scale inference, provides a suite of tools for accurately estimating the parameters of the majority normal distribution directly from the data [8, 9, 19].

Despite the power of these refinements, it is important to contextualize the role of the boxplots discussed here. They remain, at their heart, tools for exploratory data analysis. While their designs are motivated by and aligned with formal notions of Type I error, we do not claim that they rigorously control these error rates in a formal inferential sense. The procedures we have outlined rely on robust but simple estimates of an assumed distribution for the bulk of the data, and true statistical control would require a more formal handling of parameter estimation uncertainty and potential dependencies among observations. Such challenges are best addressed in subsequent, more formal modeling stages rather than at the exploratory phase.

Acknowledging this important distinction, the conceptual bridge we have drawn between graphical diagnostics and multiple testing remains powerful. It provides a principled language for discussing, comparing, and critiquing different outlier detection rules. More broadly, this framework offers a powerful lens for innovation that extends in two key directions. First, its principles can be applied to other graphical diagnostics; many plots, from Q-Q plots to residual plots, involve multiple visual comparisons that could be enhanced with ideas from adaptive error control. Second, while this paper has focused on univariate data, the core pipeline is highly general. We believe it offers a promising path for creating sample-size-aware boxplots for more complex data structures, such as functional data [7, 18, 24, 29], circular data [1, 6], and curve data or paths

[22, 25]. Exploring these avenues is a rich direction for future research and lies beyond the scope of this paper.

ACKNOWLEDGMENTS

The authors wish to thank the editor, the associate editor and the anonymous reviewer, whose careful reading and constructive suggestions substantially strengthened this work. The authors are listed in alphabetical order. Bowen Gang is the corresponding author. Hongmei Lin's research was supported in part by the National Natural Science Foundation of China (12171310). Tiejun Tong's research was supported in part by the General Research Fund of Hong Kong (HKBU12300123 and HKBU12303421) and the Initiation Grant for Faculty Niche Research Areas of Hong Kong Baptist University (RC-FNRA-IG/23-24/SCI/03).

REFERENCES

[1] ABUZAID, A. H., MOHAMED, I. B. and HUSSIN, A. G. (2012). Boxplot for circular variables. *Computational Statistics* **27** 381–392.

[2] BANERJEE, S. and IGLEWICZ, B. (2007). A simple univariate outlier identification procedure designed for large samples. *Communications in Statistics– Simulation and Computation* **36** 249–263.

[3] BARBATO, G., BARINI, E., GENTA, G. and LEVI, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics* **38** 2133–2149.

[4] BARBATO, G., GENTA, G. and LEVI, R. (2009). Outlier Detection. *CIRP STC P–Precision Engineering and Metrology, Paris*.

[5] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57** 289–300.

[6] BUTTARAZZI, D., PANDOLFO, G. and PORZIO, G. C. (2018). A boxplot for circular data. *Biometrics* **74** 1492–1501.

[7] DAI, W. and GENTON, M. G. (2018). Functional boxplots for multivariate curves. *Stat* **7** e190.

[8] EFRON, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99** 96–104.

[9] GAURAN, I. I. M., PARK, J., LIM, J., PARK, D., ZYLSTRA, J., PETERSON, T., KANN, M. and SPOUGE, J. L. (2018). Empirical null estimation using zero-inflated discrete mixture distributions and its application to protein domain data. *Biometrics* **74** 458–471.

[10] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (2005). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Chichester, UK. https://doi.org/10.1002/9781118186435

[11] HIGGINS, J. P. T., THOMAS, J., CHANDLER, J., CUMPSTON, M., LI, T., PAGE, M. J. and WELCH, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed. John Wiley & Sons, Ltd, Chichester, UK.

[12] HOAGLIN, D. C. and IGLEWICZ, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association* **82** 1147–1149.

[13] HOAGLIN, D. C., IGLEWICZ, B. and TUKEY, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association* **81** 991–999.

[14] HOFMANN, H., WICKHAM, H. and KAFADAR, K. (2017). Letter-value plots: boxplots for large data. *Journal of Computational and Graphical Statistics* **26** 469–477.

[15] HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6** 65–70.

[16] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Chichester, UK. https://doi.org/10.1002/9780470434697

[17] HYNDMAN, R. J. and FAN, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50** 361–365.

[18] HYNDMAN, R. J. and SHANG, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* **19** 29–45.

[19] JIN, J. and CAI, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102** 495–506.

[20] LIN, H., ZHANG, R. and TONG, T. (2025). When Tukey meets Chauvenet: a new boxplot criterion for outlier detection. *Journal of Computational and Graphical Statistics* **in press** 1–21. https://doi.org/10.1002/9781118186435

[21] MARONNA, R. A., MARTIN, R. D., YOHAI, V. J. and SALIBIÁN-BARRERA, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons, Chichester, UK.

[22] MIRZARGAR, M., WHITAKER, R. T. and KIRBY, R. M. (2014). Curve boxplot: generalization of boxplot for ensembles of curves. *IEEE Transactions on Visualization and Computer Graphics* **20** 2654–2663.

[23] PHAM-GIA, T. and HUNG, T. L. (2001). The mean and median absolute deviations. *Mathematical and Computer Modelling* **34** 921–936.

[24] QU, Z. and GENTON, M. G. (2022). Sparse functional boxplots for multivariate curves. *Journal of Computational and Graphical Statistics* **31** 976–989.

[25] RAJ, M., MIRZARGAR, M., RICCI, R., KIRBY, R. M. and WHITAKER, R. T. (2017). Path boxplots: a method for characterizing uncertainty in path ensembles on a graph. *Journal of Computational and Graphical Statistics* **26** 243–252.

[26] SCHWERTMAN, N. C. and DE SILVA, R. (2007). Identifying outliers with sequential fences. *Computational Statistics & Data Analysis* **51** 3800–3810.

[27] SCHWERTMAN, N. C., OWENS, M. A. and ADNAN, R. (2004). A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis* **47** 165–174.

[28] SIM, C. H., GAN, F. F. and CHANG, T. C. (2005). Outlier labeling with boxplot procedures. *Journal of the American Statistical Association* **100** 642–652.

[29] SUN, Y. and GENTON, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics* **20** 316–334.

[30] TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* 448–485. Stanford University Press.

[31] TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

[32] WILSON, E. B. and HILFERTY, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences* **17** 684–688.