Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

Comparing Various Descriptive Statistics for Two Independent Groups in Agricultural Experiments

Zeynel Cebeci[a], Melis Celik Guney[a]*, Ugur Serbester[a]

[a]*Cukurova University, Faculty of Agriculture, Department of Animal Science, Adana, TÜRKIYE*

ARTICLE INFO
Research Article
Corresponding Author: Melis Celik Guney, E-mail: celikm@cu.edu.tr
Received: 25 June 2025 / Revised: 17 August 2025 / Accepted: 22 August 2025 / Online: 20 January 2026

Cite this article
Cebeci Z, Guney M S, Serbester U (2026). Comparing Various Descriptive Statistics for Two Independent Groups in Agricultural Experiments. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 32(1):130-144. DOI: 10.15832/ankutbd.1727052

ABSTRACT
Comparing two groups holds significant importance in statistical analysis across various fields, including agriculture, where performance evaluations are commonly conducted. Conventional parametric and non-parametric tests primarily focus on differences in central tendencies, often neglecting distributional variations, especially across quantiles. This study explores the potential of resampling techniques, such as bootstrapping and permutation tests, in enhancing group comparisons beyond traditional measures like means and medians. Using both simulated and real agricultural datasets, the study demonstrates how quantile-based comparisons supported by resampling methods provide deeper insights into group differences. The results highlight that relying solely on central measures can overlook substantial variations in the distributional tails, emphasizing the necessity of a comprehensive examination of quantiles. Moreover, the findings show that permutation tests, which align with classical methods for central measures, offer a robust, assumption-free alternative that captures higher-order characteristics like skewness and kurtosis. Bootstrapping with the Bias-corrected and Accelerated (BCa) bootstrap confidence intervals also prove to be a reliable tool for estimating sampling distributions. A notable observation is that quantile-based comparisons can detect significant differences in distributional tails, even when conventional tests fail to identify discrepancies in medians. This underscores the power of quantile comparison in unveiling subtle yet critical differences between groups, offering a richer and more precise inferential framework. The study concludes by advocating the adoption of resampling-based quantile comparison methods in agricultural research, as they significantly enhance the detection of distributional differences and bolster the overall rigor of statistical analyses.

Keywords: Bootstrap method, Permutation tests, Quantile comparison, Resampling techniques, T-test

1. Introduction

Comparing two groups is a fundamental statistical analysis to ascertain if differences exist between them. This comparison is not only prevalent in fields such as medicine, social sciences, and engineering, but also in agriculture, food science, and environmental studies. Understanding group differences is crucial as it informs decision-making, policies, interventions, and the efficacy of treatments. In agricultural research, performance trials comparing two independent groups, e.g. a control and a treatment, are frequently employed due to their operational simplicity and practical relevance. For instance, evaluating the efficacy of drip irrigation versus conventional methods in courgette cultivation provides a direct comparison of water management strategies. Furthermore, paired trials, employing a within-subject design, facilitate the assessment of treatment effects on the same individual experimental units. In dairy production, for example, quantifying milk yield before and after the introduction of a novel feed supplement allows for a precise determination of its impact on milk yield. The distinction between these two types of groups hinges on the relationship among their observations. In independent groups, observations are unrelated; subjects in one group offer no relation into those in another, often stemming from random assignment or sampling from distinct populations. Conversely, dependent groups involve observations that are directly linked, typically when the same subjects are measured multiple times (i.e., before-after studies) or when subjects are intentionally paired based on specific characteristics, meaning a direct connection exists between data points across the groups (Wilcox 1995; McCrum-Gardner 2008; Isbell & Jiang 2012). Statistical methodologies for comparing two groups are broadly categorized into parametric and nonparametric approaches, each suited to distinct analytical contexts. Parametric tests, notably the Student's t-test, are preferentially employed when data adhere to specific assumptions, such as normality of distribution and homogeneity of group variances (Hoffman 2015). An important assumption for the t-test when applied to small samples is that the observations within each group should be approximately normally distributed. Another important assumption is homoscedasticity, which is the assumption that the group variances should be equal. If the assumptions are violated, they pose serious practical concerns. These violations can lead to diminished power, inaccurate confidence intervals, and a poor characterization of the differences between groups.

In agricultural research and related disciplines, the analysis of two independent or paired groups with normally distributed data is conducted through parametric t-tests or z-tests, contingent upon sample size and variance status. Conversely, when data deviate from normality, non-parametric methodologies, such as the sign test or Wilcoxon signed-rank test, are employed to assess differences in group medians. However, distributional heterogeneity beyond measures of central tendency, including the mean and median, may be observed, irrespective of normality (Pek et al. 2018). Specifically, equivalent group means and medians do not preclude disparities in quantile values. Although non-parametric tests, such as the Mann-Whitney U (MWU) test, effectively control Type I error rates, they remain susceptible to Type II errors and power deficits (Zimmerman 1994). Moreover, both parametric and non-parametric approaches exhibit limited robustness under severe deviations from normality, especially in the presence of outliers, resulting in a reduction in their capacity to discern distributional variations across quantiles. While non-parametric tests based on median comparisons are robust to distributional assumptions and applicable to ordinal data, they are insufficient for detecting quantile discrepancies. In non-parametric tests, the assumption that the compared groups have similarly shaped distributions is often neglected. However, when distributions differ in skewness or tail behavior, the validity of traditional non-parametric tests can be compromised (Wilcox et al. 2012; Pandis 2015; MacFarland & Yates 2016; Nordstokke & Colp 2018). This underscores the need for more robust alternatives, such as bootstrap-t methods, which are better suited to handle such distributional differences. Finally, the experimental units in each group should be independent, meaning that one should not influence the other.

Despite the availability of normalization techniques to mitigate violations of assumptions such as normality, the stringent requirements for parametric t-tests, such as the assumption of homogeneity of variance, the independence of observations, and the normal distribution of data within each group, are often unmet in real-world data, significantly reducing their applicability (Marozzi 2021). Alternatively, non-parametric methods provide a more robust analytical framework when data deviate from normal distributions or exhibit unequal variances. In instances where the underlying assumptions are not met, nonparametric statistical tests are employed. For independent samples, the MWU test is applied, whereas the Wilcoxon signed-rank sum test is utilized for paired or related samples (Hollander et al. 2013). Non-parametric tests are more flexible than t-tests but are not assumption-free. They require independent observations, measurements on the same scale, unique rankings (no ties), and similar distributional shapes. Their performance weakens in skewed distributions (Hesamian et al. 2023). Tied ranks, especially in small samples, complicate probability calculations, reduce the accuracy of p-values, increase computational complexity, and lead to more conservative results. Therefore, tie-adjustment methods or specialized algorithms are necessary. Although some suggestions or solutions have been proposed to address the tied rank issue, such as the use of rank-based adjustments like the average ranks for tied observations, and the application of correction techniques using the approximation methods (Neuhäuser & Ruxton 2009; Amerise & Tarsitano 2015), the challenge still requires deeper insights.

Classical parametric methodologies, predicated on stringent distributional postulates, risk generating spurious inferences when these criteria are not met, especially in datasets exhibiting skewness or extreme values. While non-parametric alternatives offer increased resilience to such deviations, they are not devoid of assumptions, often presupposing specific distributional forms such as similarly shaped of group distributions in the Wilcoxon Rank Sum (WSRS) test. Both frameworks may exhibit diminished sensitivity in discerning quantile-specific variations, as they do not fully account for the complete shape of the distribution. This can lead to weaker statistical power (Stein 2021; Lazic 2024).

Resampling methodologies circumvent the need to validate the assumptions associated with both parametric and non-parametric statistical tests, enabling a direct focus on the observed data and the inference on derived outcomes. Since it has been introduced in Efron (1979), the bootstrap has become one of the most popular resampling methods. The bootstrap, a non-parametric resampling approach, provides a means of estimating the sampling distribution of a statistic by generating numerous replicate samples through random sampling with replacement from the empirical data (Kostanek et al. 2024). This technique enables the approximation of sampling distributions for diverse statistical measures, thereby enhancing the precision of confidence interval and standard error estimation, particularly in scenarios where analytical derivations of theoretical distributions are either unavailable or computationally challenging. The bootstrap method does not have strict assumptions like parametric tests, which makes it more flexible and powerful. However, there are some very simple assumptions and requirements. Firstly, for bootstrap results to be reliable, the original sample must represent the population well. Therefore, a biased or unbalanced sample can also bias the bootstrap results. Secondly, because the bootstrap relies on resampling (aka repeated sampling) from the original dataset, the sampling units should be drawn independently, meaning that random sampling should be applied. Aside from these two basic requirements, the bootstrap is quite flexible regarding assumptions and can be used even if the data do not follow a normal distribution. This makes bootstrapping particularly useful for complex data structures or small sample sizes (Dai et al. 2009; Wang et al. 2022). The permutation test, which involves randomizing or permuting the observed group data, is a flexible method for making statistical inferences in experimental settings. Permutation tests, in particular, provide a comprehensive approach to statistical inference, especially when there are small number of independent observations. Moreover, permutation test can be applied in scenarios such as multiple groups (Holt & Sullivan 2023). This technique facilitates the determination of statistical significance based on the observed data, eliminating the requirement for pre-specified distributional models. In summary, permutation tests are highly effective at maintaining error rates and offer simplicity in their implementation, since as highlighted by Pauly et al. (2018), bootstrap methods provide flexibility and robustness in estimating variances and confidence intervals. Troendle (2005) reported that bootstrap is particularly useful for estimating confidence intervals and can handle complex data structures, including those with covariates.

The tests described above, which rely on comparing the mean or median, have led to a focus on quantile comparisons to account for the entire distribution pioneered with some implementations by Shoemaker (1995), Pan et al. (2014), Wilcox et al. (2014) and Wilcox & Schönbrodt (2014). Quantiles are descriptive statistics that divide a set of ordered data into equal parts, providing insight into the full distribution rather than just central tendencies like the mean or median. For instance, the median (Q2) splits data into two halves, while quartiles (Q1, Q2, Q3), deciles (D1–D9), and percentiles (P1–P99) divide data into four, ten, or one hundred equal parts, respectively. Quantiles are widely used to estimate population parameters from sample data for example, using the sample median to estimate the population median. Their ability to describe specific locations in a distribution makes them valuable tools for understanding data patterns that averages may obscure. While non-parametric tests offer robustness against distributional assumptions, their inherent focus on summary statistics often obscures the nuanced variations present within specific quantiles, limiting a comprehensive understanding of distributional differences as investigated in several studies (Navruz & Ozdemir 2016; Kaplan & Goldman 2018; Rached & Larsson 2019; Banfi et al. 2022; Hesamian et al. 2023; Hird et al. 2024).

Quantiles can also be used to examine and compare data distributions. In order to determine if a model, such as a normal distribution or a Weibull distribution, fits the population from which the sample is taken, the quantiles of the sample data and the model can be examined. By matching the quantiles of the sample data to the quantiles of a specific probability distribution, a collection of paired data is obtained. These data are plotted in a scatterplot, known as a quantile-quantile (Q-Q) plot. If the resulting plot is approximately linear, the model is considered a good fit for the data. Quantiles are also employed in statistical tests to compare the value of an observed statistic with a theoretical test population of similarly derived statistics. If an observed value falls outside the range of the 95% most typical values, it is classified as significantly different from the others (Marden 2004; Dhar et al. 2014).

As evidenced by the foregoing discussion, while bootstrap and permutation tests present notable advantages in agricultural research compared to parametric t-tests and their less assumption-dependent non-parametric alternatives, the implementation of these resampling techniques in applied settings is currently restricted. When combined with parametric- and non-parametric tests such as the Welch's t-test, MWU test and WSRS test, resampling methods can improve power and handle small sample sizes effectively (Wanishsakpong et al. 2023). The present study elucidates the prospective utility of resampling methodologies in agricultural research through a comparative analysis of results obtained from traditional statistical tests, bootstrap procedures, and permutation testing, utilizing both simulated datasets based on G&H distributions and real experimental data.

Moreover, the comparative assessment of the two groups includes not only measures of central tendency, such as means and medians, but also higher-order distributional characteristics, including variances, skewness, kurtosis, interquartile ranges, and user-chosen quantiles. The aim is to achieve a holistic inferential analysis by underscoring the value of evaluating group disparities beyond central location, extending to distributional tails, dispersion, and form. For a more comprehensive comparison beyond the mean, quantile comparison techniques come into play. These methods utilize location information beyond averages to offer a deeper and more detailed comparison between groups. Leveraging resampling techniques such as permutation tests and bootstrapping can further enhance these comparisons by providing more accurate and reliable inferences. Comparing all quantiles of groups, including the median, promises a more useful solution for testing the differences. A quantile test is a simple statistical method used to compare the quantiles of two groups, which helps in identifying differences across the entire distribution rather than just the central measures like the mean or median.

2. Material and Methods

2.1 Data

In order to perform the statistical tests, live weights (LW, kg) of male lambs of the Karayaka sheep breed at the end of a two-month fattening period were simulated. Three different datasets were generated with G&H distributions using a custom script based on the formula given in Equation 1 in the R statistical computing and graphics software (R Core Team 2025). G&H distributions (Tukey 1977), named after their parameters g and h, are used in statistical modeling to simulate the skewness and kurtosis of data. Utilizing Equation 1 (Tukey 1977), the G&H distribution derives a transformed random variable from a standard normal variable Z, with parameters g and h dictating the resulting distributional shape.

$$X = A + B\ \frac{e^{gZ}-1}{g} \cdot e^{\frac{hZ^2}{2}} \quad (1)$$

Where: n is the number of random samples to generate.; A, B are the location and scale parameters, respectively. A shifts the distribution, and B scales it; g, h are the parameters controlling skewness and kurtosis, respectively; Z denotes standard normal value sampled at random.

The Simulated Dataset 1 (SDS1) contained the simulated live weight observations for two independent groups of lambs (Group A and Group B), each consisting of 25 observations. The descriptive statistics of the groups such as mean, median, and standard deviation are shown in Table 1. The observations in Group A and Group B were normally distributed according to the Shapiro-

Wilk normality test results (P = 0.991 and P = 0.718, respectively). Additionally, the variances of the groups in SDS1 were not significantly different according to the Fmax test and Levene's test applied (P = 0.414 and P = 0.435, respectively).

Table 1- Descriptive statistics for groups in the simulated datasets

Datasets	*Group*	*n*	*Mean*	*SE*	*SD*	*Median*	*Skewness*	*Kurtosis*	*IQR*	*Range*
SDS1	A	25	46.6	0.41	2.04	46.5	-0.017	-0.504	2.767	8.677
	B	25	46.6	0.34	1.72	46.4	0.267	-0.899	2.017	6.219
SDS2	A	25	47.0	0.72	3.61	47.2	0.000	-0.803	5.452	13.954
	B	25	46.9	0.23	1.14	47.0	-0.388	-0.955	1.244	4.015
SDS3	A	25	50.1	2.67	13.34	45.6	2.379	5.141	8.062	57.879
	B	25	46.8	0.19	0.95	46.7	0.035	0.714	1.354	3.755

Figure 1 illustrates the probability density plots, empirical cumulative distribution plots and Q-Q plots for each group in the simulated datasets. These plots provide a comprehensive view of the data distribution and help in identifying any underlying patterns or differences between the groups. As can be understood from the plots in the figure and the statistics and assumption checks, the SDS1 dataset is generated to meet all the assumptions of the parametric t-test.

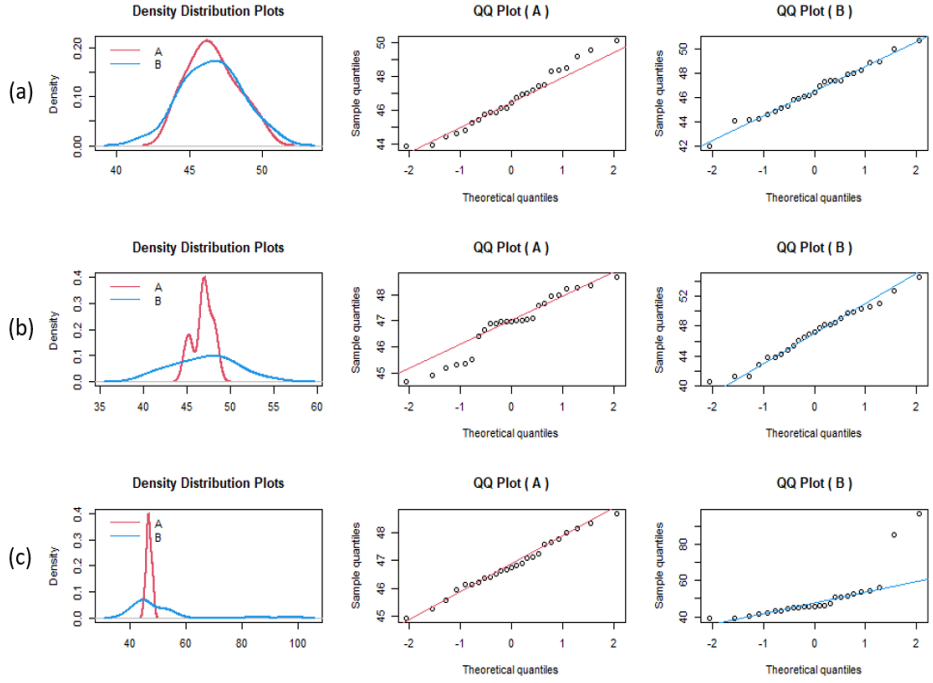


Figure 1- Descriptive plots of the groups in simulated datasets
(a) SDS1 dataset, (b) SDS2 dataset, (c) SDS3 dataset

The Simulated Dataset 2 (SDS2) was generated for the case in which the variances of the groups were different according to Fmax and Levene's tests (P<0.001). As illustrated in Figure 1 (b) and shown in Table 1, the variance of Group B was approximately three times higher than the variance of Group A, while their means and medians were close to each other. Given that SDS2 exhibits significantly different variances between Group A and Group B, a parametric t-test would be problematic because it violates the assumption of homogeneity of variances, which is crucial for the validity of the test.

Finally, in the Simulated Dataset 3 (SDS3) as shown in Figure 1 (c), while Group A was non-normally distributed, Group B was normally distributed according to Shapiro-Wilk test (P = 1.662×10^{-6} and P = 0.989, respectively). Additionally, the variance of Group A was nearly 14 times higher than that of Group B in SDS3. Based on the assumption checks, the SDS3 dataset is also not suitable for applying a parametric t-test.

Table 2 presents descriptive statistics for the amounts of trace element such zinc (RDS1), lead (RDS2), and copper (RDS3) concentrations measured in the milk of lactating cows subjected to two distinct dietary rations.

Table 2- Descriptive statistics for groups in the real datasets

Datasets	*Group*	*n*	*Mean*	*SE*	*SD*	*Median*	*Skewness*	*Kurtosis*	*IQR*	*Range*
RDS1	A	10	0.19	0.005	0.014	0.19	0.154	-1.464	0.018	0.040
	B	10	0.13	0.007	0.021	0.14	-0.426	-1.745	0.038	0.050
RDS2	A	10	0.51	0.016	0.050	0.51	0.088	-1.262	0.063	0.160
	B	10	0.53	0.014	0.043	0.55	-1.781	1.832	0.008	0.140
RDS3	A	10	5.53	0.148	0.468	5.22	0.540	-1.743	0.655	1.050
	B	10	3.40	0.162	0.512	3.15	0.655	-1.563	0.763	1.310

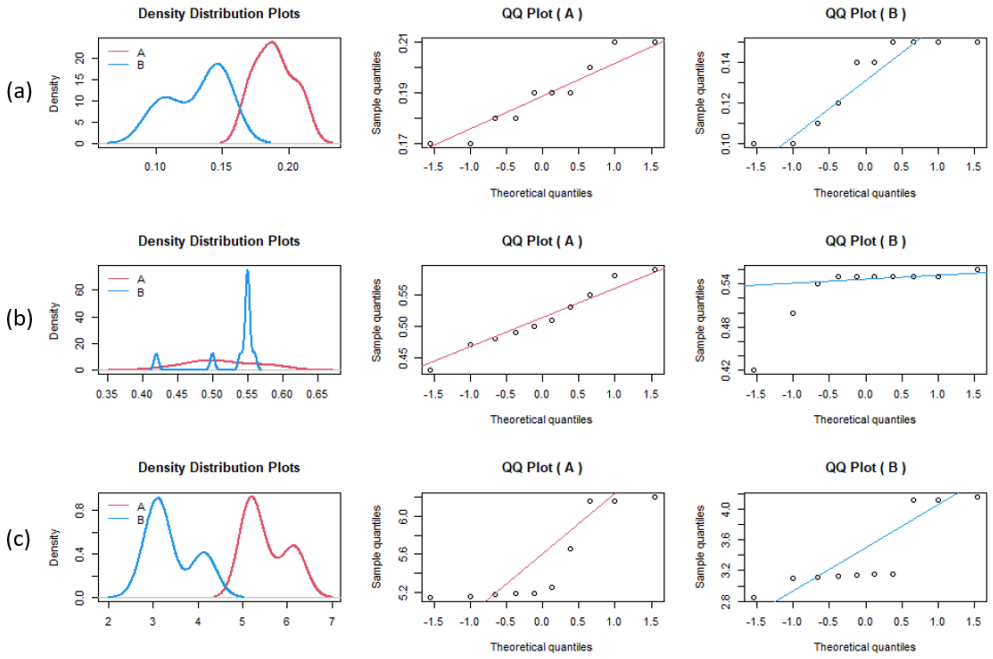


Figure 2- Descriptive plots of the groups in real datasets
(a) RDS1 dataset, (b) RDS2 dataset, (c) RDS3 dataset

As indicated in Figure 2(a), the means of the groups are significantly different in terms of zinc concentration, and there are indications that the data of the second group do not follow a normal distribution. Indeed, according to the Shapiro-Wilk test result, it was determined that Group A follows a normal distribution while Group B does not (P = 0.3296, P = 0.0168, respectively). As a result of assumption checking, the RDS1 dataset is not suitable for applying a parametric t-test due to the violation of one of the basic assumptions.

From the distribution plots in Figure 2(b), it is observed that Group A exhibits a relatively platykurtic distribution, while Group B shows a relatively leptokurtic distribution. As observed in Table 2, the kurtosis statistics for these groups were determined to be -1.262 and 1.832, respectively, which confirms this observation. However, according to both the Fmax and Levene's tests, the variances are equal (P = 0.6376 and P= 0.418, respectively) as presented in Table 5. Consequently, RDS2 violates at least one assumption of the parametric t-test, and exhibits a significant difference in kurtosis.

When examining the plots drawn from the RDS3 dataset, which contains copper composition values in milk, it is possible to conclude that the distribution shapes are very similar, the means are different, but the variances are equal (Figure 2(c)). Similarly, according to the Fmax and Levene's test results for RDS3 provided in Table 5, the variances are homogeneous (P = 0.7957 and P = 0.772, respectively). However, according to the Shapiro-Wilk's normality test, the data of Group A and Group B do not follow a normal distribution (P<0.01). These results indicate that comparing RDS3 using a parametric t-test would be problematic.

2.2. Methods and computing tools

The R statistical software environment (R Core Team 2025) was used for all data simulations and analyses. In this study, the classical t-tests and Welch's t-tests were applied using the t.test function, and the wilcox.test function for non-parametric tests from R's base stats package. In the non-parametric tests applied, average ranks were used to handle tied observations. A custom R package was authored to handle the data simulations and visualizations. Visual analytics serves as a fundamental tool for the exploration of data architecture and the detection of anomalies prior to formal analysis. This need is exemplified by Anscombe's quartet (Anscombe 1973), a seminal demonstration of how identical summary statistics can be generated from disparate data

distributions. The implications of Anscombe's quartet are frequently neglected in empirical research paradigms. Anscombe advocated for the application of exploratory data analysis techniques, including scatterplots and histograms, to ascertain data structural integrity. Moreover, Rousselet et al. (2017) posit that the derivation of precise and efficient descriptive statistics mandates the utilization of comprehensive visual displays. While these recommendations provide invaluable insights into data structure and distributional properties, the implementation of statistical hypothesis tests and robust methodologies is indispensable for quantifying inter-group differences. In this study, a script was coded to create visualizations using group data, and the following bivariate graphs were produced: density plots, empirical cumulative density plots, box-whisker plots, violin plots, Q-Q plots, symmetry plots, and empirical shift plots.

For all the tests mentioned above, the null hypothesis, $H_0: \theta_1 = \theta_2$, was tested against the two-sided alternative hypothesis, $H_1: \theta_1 \neq \theta_2$ at significance level of α = 0.05 (Type I error). Here, θ_1 represents the examined statistics in Group 1 while θ_2 represents the examined statistics in Group 2. Figure 3 outlines a flowchart for selecting statistical tests when comparing two independent groups. The process begins by testing for normality. If both groups are normal, variance homogeneity is assessed: equal variances lead to the independent samples t-test; unequal variances call for Welch's t-test. If the data is non-normal, normalization is attempted; if unsuccessful, alternative methods are used. For data that remains non-normal, non-parametric tests are applied. While less assumption-bound, these still require similar distribution shapes. If shapes align, tests like the MWU (independent) or WSRS (paired) are used; if not, resampling methods such as bootstrap or permutation tests are preferred.

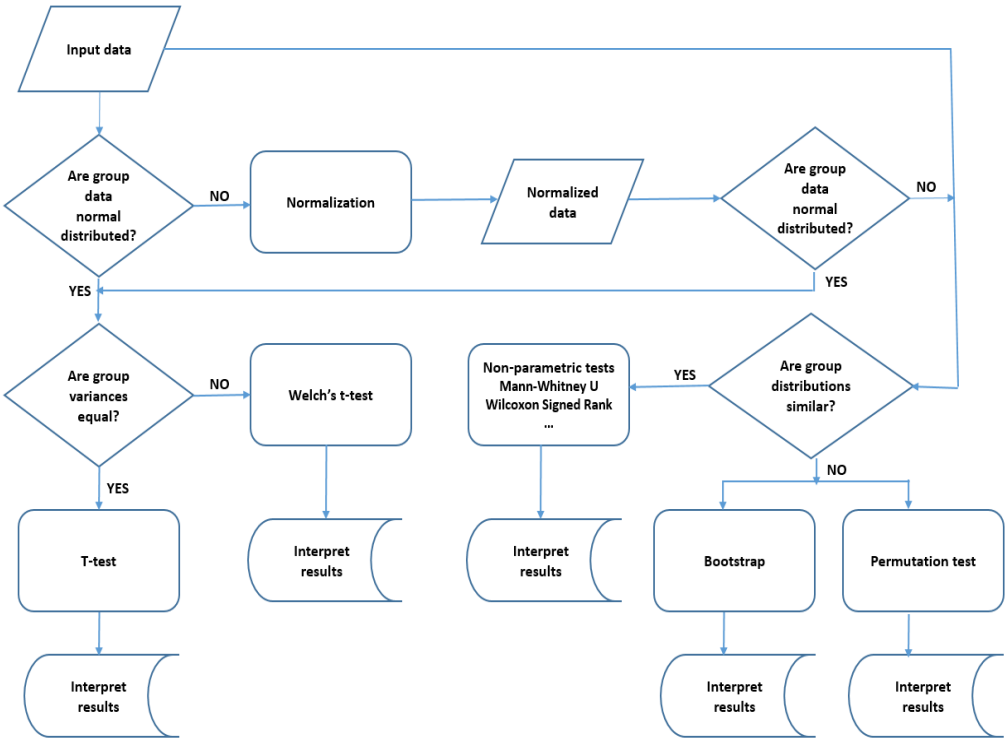


Figure 3- Flowchart for statistical test selection in two-group data analysis

In the comparison of two groups, in addition to the parametric and non-parametric tests based on means and medians, the differences between the groups were also evaluated using permutation tests and bootstrap methods. For this purpose, custom permutation test and bootstrapping scripts were developed to analyze whether the differences between basic descriptive statistics, such as the mean, median, variance, interquartile range, skewness, and kurtosis of the groups, are statistically significant.

The permutation test is a non-parametric statistical method used to evaluate the likelihood that an observed difference between two groups, in terms of a statistic of interest (e.g., mean, median, or skewness), arises purely by chance. The permutation test calculates the p-value to assess whether the observed difference between two groups is attributable to random variation or represents a statistically significant effect. A p-value smaller than the predefined significance level (commonly set at 0.05) indicates that the probability of the observed difference occurring by chance is low, thereby suggesting the presence of a meaningful difference between the groups. The pseudocode in Appendix 1 shows the steps involved in permutation test for the difference of a statistic between two groups.

The bootstrap method is a robust non-parametric statistical technique employed to evaluate differences between two groups by estimating the sampling distribution of a statistic. In contrast to traditional parametric methods that rely on theoretical assumptions such as normality, the bootstrap is particularly advantageous when data do not conform to these assumptions. It operates by repeatedly resampling the observed data, with replacement, to generate bootstrap samples. In this study, a custom

script wrapping the 'boot' function from the 'boot' package in R (Canty & Ripley 2024) in order to test whether two groups differ in terms of mean, median, variance, interquartile range, skewness, and kurtosis. Confidence intervals for the group differences of these statistics were calculated using the Normal, Basic, Percentile, and BCa methods with the 'boot.ci' function of the same R package. The bootstrap procedure in the pseudocode in Appendix 2 was implemented to estimate the sampling distribution of an interested statistic and construct a confidence interval for the difference in this statistic between two independent groups.
The Normal, Basic, Percentile, and BCa are commonly used methods for calculating confidence intervals from results obtained through bootstrapping (Cebeci 2020). If an interval does not contain the value zero, it implies that there is a statistically significant difference between the groups, as zero would represent no difference. As mentioned below, the confidence interval calculation methods have their own unique advantages and disadvantages.

The Normal Confidence Interval (Normal CI) method, detailed in Equation 2 (Jung et al. 2019; Mokhtar et al. 2023; Justus et al. 2024), builds a confidence interval based on the premise that the distribution of differences between examined statistics, the bootstrap replicates approximates a normal distribution.

$$\text{Normal CI} = [\bar{\Delta}^* - z_{\alpha/2} \cdot \text{SE}_{boot},\ \bar{\Delta}^* + z_{\alpha/2} \cdot \text{SE}_{boot}]] \quad (2)$$

Where: $\bar{\Delta}^*$: The mean of the differences of statistics calculated from the bootstrap samples; $z_{\alpha/2}$: The α/2 percentile of the standard normal distribution. For a 95% confidence interval, α = 0.05 and $z_{\alpha/2}$)≈ 1.96; SE_{boot}: The standard error of the statistics calculated from the bootstrap samples.

The Percentile Confidence Interval (Percentile CI), as formulated in Equation 3 (Jung et al. 2019; Mokhtar et al. 2023; Justus et al. 2024), is a straightforward and intuitive method for calculating confidence intervals from the bootstrap results. It directly utilizes the empirical percentiles of the bootstrap distribution to estimate the confidence interval, requiring minimal assumptions about the shape of the underlying distribution. This method is particularly appealing due to its simplicity and ease of implementation, making it a common choice in various applications such as estimating confidence intervals for sample medians. However, it assumes the bootstrap distribution adequately represents the sampling distribution and does not address potential biases or asymmetries, which might limit its accuracy in some cases (Wood 2005; Suriyakat & Panichkitkosolkul 2023).

$$\text{Percentile CI} = [\Delta^*_{(\alpha/2)}, \Delta^*_{(1-\alpha/2)}] \quad (3)$$

Where: $\Delta^*_{(\alpha/2)}$: The α/2 percentile of the bootstrap distribution of the statistic. For example, if α = 0.05 (for a 95% confidence level), then α/2 = 0.025, and this term corresponds to the 2.5% of the sorted bootstrap differences.

$\Delta^*_{(1-\alpha/2)}$: The 1-α/2 percentile of the bootstrap distribution of the statistic. For example, if α = 0.05 (95% confidence level), then 1-α/2 = 0.975, and this corresponds to the 97.5% of the sorted bootstrap differences.

The Basic Bootstrap Confidence Interval (Basic CI), as formulated in Equation 4 (Jung et al. 2019; Mokhtar et al. 2023; Justus et al. 2024), provides a straightforward application of the bootstrap method and is typically favored when the sampling distribution of the statistic of interest is approximately symmetric or when deviations from symmetry are considered negligible. However, unlike the Bias-Corrected and Accelerated Confidence Interval method, which incorporates adjustments for potential biases and skewness, the Basic CI does not account for these distributional asymmetries.

$$\text{Basic CI} = [2 \cdot \Delta_o - \Delta^*_{(1-\alpha/2)}, 2 \cdot \Delta_o - \Delta^*_{(\alpha/2)}] \quad (4)$$

Where: Δ_o: The observed difference in the statistic (e.g., the difference in means between two groups); $\Delta^*_{(1-\alpha/2)}$: The (1-α/2) percentile of the bootstrap distribution; $\Delta^*_{(\alpha/2)}$: The α/2 percentile of the bootstrap distribution.

The Bias-Corrected and Accelerated Confidence Interval (BCa CI), as shown in Equation 5, is an advanced bootstrap method designed to mitigate the effects of bias and skewness within the sampling distribution. It achieves this by integrating a bias correction, which accounts for the discrepancy between the bootstrap median and the observed statistic, and an acceleration correction, which adjusts for the rate of change in standard error as sample size fluctuates. BCa CI in Equation 5 (Jung et al. 2019; Mokhtar et al. 2023; Justus et al. 2024) is particularly advantageous when dealing with non-normal distributions or limited sample sizes, providing superior accuracy and coverage probability compared to simpler bootstrap methods like the Basic CI or Percentile CI. However, its implementation demands greater computational resources, primarily due to the calculation of the acceleration factor, and its performance can be compromised by extreme outliers or severely restricted sample sizes. Ultimately, the BCa CI is the preferred approach when precision and reliability are paramount, especially in intricate data contexts.

$$\text{BCa CI} = [\Delta_{F^{-1}}(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}), \Delta_{F^{-1}}(z_0 + \frac{z_0 - z_{\alpha/2}}{1 - a(z_0 - z_{\alpha/2})})] \quad (5)$$

Where: z_0: The bias correction factor calculated as $z_0 = \Phi^{-1}\left(\frac{count(\Delta_i^* < \Delta_o)}{B}\right)$. It quantifies the discrepancy between the median of the bootstrap distribution and the observed statistic; $\Delta_{F^{-1}}$: The inverse cumulative distribution function (CDF) of the bootstrap distribution; Φ^{-1}: The inverse cumulative density function of the standard normal distribution; a: The acceleration constant, computed via jackknife resampling, which accounts for the rate of change of the standard error of the statistic.

This study adopted a comprehensive approach to evaluating group differences by examining not only measures of central tendency, such as the mean and median, but also specific quantiles across the full distribution. Specifically, the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles (P10-P90), representing increments of 0.1 within the [0.1, 0.9] range, were compared to provide a detailed assessment of distributional disparities beyond central tendency.

Quantile calculations are performed using a range of algorithms, and these implementations differ across statistical software, potentially leading to discrepancies (Kreutzmann 2018). In this study, we implemented the Harrell-Davis Estimator (HDE) to calculate quantiles using a custom script. As a more robust estimator, the HDE, introduced by Harrell & Davis (1982), provides a smoother estimate of the quantile compared to nine common quantile estimation methods included in the 'quantile' function of R. HDE, an L-estimator based on the linear combination of order statistics, reduces the variability associated with quantile estimates, particularly for small sample sizes. Unlike traditional quantile calculations that rely on data from a specific rank, HDE considers a weighted sum of ordered statistics (data points sorted in ascending order). This implies that each data point contributes to the quantile estimate based on a specific weight (Wilcox 1995; Wilcox 2014; Mair & Wilcox 2020). To assess the variability of the data and estimate quantiles (specifically the median), bootstrap samples are utilized. For a given sample of size n, let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ represent the order statistics. The HDE of P-th quantile (Q_P) is then calculated as a weighted sum of these order statistics, as shown in Equation 6 (Harrell & Davis 1982).

$$Q_P = \sum_{i=1} w_{i(P,n)}\, x_{(i)} \tag{6}$$

Where: $w_{i(p,n)}$ are the weights associated with each order statistic, determined based on the P-th quantile and the sample size n, as defined in Equation 7.

$$\int_{F_{i-1}(n)}^{F_i(n)} B\big(u; (n+1)P, (n+1)(1-P)\big)\, du \tag{7}$$

Where: $F_i(n)$ represents the cumulative distribution function (CDF) of the i-th order statistic in a sample of size n, calculated as the cumulative distribution function of the Beta distribution with parameters $a = i$ and $b = n - i + 1$, evaluated at P. $B(u;\, a, b)$ is the Beta density function with parameters a and b, and P is the quantile of interest. The HDE is particularly valuable in statistical analyses where accurate and robust quantile estimates are essential.

3. Results and Discussion

3.1 Simulated datasets

Table 3 demonstrates that while no statistically significant differences in means were detected between Group A and Group B across all simulated datasets (SDS1, SDS2, and SDS3), using both parametric t-test and non-parametric MWU test, the datasets SDS2 and SDS3 exhibit violations of normality and/or homogeneity of variance assumptions. This is despite the simulated equality of means. Consequently, while the means are consistent, these datasets show differences in distribution shape and other measures of location and variation, highlighting the importance of considering non-parametric approaches for further analysis in SDS2 and SDS3.

Table 3- Results of assumption checks and inter-group differences in simulated datasets

Datasets	*Group*	*Shapiro-Wilk (p)*	*F_{max} test (p)*	*Levene test (p)*	*t-test (p)*	*MWU test (p)*
SDS1	A	0.991	0.414	0.435	0.941	0.985
	B	0.718				
SDS2	A	0.946	3.346e-07 ***	2.573e-05 ***	0.854	0.729
	B	0.087				
SDS3	A	1.662e-06 ***	< 2.2e-16 ***	0.0008 ***	0.238	0.280
	B	0.989				

* P<0.05, ** P<0.01, and *** P<0.001

Table 4 presents p-values obtained from permutation tests for various descriptive statistics across three simulated datasets (SDS1, SDS2, and SDS3). In SDS1, all descriptive statistics (mean, median, IQR, variance, skewness, and kurtosis) show non-significant

p-values (P>0.05), indicating that the observed values are likely due to random variation. However, in SDS2 and SDS3, the IQR and variance exhibit highly significant p-values (P<0.001), suggesting that these measures significantly deviate from what would be expected by chance. Additionally, the median in SDS3 shows a significant p-value (P<0.05). These results indicate that while SDS1 demonstrates characteristics consistent with random data, SDS2 and SDS3 exhibit statistically significant deviations in their measures of spread and central tendency, as reflected by the IQR, variance, and, in the case of SDS3, the median.

Table 4- P-values for descriptive statistics from the permutation tests in the simulated datasets

Datasets	*Mean*	*Median*	*IQR*	*Variance*	*Skewness*	*Kurtosis*
SDS1	0.937	1.000	0.390	0.378	0.624	0.596
SDS2	0.857	0.282	0.000 ***	0.000 ***	0.728	0.893
SDS3	0.312	0.029 *	0.000 ***	0.000 ***	0.236	0.231

* P<0.05, ** P<0.01, and *** P<0.001

As previously explained, applying the permutation test analyzes differences in distribution measures and shapes between two groups, thus providing more comprehensive insights into the groups. In fact, as demonstrated in Table 4, the permutation test outcomes for variance and interquartile range (IQR) are consistent with the results of Levene's test and the Fmax test, presented in Table 3. To assess differences in distributions between two groups, the tests such as Kolmogorov-Smirnov, Shapiro-Wilk, and Anderson-Darling can be employed. However, the Kolmogorov-Smirnov test primarily focuses on overall distribution differences and lacks detailed analysis of shape-specific variations. Similarly, the Shapiro-Wilk and Anderson-Darling tests are primarily used to evaluate normality and can only indirectly compare distribution shapes between groups. Conversely, permutation tests and bootstrapping enable direct testing of skewness and kurtosis differences, allowing for a more comprehensive evaluation of shape similarities or differences.

3.2. Real datasets

Table 5 presents the results from assumption checks and statistical tests for inter-group differences in the real datasets RDS1, RDS2, and RDS3. The table includes Shapiro-Wilk tests for normality, Fmax and Levene's tests for variance homogeneity, and t-tests and MWU tests for group differences. In RDS1, Group B shows a significant deviation from normality (Shapiro-Wilk P<0.05). Despite this, Levene's test indicates homogeneity of variances. Notably, both the t-test and MWU test reveal significant differences between Groups A and B (P<0.001), indicating a true difference in the distribution of the groups. In RDS2, Group B again fails the normality test (Shapiro-Wilk P<0.001). However, both variance tests are non-significant, suggesting homogeneity. Neither the t-test nor the MWU test shows a significant difference between the groups. In RDS3, both groups fail the normality test. The variance tests are non-significant, again suggesting homogeneity. While the t-test shows no significant difference, the MWU test shows a highly significant difference between the groups. Overall, the results highlight that the choice of test (parametric vs. non-parametric) can significantly impact the conclusions drawn, particularly when normality assumptions are violated. The MWU test, a non-parametric alternative, detects differences in RDS1 and RDS3 where the t-test does not, indicating that the groups may differ in ways other than their means (e.g., distribution shape).

Table 5- Results of assumption checks and inter-group differences in real datasets

Datasets	*Group*	*Shapiro-Wilk (p)*	*F_{max} test (p)*	*Levene test (p)*	*t-test (p)*	*MWU test (p)*
RDS1	A	0.3296	0.2656	0.051	1.247e-06 ***	0.000164 ***
	B	0.0168 *				
RDS2	A	0.9145	0.6376	0.418	0.3733	0.2794
	B	5.238e-05 ***				
RDS3	A	0.002139 **	0.7957	0.772	0.9985	0.0001786 ***
	B	0.001739 **				

* P<0.05, ** P<0.01, and *** P<0.001

Table 6 shows p-values obtained from permutation tests for various descriptive statistics in the datasets RDS1, RDS2, and RDS3. The results indicate varying degrees of statistical significance across the datasets. In RDS1, the mean and median exhibit highly significant p-values (P<0.001), suggesting that these measures are significantly different from what would be expected by chance. The IQR, variance, skewness, and kurtosis show non-significant p-values (P>0.05). In RDS2, the skewness and kurtosis show significant p-values (P<0.01), indicating deviations from what would be expected by chance. The mean, median, IQR, and variance show non-significant p-values. In RDS3, the mean and median show highly significant p-values (P<0.001 and P<0.01, respectively), while the IQR, variance, skewness, and kurtosis show non-significant p-values. As an overall evaluation, the results reveals that RDS1 and RDS3 exhibit significant differences in their measures of central tendency (mean and median), while RDS2 shows significant differences in measures of shape (skewness and kurtosis).

Table 6- P-values for descriptive statistics from the permutation tests in the real datasets

Datasets	*Mean*	*Median*	*IQR*	*Variance*	*Skewness*	*Kurtosis*
RDS1	0.0000 ***	0.0008 ***	0.3676	0.6743	0.2281	0.5261
RDS2	0.4092	0.1712	0.0621	0.6216	0.0034**	0.0066 **
RDS3	0.0000 ***	0.0019 **	0.9890	0.9409	0.8229	0.8229

* P<0.05, ** P<0.01, and *** P<0.001

3.3. Comparison of quantiles

The quantile comparisons for SDS1 presented in Table 7 indicate that there are no differences between the groups, both in terms of the median and other quantiles. On the other hand, when applying the permutation tests for SDS2, there are significant differences between groups in the lower and upper percentiles such as P10, P20, P70, P80 and P90 (P<0.05) but not for the other quantiles as observed in Table 8. In Figure 1, this phenomenon is even apparent in the distribution plots for SDS2, where Group A and Group B data have been simulated with markedly different distributions. As shown in Table 9, the quantile comparison with permutation test revealed strong significant differences for the lower quantiles P10, P20, P30 and upper quantiles P70, P80, P90.

While all bootstrap confidence intervals yielded findings consistent with the permutation test results for SDS1, only the BCa confidence interval fully agreed with the permutation test results for SDS2. For this dataset, as illustrated in Table 8, Basic CI, Normal CI, and Percent CI yielded inferences consistent with the permutation tests at the P10, P80, and P90 percentiles, with the Basic CI also demonstrating consistency at the P70 percentile. Upon examing the bootstrap results for the SDS3 presented in Table 9, it is evident that the Normal CI did not detect any significant percentile differences. Conversely, the Basic CI, Percent CI, and BCa CI detected significant differences at the lower percentiles P10, P20, and P30, and the upper percentiles P80 and P90, which are in agreement with the permutation test results. The observed poor performance of the Normal CI in this dataset, as depicted in Figure 1, is attributed to the simulation of one group's data with an exceptionally steep and narrow distribution, coupled with the simulation of the other group's data with high variability and a right-skewed distribution. This disparity results in a non-normal bootstrap sampling distribution.

The detection of differences for one or more quantiles confirms that group data may appear different at lower and upper end of distributions even if the group medians are not different. Therefore, quantile comparisons can provide more informative insights indicating the tendency to variational locations.

Table 7- Comparison of quantile differences of the groups in SDS1

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	43.942	44.314	-0.372	[-1.629, 1.117]	[-1.446, 1.257]	[-2.002, 0.701]	[-2.207, 0.593]	0.577
P20	44.728	45.013	-0.284	[-1.550, 0.937]	[-1.573, 0.895]	[-1.464, 1.005]	[-1.517, 0.953]	0.669
P30	45.388	45.585	-0.197	[-1.477, 1.008]	[-1.522, 0.960]	[-1.353, 1.129]	[-1.407, 1.050]	0.768
P40	45.997	46.019	-0.022	[-1.256, 1.219]	[-1.238, 1.235]	[-1.278, 1.195]	[-1.283, 1.190]	0.971
P50	46.585	46.455	0.130	[-1.093, 1.422]	[-1.032, 1.456]	[-1.197, 1.291]	[-1.156, 1.317]	0.851
P60	47.146	46.942	0.204	[-1.019, 1.530]	[-0.940, 1.621]	[-1.213, 1.348]	[-1.167, 1.396]	0.758
P70	47.648	47.505	0.143	[-1.203, 1.494]	[-1.186, 1.530]	[-1.245, 1.471]	[-1.285, 1.439]	0.833
P80	48.298	48.221	0.078	[-1.474, 1.527]	[-1.508, 1.485]	[-1.330, 1.663]	[-1.373, 1.606]	0.923
P90	49.389	49.155	0.234	[-1.349, 1.788]	[-1.346, 1.772]	[-1.305, 1.813]	[-1.244, 1.869]	0.794

A and B are the quantile values for Group 1 and Group 2, respectively; The values between the square brackets are the lower and upper limits of confidence intervals by the CI methods; The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

Table 8- Comparison of quantile differences of the groups in SDS2

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	41.843	45.117	-3.274	**[-5.318, -1.631**]	**[-5.511, -1.964]**	**[-4.584, -1.037]**	**[-4.867, -1.356]**	0.015*
P20	43.580	45.699	-2.119	[-4.290, 0.062]	[-4.308, 0.063]	[-4.302, 0.069]	**[-4.448, -0.113]**	0.040*
P30	44.907	46.401	-1.494	[-3.749, 0.536]	[-3.767, 0.499]	[-3.487, 0.779]	[-3.758, 0.555]	0.100
P40	46.144	46.835	-0.691	[-2.750, 1.226]	[-2.663, 1.258]	[-2.639, 1.282]	[-2.820, 1.094]	0.342
P50	47.232	46.998	0.234	[-1.489, 2.076]	[-1.358, 2.231]	[-1.763, 1.826]	[-1.753, 1.845]	0.643
P60	48.158	47.165	0.993	[-0.566, 2.749]	[-0.479, 2.857]	[-0.870, 2.466]	[-0.643, 2.638]	0.077
P70	49.090	47.508	1.581	[-0.002, 3.275]	**[0.076, 3.305]**	[-0.142, 3.087]	**[0.014, 3.259**]	0.011*
P80	50.103	47.933	2.169	**[0.387, 3.854]**	**[0.258, 3.842]**	**[0.496, 4.081]**	**[0.614, 4.270]**	0.007**
P90	51.786	48.295	3.491	**[1.380, 5.698]**	**[1.376, 5.526]**	**[1.456, 5.606]**	**[1.883, 6.076]**	0.003**

A and B are the quantile values for Group 1 and Group 2, respectively; The values between the square brackets are the lower and upper limits of confidence intervals by the CI methods (significant results marked bold); The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

Table 9- Comparison of quantile differences of the groups in SDS3

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	40.387	45.542	-5.155	[-7.561, -3.377]	**[-7.901, -3.873]**	**[-6.438, -2.41]**	**[-6.700, -3.033]**	0.000***
P20	42.493	46.068	-3.575	[-5.980, -1.331]	**[-5.875, -1.372]**	**[-5.778, -1.274]**	**[-6.121, -1.517]**	0.000***
P30	44.051	46.340	-2.289	[-4.223, -0.224]	**[-3.983, -0.068]**	**[-4.510, -0.595]**	**[-4.778, -0.757]**	0.000***
P40	45.040	46.559	-1.519	[-3.604, 0.408]	[-4.020, 0.286]	[-3.323, 0.983]	[-3.420, 0.766]	0.005**
P50	46.026	46.776	-0.750	[-3.911, 1.715]	[-4.717, 0.901]	[-2.402, 3.216]	[-2.381, 3.253]	0.182
P60	47.894	47.030	0.864	[-3.356, 4.524]	[-3.617, 3.559]	[-1.831, 5.346]	[-1.788, 5.460]	0.229
P70	50.622	47.355	3.267	[-2.598, 8.639]	[-3.849, 7.626]	[-1.092, 10.382]	[-1.027, 10.777]	0.010**
P80	54.209	47.740	6.469	[-6.873, 16.439]	[-12.345, 11.839]	**[1.100, 25.284]**	**[2.201, 31.098]**	0.008**
P90	69.768	48.195	21.572	[-0.064, 44.264]	**[0.282, 38.628]**	**[4.517, 42.863]**	**[5.850, 46.524]**	0.001***

A and B are the quantile values for Group 1 and Group 2, respectively; The values between the square brackets are the lower and upper limits of confidence intervals by the CI methods; The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

Table 10 reveals statistically significant differences between groups A and B in RDS1 across all percentiles according to both permutation tests and bootstrapping. The differences are relatively consistent and supported by confidence intervals and p-values. This supports the conclusion that the distributions of the two groups are distinct and not due to random chance (P<0.001).

Table 10- Comparison of quantile differences of the groups in RDS1

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	5.157	2.947	2.210	**[2.035, 2.450]**	**[2.102, 2.400]**	**[2.020, 2.318]**	**[2.038, 2.438]**	0.000***
P20	5.170	3.053	2.117	**[1.902, 2.302]**	**[1.906, 2.246]**	**[1.906, 2.246]**	**[1.989, 2.329]**	0.000***
P30	5.193	3.113	2.080	**[1.770, 2.325]**	**[1.722, 2.325]**	**[1.834, 2.437]**	**[1.727, 2.339]**	0.000***
P40	5.244	3.148	2.096	**[1.648, 2.506]**	**[1.576, 2.574]**	**[1.619, 2.616]**	**[1.623, 2.618]**	0.000***
P50	5.364	3.222	2.143	**[1.547, 2.759]**	**[1.519, 2.824]**	**[1.462, 2.766]**	**[1.544, 2.817]**	0.000***
P60	5.575	3.394	2.181	**[1.469, 2.969]**	**[1.513, 3.000]**	**[1.363, 2.849]**	**[1.414, 2.878]**	0.000***
P70	5.836	3.672	2.164	**[1.398, 2.984]**	**[1.447, 2.986]**	**[1.342, 2.881]**	**[1.370, 2.900]**	0.000***
P80	6.055	3.956	2.099	**[1.398, 2.779]**	**[1.300, 2.786]**	**[1.412, 2.899]**	**[1.446, 2.929]**	0.000***
P90	6.165	4.113	2.052	**[1.490, 2.566]**	**[1.105, 2.610]**	**[1.495, 3.000]**	**[1.501, 3.005]**	0.000***

A and B are the quantile values for Group 1 and Group 2, respectively; The values between square brackets are the lower and upper limits of confidence intervals by the CI methods (significant results marked bold); The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

Table 11 reveals no statistically significant differences between groups A and B in RDS2 across all percentiles. The differences are small and supported by confidence intervals and p-values. This supports the conclusion that the distributions of the two groups are similar and likely due to random chance. Table 11 reveals a nuanced observation at the P40 percentile, where the Normal CI, Basic CI, and BCa CI indicate statistically significant differences. This finding aligns with the visual trends observed in the distribution plots of Figure 2, suggesting a potential divergence between the groups around this specific percentile in the second group data. The ability of bootstrap methods to identify these subtle differences underscores their enhanced sensitivity in capturing variations in group quantiles, particularly in regions where traditional methods might overlook statistical significance.

Table 11- Comparison of quantile differences of the groups in RDS2

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	0.447	0.456	-0.009	[-0.086, 0.087]	[-0.071, 0.089]	[-0.106, 0.054]	[-0.096, 0.067]	0.681
P20	0.468	0.501	-0.033	[-0.107, 0.037]	[-0.109, 0.029]	[-0.095, 0.043]	[-0.088, 0.054]	0.397
P30	0.484	0.532	-0.048	[-0.113, 0.001]	[-0.125, -0.009]	[-0.087, 0.029]	[-0.089, 0.024]	0.140
P40	0.496	0.544	-0.049	**[-0.102, -0.012]**	**[-0.111, -0.002]**	[-0.077, 0.014]	**[-0.087, -0.004]**	0.083
P50	0.509	0.548	-0.040	[-0.086, -0.006]	[-0.093, -0.012]	[-0.068, 0.013]	[-0.073, 0.004]	0.075
P60	0.524	0.549	-0.026	[-0.069, 0.013]	[-0.072, 0.008]	[-0.060, 0.021]	[-0.059, 0.022]	0.103
P70	0.543	0.550	-0.007	[-0.047, 0.037]	[-0.044, 0.036]	[-0.051, 0.029]	[-0.045, 0.033]	0.513
P80	0.565	0.552	0.012	[-0.020, 0.058]	[-0.010, 0.064]	[-0.039, 0.035]	[-0.026, 0.038]	0.477
P90	0.582	0.556	0.026	[-0.002, 0.067]	[0.012, 0.082]	[-0.030, 0.039]	[-0.011, 0.040]	0.470

A and B are the quantile values for Group 1 and Group 2, respectively; The values between square brackets are the lower and upper limits of confidence intervals by the CI methods (significant results marked bold); The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

Table 12 presents compelling evidence of statistically significant disparities between groups A and B within the RDS3 dataset, observed across all quantiles examined. The uniformity of these findings, substantiated by significant results from the confidence intervals and consistently small p-values derived from the permutation tests, rigorously supports the robustness and reliability of

the observed differences. This data implies a systematic and substantial divergence in the distributional characteristics of the two groups, indicative of distinct underlying processes or factors influencing their respective quantile distributions.

Table 12- Comparison for quantile differences of the groups in RDS3

Q	*A*	*B*	*A-B*	*Normal CI*	*Basic CI*	*Percentile CI*	*BCa CI*	*P.val*
P10	5.157	2.947	2.210	**[2.035, 2.450]**	**[2.102, 2.400]**	**[2.020, 2.318]**	**[2.038, 2.438]**	0.000***
P20	5.170	3.052	2.117	**[1.902, 2.302]**	**[1.906, 2.246]**	**[1.989, 2.329]**	**[2.019, 2.428]**	0.000***
P30	5.193	3.113	2.080	**[1.770, 2.325]**	**[1.722, 2.325]**	**[1.834, 2.437]**	**[1.727, 2.339]**	0.000***
P40	5.244	3.148	2.096	**[1.648, 2.506]**	**[1.576, 2.574]**	**[1.619, 2.616]**	**[1.623, 2.618]**	0.000***
P50	5.364	3.222	2.143	**[1.547, 2.759]**	**[1.519, 2.824]**	**[1.462, 2.766]**	**[1.544, 2.817]**	0.000***
P60	5.575	3.394	2.181	**[1.469, 2.969]**	**[1.513, 3.000]**	**[1.363, 2.849]**	**[1.414, 2.878]**	0.000***
P70	5.836	3.672	2.164	**[1.398, 2.984]**	**[1.447, 2.986]**	**[1.342, 2.881]**	**[1.370, 2.900]**	0.000***
P80	6.055	3.956	2.099	**[1.398, 2.779]**	**[1.300, 2.786]**	**[1.412, 2.899]**	**[1.446, 2.929]**	0.000***
P90	6.165	4.113	2.052	**[1.490, 2.566]**	**[1.105, 2.610]**	**[1.495, 3.000]**	**[1.501, 3.005]**	0.000***

A and B are the quantile values for Group 1 and Group 2, respectively; The values between square brackets are the lower and upper limits of confidence intervals by the CI methods (significant results marked bold); The values in P.val column are the calculated *p*-values with permutation tests; * P<0.05, ** P<0.01, and *** P<0.001

4. Conclusions

In this study, we illustrated how bootstrap and permutation tests address common challenges by drawing examples from the agricultural field. We went to beyond the widely used assumption-based parametric and non-parametric tests and help make permutation tests and bootstrap confidence intervals more visible as a flexible framework for statistical inference. This study also demonstrated how effective quantile comparisons are in identifying differences between groups, even when traditional parametric and non-parametric tests fail due to their assumptions. Based on the results obtained from both artificial and real datasets, we can draw the following conclusions:

I. Inferences regarding medians and means do not represent the lower and upper tails of the distribution, thereby failing to fully capture the differences between the two groups. Consequently, rather than solely testing central measures of distribution, such as medians and means, it is more informative to examine the entire distribution across all quantiles. This approach ensures a more comprehensive inference concerning the groups.

II. With a few exceptions, the permutation test results for means and medians align with the results of classical tests. This demonstrates that performing a permutation test alone is sufficient, eliminating the need to rely on classical parametric and non-parametric tests that require assumption checking. Furthermore, performing permutation tests not only for the mean and median but also for statistics such as IQR, variance, skewness, and kurtosis would enable a more comprehensive comparison of the two groups regarding both variability and central tendencies combined.

III. As an alternative, or rather an additional option, bootstrap has yielded similar inferences to permutation tests. Comparing the confidence intervals to permutation test results reveals that the BCa CI method is more effective than other confidence interval methods. The findings of this study indicate that, from a practical standpoint, the Percentile CI and the analogously performing Basic CI exhibit performances approximating those of the BCa CI. However, the Normal CI method should only be employed when the sampling distribution is normal or approximately normal.

IV. Permutation tests and bootstrapping enable direct testing of skewness and kurtosis differences, allowing for a more comprehensive evaluation of shape similarities or differences.

V. We acknowledge that interpreting specific quantile estimates, especially those in the extreme tails, can sometimes require careful consideration. However, we believe the benefits of a comprehensive view of the entire distribution often outweigh this complexity. Quantile comparisons provide a more detailed understanding of data by capturing variability across the entire distribution, not just at the center, proving especially useful for analyses focusing on extreme values. For instance, in plant protection research, they help uncover how treatments affect specific subgroups, such as those with delayed recovery or severe side effects. This analysis allows researchers to monitor the progress of both low-performing and high-performing groups, offering a more detailed and equitable assessment of outcomes along the entire distribution. While quantile estimates can be sensitive to small sample sizes, particularly in the tails of the distribution, leading to potential instability, this challenge can be mitigated by employing robust methods. In this study, we estimated the quantiles using the Harrell-Davis Quantile Estimator which is known for its improved efficiency and stability in such situations.

VI. The analyses show that, even in cases where no statistically significant differences are detected between two groups using the t-test and the MWU test, quantile comparisons can provide additional insights by revealing significant differences in the lower and upper quantiles, where the groups' distributions differ markedly. The results emphasize the

importance of considering quantile differences, as significant variations can be detected even when group medians are not different. This finding highlights the potential of quantile comparisons to capture more subtle differences in data distributions, which may be overlooked by traditional methods. The ability to detect such differences is especially valuable in scenarios where the underlying data distributions are expected to vary between groups. As a general conclusion, the utilization of bootstrap and permutation-based methodologies for quantile comparison facilitates a rigorous and assumption-minimized approach to the identification of statistically significant distributional heterogeneity across the entire quantile range. This framework effectively elucidates disparities that may remain undetected by traditional parametric and non-parametric statistical procedures, thereby enhancing the precision and comprehensiveness of comparative analyses. Although our aim was to demonstrate the applicability of bootstrap and permutation approaches under realistic data limitations where traditional methods may fall short. In this context, focusing on large-scale datasets may be the primary objective of the future works.

Author Declarations

Authors' contributions: Concept: Z.C., Design: Z.C., Data collection: U.S., Analysis: Z.C., M.C.G., Writing: Z.C., M.C.G., U.S.

Conflict of interest/Competing interests: The author(s) declare no conflict of interest.

Ethics approval/declarations: No need.

Funding: This research received no external funding.

Consent to participate: No need.

Consent for publication: No need.

Data availability: Data are available on request due to privacy or other restrictions.

References

Amerise I L & Tarsitano A (2015). Correction methods for ties in rank correlations. Journal of Applied Statistics, 42(12): 2584–2596.

Anscombe F J (1973). Graphs in statistical analysis. The American Statistician 27(1): 17-21

Banfi F, Cazzaniga G & De Michele C (2022). Nonparametric extrapolation of extreme quantiles: a comparison study. Stochastic Environmental Research and Risk Assessment 36(6): 1579-1596

Canty A & Ripley B (2024). boot: Bootstrap R (S-Plus) functions. R package version 1.3-31

Cebeci Z (2020). Applied Resampling Techniques Using R. 2nd Ed., Pegem Akademi Yayıncilik, Ankara. 650 p. ISBN: 9786257052603 (In Turkish).

Dai S, Gao H M & Xiao Z C (2009). Research on data analysis of small sample based on bootstrap method. *Journal of Naval Acronautical and Astronautical University* 24(1): 27-30

Dhar S S, Chakraborty B & Chaudhuri P (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. Bernoulli 20(3): 1484-1506

Efron B (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7(1): 1-26

Harrell F E & Davis C E (1982). A new distribution free quantile estimator. Biometrika 69(3): 635-640

Hesamian G, Chukhrova N & Johannssen A (2023). Statistical inference on quantiles of two independent populations under uncertainty. Soft Computing 27(23): 17573-17583

Hird C, Barham K E & Franklin C E (2024). Looking beyond the mean: quantile regression for comparative physiologists. *Journal of Experimental Biology* 227(5)

Hoffman J (2015). Chapter 22-Comparison of two groups: t-tests and nonparametric tests. Biostatistics for Medical and Biomedical Practitioners, pp. 337-362

Hollander M, Wolfe D A & Chicken E (2013). Nonparametric statistical methods. John Wiley & Sons.

Holt C A & Sullivan S P (2023). Permutation tests for experimental data. Experimental Economics 26(4): 775-812

Isbell D R & Jiang X (2012). Comparing two related samples. The Encyclopedia of Applied Linguistics pp. 1-7

Jung K, Lee J, Gupta V & Cho G (2019). Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation. Frontiers in Psychology, 10, 2215

Justus V L, Rodrigues V B & Sousa A R D S (2024). Bootstrap confidence intervals: A comparative simulation study. arXiv preprint arXiv:2404.12967.

Kaplan D M & Goldman M (2018). Comparing distributions by multiple testing across quantiles or CDF values, Working Papers 1801, Department of Economics, University of Missouri.

Kostanek J, Karolczak K, Kuliczkowski W & Watala C (2024). Bootstrap method as a tool for analyzing data with atypical distributions deviating from parametric assumptions: Critique and effectiveness evaluation. Data 9(8): 95

Kreutzmann A K (2018). Estimation of sample quantiles: challenges and issues in the context of income and wealth distributions. AStA Wirtschafts-und Sozialstatistisches Archiv 12(3): 245-270

Lazic S E (2024). Ditching the norm: Using alternative distributions for biological data analysis. Laboratory Animals, 58(5): 438-442

Mair P & Wilcox R (2020). Robust statistical methods in R using the WRS2 package. Behavior Research Methods 52: 464-488

Marden J I (2004). Positions and QQ plots. Statistical Science 19(4): 606-614

Marozzi M (2021). A combined bootstrap test for the two-sample location problem. *Journal of Statistical Computation and Simulation* 91(1): 180-196

McCrum-Gardner E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery* 46(1): 38-41

MacFarland T W & Yates J M (2016). Mann–Whitney U test. In Introduction to nonparametric statistics for the biological sciences using R (pp. 103-132). Cham: Springer International Publishing.
Mokhtar S F, Yusof Z M & Sapiri H (2023). Confidence intervals by bootstrapping approach: a significance review. *Malaysian Journal of Fundamental and Applied Sciences* 19(1): 30-42
Navruz G & Ozdemir A F (2016). Comparison of two independent groups by using the lower and upper quantiles and percentile bootstrap. Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering 17(3): 506-511
Neuhäuser M & Ruxton G D (2009). Round your numbers in rank tests: exact and asymptotic inference and ties. Behavioral Ecology and Sociobiology 64: 297-303
Nordstokke D W & Colp S M (2018). A note on the assumption of identical distributions for nonparametric tests of location. Practical Assessment, Research & Evaluation 23(3): 1-9
Pan Y, Caudill S P, Li R & Caldwell K L (2014). Median and quantile tests under complex survey design using SAS and R. Computer Methods and Programs in Biomedicine, 117(2): 292-297
Pandis N (2015). Nonparametric methods. American Journal of Orthodontics and Dentofacial Orthopedics 148(4): 695.
Pauly M, Umlauft M & Unlu A (2018). Resampling-based inference methods for comparing two coefficients alpha. Psychometrika 83: 203-222
Pek J, Wong O & Wong A C (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. Frontiers in Psychology 9, 2104
R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rached I & Larsson E (2019). Tail distribution and extreme quantile estimation using non-parametric approaches. High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet pp. 69-87
Rousselet G A, Pernet C R, Wilcox R R (2017). Beyond differences in means: robust graphical methods to compare two groups in neuroscience. European Journal of Neuroscience, 46(2): 1738-1748
Shoemaker L H (1995). Tests for differences in dispersion based on quantiles. The American Statistician 49(2): 179-182
Stein M L (2021). Parametric models for distributions when interest is in extremes with an application to daily temperature. Extremes 24(2): 293-323
Suriyakat W & Panichkitkosolkul W (2023). Nonparametric bootstrap confidence intervals for the population mean of a zero-truncated Poisson-Lindley distribution and their application. *Journal of Current Science and Technology* 13(3): 725-733
Troendle J F (2005). Multiple comparisons between two groups on multiple Bernoulli outcomes while accounting for covariates. Statistics in Medicine 24(23): 3581-3591
Tukey J W (1977). Modern techniques in data analysis. In Proc. of the NSF-Sponsored Regional Research Conference (Vol. 7). North Dartmouth, MA, USA: Southern Massachusetts University.
Wang Z, Peng L & Kim J K (2022). Bootstrap inference for the finite population mean under complex sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(4): 1150-1174
Wanishsakpong W, Sodrung K & Thongteeraparp A (2023). A comparison of nonparametric statistics and bootstrap methods for testing two independent populations with unequal variance. *International Journal of Analysis and Applications* 21: 36-36
Wilcox R R (1995). Comparing two independent groups via multiple quantiles. *Journal of the Royal Statistical Society: Series D (The Statistician)* 44(1): 91-99
Wilcox R R, Erceg-Hurn D M, Clark F & Carlson M (2012). Comparing two independent groups via the lower and upper quantiles. Journal of Statistical Computation and Simulation 84(7): 1543-1551.
Wilcox R R & Schönbrodt F D (2014). The WRS package for robust statistics in R (version 0.24). Retrieved from http://r-forge.r-project.org/projects/wrs/
Wood M (2005). Bootstrapped confidence intervals as an approach to statistical inference. Organizational Research Methods, 8(4): 454-470
Zimmerman D W (1994). Note on the influence of distribution shape on nonparametric tests. Perceptual and Motor Skills, 79(3): 1160-1162

Appendix 1. Pseudocode for the steps involved in permutation test

Input:
Group 1: $X = \{x_1, x_2, \dots, x_n\}$ // n observations
Group 2: $Y = \{y_1, y_2, \dots, y_m\}$ // m observations
Statistic function: $\theta(data)$ // e.g., mean, median, variance, skewness …
Number of permutations: B // e.g., 20000
Significance level: α // 0.05 for a 5% significance level

Algorithm:
Compute the original statistic $\theta(X)$ // Statistic for Group 1
Compute the original statistic $\theta(Y)$ // Statistic for Group 2
Calculate the observed difference: $\Delta_o = \theta(X) - \theta(Y)$
Pool the observed data
Combine all observations: $W = X \cup Y = \{w_1, w_2, \dots, w_{n+m}\}$
Perform permutation test // Iterative process
For i = 1 to B:
Randomly shuffle W to generate a permuted data, W_i^*
Divide W_i^* into two groups:
X_i^*, the first n observations in W_i^*
Y_i^*, remaining m observations in W_i^*
Calculate permutations statistics

Compute the statistic $\theta(X_i^*)$ // Permutation statistic for Group 1
Compute the statistic $\theta(Y_i^*)$ // Permutation statistic for Group 2
Compute the difference: $\Delta_i^* = \theta(X_i^*) - \theta(Y_i^*)$
Merge the differences from all permutations: $\{\Delta_1^*, \Delta_2^*, \ldots, \Delta_B^*\}$

Output:

Calculate the *p*-value:
Count the number of permuted differences for $|\Delta_i^*| \geq |\Delta_o|$ and store in b
Calculate *p*-value: $p = b/B$

Inference:

Assess statistical significance:
If $p \leq \alpha$, reject the null hypothesis that there is no difference between the groups.

Appendix 2. Pseudocode for the steps involved in bootstrap

Input:

Group 1: $X = \{x_1, x_2, \ldots, x_n\}$ // *n* observations
Group 2: $Y = \{y_1, y_2, \ldots, y_m\}$ // *m* observations
Statistic function: $\theta(data)$ // e.g., mean, median, variance, skewness …
Number of bootstrap iterations: B // e.g., 10000
Confidence level: $1 - \alpha$ // 0.95 for a 95% confidence interval

Algorithm:

Compute the original statistic $\theta(X)$ // Statistic for Group 1
Compute the original statistic $\theta(Y)$ // Statistic for Group 2
Calculate the observed difference: $\Delta_o = \theta(X) - \theta(Y)$
Perform bootstrap // Iterative process
For i = 1 to B:
Draw a bootstrap sample X_i^* by sampling *n* observations with replacement from X
Draw a bootstrap sample Y_i^* by sampling *m* observations with replacement from Y
Calculate bootstrap statistics
Compute the statistic $\theta(X_i^*)$ // Bootstrap statistic for Group 1
Compute the statistic $\theta(Y_i^*)$ // Bootstrap statistic for Group 2
Compute the difference: $\Delta_i^* = \theta(X_i^*) - \theta(Y_i^*)$
Create the distribution of bootstrap differences: $\{\Delta_1^*, \Delta_2^*, \ldots, \Delta_B^*\}$

Output:

Compute confidence intervals (CIs)
Calculate CIs using the methods such as Normal, Basic, Percentile, BCa
For example, with the percentile method:
Determine the $\alpha/2$ and $1 - \alpha/2$ percentiles of the sorted bootstrap distribution
Obtain percentile CI as $[\Delta_{\alpha/2}^*, \Delta_{1-\alpha/2}^*]$ at $(1 - \alpha)\%$ confidence level

Inference:

Assess statistical significance:
If a CI does not contain zero, reject the null hypothesis at the α significance level.

Copyright © 2026 The Author(s). This is an open-access article published by Faculty of Agriculture, Ankara University under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.