Persistent confusion in nutrition and obesity research about the validity of classic nonparametric tests in the presence of heteroscedasticity: evidence of the problem and valid alternatives

Cynthia M Kroeger,1,2 Keisuke Ejima,2 Bridget A Hannon,3 Tanya M Halliday,4 Bryan McComb,5 Margarita Teran-Garcia,3 John A Dawson,6 David B King,4 Andrew W Brown,7 and David B Allison2

1Charles Perkins Centre, Central Clinical School, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia; 2Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA; 3Division of Nutritional Sciences, University of Illinois, Urbana-Champaign, Urbana, IL, USA; 4Department of Health, Kinesiology, and Recreation, University of Utah, Salt Lake City, UT, USA; 5Division of Biostatistics, New York University School of Medicine, New York, NY, USA; 6Department of Nutritional Sciences, Texas Tech University, Lubbock, TX, USA; and 7Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA

ABSTRACT

The use of classic nonparametric tests (cNPTs), such as the Kruskal–Wallis and Mann–Whitney U tests, in the presence of unequal variance for between-group comparisons of means and medians may lead to marked increases in the rate of falsely rejecting null hypotheses and decreases in statistical power. Yet, this practice remains prevalent in the scientific literature, including nutrition and obesity literature. Some nutrition and obesity studies use a cNPT in the presence of unequal variance (i.e., heteroscedasticity), sometimes because of the mistaken rationale that the test corrects for heteroscedasticity. Herein, we discuss misconceptions of using cNPTs in the presence of heteroscedasticity. We then discuss assumptions, purposes, and limitations of 3 common tests used to test for mean differences between multiple groups, including 2 parametric tests: Fisher’s ANOVA and Welch’s ANOVA; and 1 cNPT: the Kruskal–Wallis test. To document the impact of heteroscedasticity on the validity of these tests under conditions similar to those used in nutrition and obesity research, we conducted simple simulations and assessed type I error rates (i.e., false positives, defined as incorrectly rejecting the null hypothesis). We demonstrate that type I error rates for Fisher’s ANOVA, which does not account for heteroscedasticity, and Kruskal–Wallis, which tests for differences in distributions rather than means, deviated from the expected significance level. Greater deviation from the expected type I error rate was observed as the heterogeneity increased, especially in the presence of an imbalanced sample size. We provide brief tutorial guidance for authors, editors, and reviewers to identify appropriate statistical tests when test assumptions are violated, with a particular focus on cNPTs. Am J Clin Nutr 2021;113:517–524.

Keywords: nonparametric tests, heteroscedasticity, research rigor, statistical methods, causation, association, nutrition, obesity

Introduction

In science (1–3) generally, and nutrition and obesity research specifically (4–6), concerns have been raised about the extent to which research methods are rigorous, meet modern standards, and can be expected to produce reproducible results. Whether such concerns reflect a growing lapse in rigor or simply a greater recognition of the gaps that have riddled science since its beginning remains uncertain. What is clear, however, is that there is always room for improvement, and scientific standards for increasing rigor, reproducibility, and transparency are developing in a positive direction. In 2015, the NIH updated its instructions for grant applications to refine the standards for biomedical research (7). Various suggestions for improvements have also recently been published within the nutrition and obesity fields (8–10).

One aspect that merits further attention is the use of statistical procedures in a manner that leads to valid inferential testing. Inappropriate choices of statistical models or incorrect implementation of statistical procedures in nutrition and obesity research have been noted elsewhere (4, 11, 12). In addition to perpetuating confusion within the scientific process, errors of...
this type can cultivate mistrust in nutrition science (8). Here, we address one area where improvements in the choice of statistical tests, and in the understanding of their validity, may be warranted: the validity (or lack thereof) of using classic nonparametric tests (cNPTs) to test for differences in central tendency (e.g., means or medians) between groups when heteroscedasticity is present. Heteroscedasticity means a difference in statistical dispersion between groups. Variance is one measure of statistical dispersion and is the measurement of heteroscedasticity we use here. Supplemental Figure 1 depicts a simple example of heteroscedasticity between 2 groups. Note that we do not discuss single-sample or paired-sample tests here. We define cNPTs as tests that do not rely on any assumptions about the data having a particular distribution, other than having a finite mean and a finite variance. The terms nonparametric and distribution-free do not always refer to the same thing. A distribution-free test is a test that does not depend on the data having any particular distributional form (e.g., the data do not need to fit a normal distribution). In contrast, the term nonparametric has historically been used in multiple ways. Herein, we refer to nonparametric in a way synonymous with distribution-free [cf. Hollander et al. (13)]. Further, cNPTs (as opposed to nonparametric tests in general) are statistical procedures that do not involve explicit resampling, thereby excluding bootstrap, randomization, or permutation testing, and are usually rank-based.

The aims of the present article are to 1) describe various misconceptions about cNPTs in the context of heteroscedasticity, 2) simulate the impact of such use on the validity of significance testing, and 3) provide a tutorial on how to conduct cNPTs validly. We focus on tests for ≥3 groups; however, our arguments also apply to tests for 2 groups.

**Misconceptions of Using Nonparametric Tests in the Presence of Heteroscedasticity**

The main misconceptions about cNPTs are that they are “assumption free” or “less powerful than their parametric counterparts” (4). To be clear, when the assumptions of the parametric test (e.g., normality and equal variance for Student’s t test and ANOVA; see Table 1) are met, the parametric test will generally be the most powerful test available. When the assumptions of the parametric test are not met, nonparametric tests will often (but not always) be more powerful than parametric tests, depending on multiple factors including the magnitude of violation of test assumptions of parametric tests and the sample size (14–16). However, nonparametric tests are generally sensitive to distributional differences beyond mean or median differences between groups. Thus, one needs to ensure that there is no heteroscedasticity when using cNPTs (that is, when there is homoscedasticity). Further, this highlights the difference between cNPTs and other nonparametric tests, such as bootstrap tests and permutation tests, which do not require the homoscedasticity assumption.

A common, yet incorrect, procedure used to test for a difference in means or medians between 2 groups is as follows:

1) A test of normality is conducted for each group, or for the combined residuals from a model including a term for group, with the Kolmogorov–Smirnov or Shapiro–Wilk test (valid).
2) If the normality assumption is supported, an $F$ test is used to test for heteroscedasticity (valid).
   a) If the homoscedasticity assumption is supported, Student’s $t$ test is used (valid). It is important to note that our use of the word “valid” here does not necessarily mean “optimal.” The optimal choice of test depends on many factors (17).
   b) If the homoscedasticity assumption is not supported, the Mann–Whitney $U$ test is used (incorrect).
3) If the normality assumption is not supported, Mann–Whitney $U$ is used (valid, if homoscedasticity is present).

The challenge with this erroneous procedure, particularly step 2b, is that Student’s $t$ test assumes that the 2 groups follow a normal distribution with the same variance (i.e., homoscedastic). The Mann–Whitney $U$ test actually tests for differences in distributions, and therefore does not assume any specific distributional shape. Rather, it assumes the 2 groups are drawn from populations with identical distributions under the null hypothesis, including mean, median, variance, and shape. The only way that the Mann–Whitney $U$ test can test for differences in central tendency (e.g., means and medians), then, is for the distributions of the 2 groups to be identical except that the central tendency is different. If the test is used improperly (e.g., in the presence of heteroscedasticity), a Mann–Whitney $U$ test can be statistically significant even with identical central tendency. Thus, cNPTs are not a cure for heteroscedasticity, and generally should not be used in the presence of important heteroscedasticity.

With large sample sizes and higher statistical power, highly statistically significant evidence for a departure from the assumptions (normality and homoscedasticity) does not mean that the magnitude of assumption violation is large or concerning. Therefore, the analyst should attend to the magnitude of the apparent heteroscedasticity and not just its statistical significance.

The use of cNPTs in cases of heteroscedasticity is of particular concern in nutrition and obesity research, where heteroscedasticity is common. Unequal variances between groups can occur for multiple reasons. In a study of diet and exercise, for example, participants in a control group may continue their usual diet and exercise routine, whereas those in the intervention group vary on the spectrum of compliance to noncompliance over time. This could lead to a smaller variance in body weight change scores in the control group than in the intervention group (18). Diet interventions may have the opposite effect when the outcome of interest is dietary diversity. A control group consuming a usual diet may demonstrate higher variance than a more restrictive intervention group.

Assumptions, Purposes, and Limitations of 3 Common Tests

This section describes misconceptions about cNPTs in the presence of heteroscedasticity by comparing them with their parametric counterparts. We focus on 3 tests for differences between ≥3 groups: 1) Fisher’s ANOVA, usually referred to as just ANOVA or Fisher’s $F$ test, which is a parametric test; 2) the Kruskal–Wallis test, a cNPT; and 3) Welch’s ANOVA, a variant of Fisher’s ANOVA that addresses heteroscedasticity. Table 1 summarizes assumptions and the hypothesis of each test. The descriptions here are basic overviews; interested readers should refer to more detailed discussions for more information (19, 20).

Fisher’s ANOVA

Fisher’s ANOVA refers to a parametric analysis of variance. Fisher’s ANOVA is often what researchers try first with continuous data by default and is the subject of many basic statistical tutorials. It is a powerful and useful test but can have undesirable statistical properties when basic conditions or assumptions are violated. Fisher’s ANOVA assumes normality of residuals, independence of observations, and homogeneity of variance. Although an equal sample size is not required, unequal sample sizes among groups can exacerbate the bias in type I and type II errors in the presence of heteroscedasticity, as we discuss later in our simulations.

The fact that unequal group variances and sample sizes occur commonly in practice has led some to suggest that Fisher’s ANOVA should not be used at all, instead moving directly to Welch’s ANOVA (described later). Despite the challenges with Fisher’s ANOVA, 2 aspects allow it to remain useful:

1) the denominator df is larger than Welch’s ANOVA, and
2) although Fisher’s ANOVA develops serious problems (bias, type I error rate increase) in the presence of heteroscedasticity, data can potentially be transformed to fix variance heterogeneity, for example, by using a Box–Cox transformation (21); common transformations like log, square root, and inverse are special cases of the Box–Cox transformation.

The first point means that, when assumptions are met, Fisher’s ANOVA is more powerful for testing hypotheses than other tests like Welch’s ANOVA. The second point means that when $\geq$1 assumption, such as homogeneity of variance, is not met, it may be possible to transform data to be used in this more powerful statistical approach. When these assumptions and conditions do not hold, other tests should be used.

Kruskal–Wallis test

When the assumption of homogeneity of variance fails, some researchers erroneously substitute a nonparametric test, such as the Kruskal–Wallis test, that relies on the ordered ranks of the data. The first step of the Kruskal–Wallis test is to sort the data from smallest to largest, regardless of the group that the observations belong to, and replace the actual observed data with their sorted ranks. The advantage of the Kruskal–Wallis test is that no assumptions about the specific distributions of the data need to be made for the test procedure to be valid, other than that whatever the distributions are, they are the same for all groups being compared. However, as a general rule, both Kruskal–Wallis and Fisher’s ANOVA tend to increase the type I error rate in the presence of heteroscedasticity with equal sample size (20).

Kruskal–Wallis can be less powerful than either Fisher’s or Welch’s ANOVA for 2 reasons. First, the Kruskal–Wallis test procedure loses statistical information in the first step when the actual data are replaced by sample ranks. Second, the Kruskal–Wallis test has a more ambitious goal than either Fisher’s or Welch’s ANOVA. Whereas Fisher’s and Welch’s ANOVA evaluate differences in central tendency (e.g., means), the Kruskal–Wallis test compares differences in distributions between groups, including central tendency, variance, and other distributional characteristics. By definition, if we know that the groups have
differences in variances, then we know that the groups have
different distributions (although not necessarily differences in
central tendency), and an appropriately powered, but possibly
inappropriately used, Kruskal–Wallis test is expected to show
a statistically significant result, even if the measure of central
tendency is equal. Thus, the belief that nonparametric test
procedures, such as Kruskal–Wallis or Mann–Whitney U, are not
influenced by heterogeneous variances in the data is erroneous.
Published articles, however, continue to state that Kruskal–Wallis
tests were used when the equal variance assumption was violated.
In many cases, utilizing the Kruskal–Wallis test when variance
heterogeneity is present will result in an increased type I error
rate for testing differences in central tendency, often expressed at
the median (10, 12).

**Welch’s ANOVA**

Welch’s ANOVA (sometimes called Welch’s F test) is
designed to be robust in the presence of normality and when
the homogeneity of variance assumption does not hold. Welch’s
ANOVA employs 2 innovations compared to Fisher’s ANOVA.
First, it weights observations within each group by the inverse
of the estimated group variances. That is, observations that come
from groups with higher variances have smaller weights than do
observations that come from groups with small variances. The
second innovation of Welch’s ANOVA is that the denominator
df employs the Welch–Satterthwaite adjustment to approximate
the F-distribution. Because Welch’s ANOVA is based upon the
general linear model, it can be viewed as a generalization of
Fisher’s ANOVA that is robust to the heterogeneity of variance.
Because Welch’s ANOVA places weights on observations
from groups that are inversely proportional to the variance in
each group, the effect of heterogeneity is mitigated. Hence,
the procedure is better able to preserve the type I error
rate.

One disadvantage of Welch’s ANOVA is that it estimates
group variances in addition to group means. Thus, this model
spends more df than Fisher’s ANOVA, and the denominator df for
Fisher’s ANOVA will always be larger than for Welch’s ANOVA.
As a result, Welch’s ANOVA tends to have lower power when the
heteroscedasticity assumption holds. Thus, the gain in robustness
(i.e., useful when there is violation of homoscedasticity) comes at
the cost of statistical power. Although it is important to recognize
that Welch’s ANOVA is robust to heterogeneity in variance, it
may still be useful to use the Kruskal–Wallis model in some
circumstances, such as when the variances are homogeneous, but
the residual distribution is not normal.

**Simulated Impact of Using Nonparametric Tests in
the Presence of Heteroscedasticity**

To document the effect of heteroscedasticity on the validity of
cNPT under conditions similar to those used in nutrition and
obesity research, we conducted simple simulations and assessed
the empirical probability of rejecting the null hypothesis (H0) of
no differences in central tendency between groups (i.e., the type
I error rate). In the simulations, samples were randomly sampled
from normal distributions with the same mean and different
variances between groups. Because it is known that imbalanced
sample sizes influence the type I error rate in the presence
of heteroscedasticity, we separately performed simulations with
balanced and imbalanced sample sizes. **Supplemental Figure 2**
summarizes the overall simulation process.

**Data generation**

A single data set was composed of 40 samples for each of 5
groups (i.e., same sample size), randomly sampled from normal
distributions with mean 0 (i.e., equal means) and group-specific
variances: Norm(0, σi2) (i = 1, . . . , 5). We set the SDs to
maximize the inflation of type I error for Fisher’s ANOVA under
equal sample sizes as proved by Box (22) by setting the SD of 4 of
the groups to 1 and the other group ≥ 1 (σ1, σ2, σ3, σ4, σ5) =
{1, 1, 1, 1, a} , where a ≥ 1). The data generation and statistical
tests (see below) were repeated 10,000 times for each parameter
setting (i.e., varying maximum SD: a = {1, 3, 5, 7, 9}).

Furthermore, to investigate the effect of unequal sample
sizes among groups on type I error rate when combined with
heteroscedasticity, we designed the simulation with 2 groups,
2 different SDs ( {σ1, σ2} = {1, 2}), and a fixed total sample
size of 60: Case A: [n1, n2] = [20, 40], Case B: [n1, n2] =
{30, 30}, and Case C: [n1, n2] = {40, 20}.

**Statistical tests**

We performed 3 different tests on each of the 10,000 generated
data sets: Kruskal–Wallis test, Fisher’s ANOVA (i.e., with equal
variance assumption) (23), and Welch’s ANOVA (i.e., without
equal variance assumption) (23). Type I error rate was defined as
the number of simulations with P value (pi; i = 1, 2, . . . , N)
below significance level (α = 0.1, 0.05, 0.01): ˆp = \sum_{i=1}^{N} I(p_i < α)/N, where I is the identity function. The 95% CI of empirical
type I error rates (ˆp) was computed as ˆp ± 1.96/√N(1− p)/N.
For data generation and statistical analysis, we used the statistical
computing software R 3.3.1 (24). The code is available in our
study repository (25).

**Results**

The type I error rate of Fisher’s 1-factor ANOVA and
Kruskal–Wallis increased as the maximum SD (a) increased
(that is, greater heteroscedasticity), whereas that of Welch’s 1-
factor ANOVA was preserved even when a increased (**Figure
1**). Although the deviation was not clearly observed for the
inappropriate tests under an equal sample size, this simply means
the error is minimized if the sample size is equal under the
same parameter settings. The results do not mean the deviation
is negligible even under the same sample size, as we have
demonstrated in **Figure 1** (equal sample size but different SD).
Deviation of type I error rate from the nominal significance level
was most profound with unequal sample sizes in different groups
(**Figure 2**). The type I error rate increased when the sample size
of the group with larger variance was smaller and inappropriate
tests (Kruskal–Wallis and Fisher’s ANOVA) were used. It is worth
noting that Fisher’s ANOVA with balanced sample sizes (i.e.,
30;30) and different SDs appears to preserve the type I error rate;
however, this is because of the relatively small difference in SE
and the small number of groups.
Nonparametric tests and heterogeneous variances

FIGURE 1  Type I error rate with maximum SD for 3 different nominal significance levels. The type I error rates with the 95% CIs for 3 tests (Fisher’s ANOVA, Welch’s ANOVA, and Kruskal–Wallis) are shown. Each panel corresponds to the cases with different significance levels ($\alpha$): 0.1, 0.05, and 0.01. Each simulation included 5 groups, with $n = 40$ per group. The simulation was repeated 10,000 times. Type I error rate was defined as the number of simulations with $P$ value ($p_i; i = 1, 2, \ldots, n$) below significance level ($\alpha$): $\hat{p} = \sum_{i=1}^{N_i} I(p_i < \alpha)/n$. The 95% CI of type I error rates ($\hat{p}$) was computed as $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. Each statistical test was of the overall null hypothesis of no differences in means between groups. Although Kruskal–Wallis tests for differences among distributions, researchers sometimes use it to test for differences in central tendency, which is invalid when there is heteroscedasticity.

We acknowledge 2 limitations of our approach. We used 0.05 as a prespecified significance level because it is frequently used. Although threshold-based inference like “statistical significance” at the 0.05 level has limitations, using type I error permits us to investigate the influence of incorrect analysis on the distribution of $P$ values. If the type I error rate does not match the expected rate, it implies that the distribution of the $P$ values is not what it should be, that the expected value of the $P$ values is also unlikely to be what it should be, and therefore that the tests will be biased in general, regardless of whether one uses specific significance thresholds. The degree and direction of bias may depend on the significance level used. Given that $P$ values and estimates of uncertainty, like CIs, use the same underlying math, incorrect type I error rates also imply biased CIs. Furthermore, we have not discussed the influence of heteroscedasticity on power or type II error (failure to reject null hypotheses that are actually wrong). Given that Welch’s ANOVA considers the difference in variance, whereas Fisher’s ANOVA does not,

FIGURE 2  Type I error rate with imbalanced sample sizes in the presence of heteroscedasticity. The type I error rates with the 95% CIs for 3 tests (Fisher’s ANOVA, Welch’s ANOVA, and Kruskal–Wallis) are shown. Different colors correspond to different settings for sample size: dark gray $[n_1, n_2] = [20, 40]$, gray $[n_1, n_2] = [30, 30]$, and light gray $[n_1, n_2] = [40, 20]$. SD: $[\sigma_1, \sigma_2] = [1, 2]$; significance level is 0.05. The simulation was repeated 10,000 times. Type I error rate was defined as the number of simulations with $P$ value ($p_i; i = 1, 2, \ldots, n$) below significance level ($\alpha$): $\hat{p} = \sum_{i=1}^{N} I(p_i < \alpha)/n$. The 95% CI of type I error rates ($\hat{p}$) was computed as $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. Each statistical test was of the overall null hypothesis of no differences in means between groups. Although Kruskal–Wallis tests for differences among distributions, researchers sometimes use it to test for differences in central tendency, which is invalid when there is heteroscedasticity.
FIGURE 3 Diagnostic checks for an ANOVA model. Key questions, hypothesis tests, and visual plots that could be conducted whenever one considers diagnostics associated with any ANOVA model–based analysis. Note that each of these assessments is dependent on a sufficiently large sample, just like the overall statistical test of interest.

Independence

Key Question:
Did your data collection process involve repeated measures on same subjects or experimental units?

Tests for independence if observation time or observation sequence is known:
1. Wald-Wolfowitz Runs Test
2. Durbin-Watson Autocorrelation Test

Visual plots that might assist diagnosis:
1. Plot residuals vs observation or data collection time and look for trends in residuals.

If data are not independent then you should consult statistician who can run longitudinal models which account for autocorrelation in data.

Constant Variance

Tests for equal group variance:
1. Utilize the Levene Test or the Brown-Forsythe modification to the Levene test on the residuals. This method does not depend on normality of residuals.
2. If residuals are Normally Distributed one can utilize the Breusch-Pagan Test.

Visual plots that might assist diagnosis:
1. Plot residuals vs fitted values and look for "fan" shaped spread changes in dots around zero.
2. Create box plots and/or strip plots of residuals by group and look for width changes.

Normality

Tests for normality of residuals:
1. Conduct a correlation test between residual quantiles and theoretical quantiles from QQ plot
2. Shapiro-Wilk Normality test
3. Kolmogorov-Smirnov test
4. Anderson-Darling Test
5. Cramer-von Mises Test
6. Lilliefors Test

Visual plots that might assist diagnosis:
1. Create QQ-plots of residuals. Look for departures from the 45° line
2. Histogram plots or nonparametric density plots of residuals.

Omission of Other Important Predictor Variables

Key Question:
Does the model exclude important variables which are readily available in the data?

Tests for Variable Omission:
1. Build another linear model with the other variable(s) included in the model and then conduct likelihood ratio tests to see if these variables can be removed.

Visual plots that might assist diagnosis:
1. Plot the residuals from the ANOVA model versus the other potential excluded predictor variables and look for trends.

The power of Welch’s ANOVA could be theoretically lower when heteroscedasticity is not present. Thus, the choice of appropriate tests is dependent on the balance between type I and II error rates, which should be done on a case-by-case basis.

Tutorial of Valid Statistical Procedures in the Context of Heteroscedasticity

Some experts who analyze their data but are not formally or extensively trained in statistics may not be as familiar with the intricacies and assumptions involved with statistical methods. To assist in such situations, we present a flowchart that can provide guidance on which statistical method is best to use. These charts are meant as guides, not as edicts for the user to follow uncritically. The charts can be thought of as a framework for the structured thinking that one should engage in when deciphering which tool is best used and in which circumstance. Moreover, this guidance is not a comprehensive assembly of all procedures that could be applied when comparing groups (26–28).

In Figure 3, we display guidance on some diagnostics checks that should be used whenever one evaluates whether any ANOVA model is appropriate. We present 4 main diagnostics for the ANOVA model with accompanying tools and visuals to evaluate these checks: Independence, Constant Variance, Normality, and Omission of Other Important Predictor Variables. Note that most of these tools and visuals are readily available in most statistical software but require prior planning by the investigator to implement successfully. The implementation of these guidelines should consider the characteristics of the statistical tests in addition to the magnitude of heteroscedasticity (e.g., the difference in variances between groups). In very large samples, for instance, tests of heteroscedasticity may be statistically significant even if the difference in variances is trivial. However, if the magnitude of the difference in variances between groups is statistically significant but trivial, the influence of the choice of statistical test (Fisher’s ANOVA rather than Welch’s ANOVA) on the type I error rate may also be trivial, and the choice of statistical test (Welch’s ANOVA rather than Fisher’s ANOVA) may also include considerations for loss of power. In contrast, if the sample size is small, tests for heteroscedasticity may not be powerful enough to detect even meaningful amounts of heteroscedasticity if it exists. Therefore, investigators should consider adopting statistical methods that allow for heteroscedasticity and nonnormality when sample sizes are small, and consider the magnitude of difference in variance between groups, not only the significance of tests of heteroscedasticity, to decide on tests to use when sample sizes are large.

Figure 4 contains a decision tree along with some remedial measures and various models that could be fit to the data to obtain the most valid analysis. In the case that the data neither are normally distributed nor have equal variance, a nonparametric bootstrapping approach could be used to test for significant differences in the means. Nonparametric bootstrapping approaches do not rely on distributional assumptions, because they rely on the empirical distribution (29).

Conclusion

We introduced the assumptions of the statistical tests for the difference in central tendency (e.g., Fisher’s ANOVA, Welch’s ANOVA, and Kruskal–Wallis) while discussing the misconception of cNPTs (Kruskal–Wallis) among those tests. Our simulations indicated that Kruskal–Wallis and Fisher’s 1-factor ANOVA tests provide (in many cases positively) biased type I error rates in the presence of heteroscedasticity, whereas Welch’s 1-factor ANOVA does not. We further provided a tutorial on selecting statistical tests under heteroscedasticity. More training in these common errors may be needed for nutrition and obesity researchers.
FIGURE 4  Decision tree for statistical inference. A decision tree containing some remedial measures that could be taken to analyze data in which group mean comparisons or tests of whether samples come from the same distribution are desired. ¹See (30–33). ²See Boos and Brownie (34) and Miller (35) for bootstrap and jackknife methods, respectively.

The authors’ responsibilities were as follows—CMK, KE, BAH, TMH, MT-G, DBK, and DBA: designed the research; CMK, KE, BAH, TMH, and DBA: conducted the research; CMK and KE: analyzed the data; CMK, KE, BAH, BM, JAD, AWB, DBK, and DBA: wrote the paper; CMK, KE, AWB, and DBA: had primary responsibility for the final content; and all authors: read and approved the final manuscript. The authors report no conflicts of interest.

Data Availability

Data described in the article, code book, and analytic code will be made publicly and freely available under the Creative Commons Attribution-ShareAlike license at https://zenodo.org/record/3625431 (study repository DOI: 10.5281/zenodo.3625431).

References

1. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA. A manifesto for reproducible science. Nat Hum Behav 2017;1:0021.
2. Hsieh T, Vaickus MH, Remick DG. Enhancing scientific foundations to ensure reproducibility: a new paradigm. Am J Pathol 2018;188(1):6–10.
3. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. Nature 2016;530(7588):27–9.
4. George BJ, Beasley TM, Brown AW, Dawson J, Dimova R, Divers J, Goldsby TU, Heo M, Kaiser KA, Keith SW, et al. Common scientific and statistical errors in obesity research. Obesity 2016;24(4):781–90.
5. Byrne JL, Yee T, O’Connor K, Dyson MP, Ball GD. Registration status and methodological reporting of randomized controlled trials in obesity research: a review. Obesity 2017;25(4):665–70.
6. Brown AW, Ioannidis JPA, Cope MB, Bier DM, Allison DB. Unscientific beliefs about scientific topics in nutrition. Adv Nutr 2014;5(5):563–5.

7. National Institutes of Health. Implementing rigor and transparency in NIH & AHRQ research grant applications [Internet]. Rockville, MD: NIH Office of Extramural Research; 2015. Available from: https://grants.nih.gov/grants/guide/notice-files/not-od-16-011.html. [accessed 22 December 2020].

8. Kroeger CM, Garza C, Lynch CJ, Myers E, Rowe S, Schneeman BO, Sharma AM, Allison DB. Scientific rigor and credibility in the nutrition research landscape. Am J Clin Nutr 2018;107(3):484–94.

9. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. Proc Natl Acad Sci U S A 2018;115(11):2563–70.

10. Wood AC, Wren JD, Allison DB. The need for greater rigor in childhood nutrition and obesity research. JAMA Pediatr 2019;173:311.

11. George BJ, Brown AW, Allison DB. Errors in statistical analysis and questionable randomization lead to unreliable conclusions. J Paramed Sci 2015;6(3):153–4.

12. Mehta T, Allison DB. From measurement to analysis reporting: grand challenges in nutritional methodology. Front Nutr 2014;1(6):6.

13. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2015.

14. Blair RC, Higgins JJ. A comparison of the power of Wilcoxon’s rank-sum statistic to that of Student’s t statistic under various nonnormal distributions. J Educ Stat 1980;5(4):309–35.

15. Zimmerman DW. Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. J Exp Educ 1987;55(3):171–4.

16. Wilcox RR. Introduction to robust estimation and hypothesis testing: statistical modeling and decision science. 3rd ed. London: Academic Press; 2011.

17. Nussbaum EM. Categorical and nonparametric data analysis: choosing the best statistical technique. New York: Routledge; 2015.

18. Kaiser KA, Gadbury GL. Estimating the range of obesity treatment response variability in humans: methods and illustrations. Hum Hered 2013;75(2–4):127–35.

19. Feir-Walsh BJ, Toothaker LE. An empirical comparison of the Anova F-Test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educ Psychol Meas 1974;34(4):789–99.

20. Tomarken AJ, Serlin RC. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol Bull 1986;99(1):90–9.

21. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc B 1964;26(2):111–52.26.

22. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. Ann Math Statist 1954;25(2):290–302.

23. Welch BL. On the comparison of several mean values: an alternative approach. Biometrika 1951;38(3–4):330–6.

24. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/.

25. Kroeger CM, Ejima K. Data and analysis code for manuscript: persistent confusion in nutrition and obesity research about the validity of classic nonparametric tests in the presence of heteroscedasticity: evidence of the problem and valid alternatives. Version v2.2. Zenodo 2019. doi: 10.5281/zenodo.3625431.

26. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Clinical Trials 2005;2(4):282–90; discussion 301–4, 364–78.

27. du Pree I-B, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107(19):343–8.

28. Nayak BK, Hazra A. How to choose the right statistical test? Indian J Ophthalmol 2011;59(2):85–6.

29. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge, UK: Cambridge University Press; 1997.

30. Bryk AS, Raudenbush SW. Heterogeneity of variance in experimental studies: a challenge to conventional interpretations. Psychol Bull 1988;104(3):396–404.

31. Jones LV, Tukey JW. A sensible formulation of the significance test. Psychol Methods 2000;5(4):411–14.

32. Tukey JW. The philosophy of multiple comparisons. Statist Sci 1991;6(1):100–16.

33. Wilcox RR. Understanding and applying basic statistical methods using R. Hoboken, NJ: John Wiley & Sons; 2016.

34. Boos DD, Brownie C. Bootstrap methods for testing homogeneity of variances. Technometrics 1989;31(1):69–82.

35. Miller RG. The jackknife – a review. Biometrika 1974;61(1):1–15.