

Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited?

T. Mark Beasley · Stephen Erickson ·
David B. Allison

Received: 3 September 2008 / Accepted: 22 May 2009 / Published online: 14 June 2009
© Springer Science+Business Media, LLC 2009

Abstract Many complex traits studied in genetics have markedly non-normal distributions. This often implies that the assumption of normally distributed residuals has been violated. Recently, inverse normal transformations (INTs) have gained popularity among genetics researchers and are implemented as an option in several software packages. Despite this increasing use, we are unaware of extensive simulations or mathematical proofs showing that INTs have desirable statistical properties in the context of genetic studies. We show that INTs do not necessarily maintain proper Type 1 error control and can also reduce statistical power in some circumstances. Many alternatives to INTs exist. Therefore, we contend that there is a lack of justification for performing parametric statistical procedures on INTs with the exceptions of simple designs with moderate to large sample sizes, which makes permutation testing computationally infeasible and where maximum likelihood testing is used. Rigorous research evaluating the utility of INTs seems warranted.

Keywords Blom · Inverse normal transformation · Robustness · Type 1 error rate

Edited by Stacey Cherny.

T. M. Beasley (✉) · S. Erickson · D. B. Allison
Department of Biostatistics, Section on Statistical Genetics,
University of Alabama at Birmingham, Ryals Public Health
Building, Suite 327, Birmingham, AL 35294, USA
e-mail: MBeasley@UAB.edu

D. B. Allison
Department of Nutrition Sciences, University of Alabama
at Birmingham, Birmingham, AL, USA

D. B. Allison
Clinical Nutrition Research Center, University of Alabama
at Birmingham, Birmingham, AL, USA

Introduction

The validity of many statistical tests depends on the assumption that residuals from a fitted model are normally distributed (Berry 1993). In contrast, many complex traits studied in genetics have markedly non-normal distributions (Micceri 1989; Allison et al. 1999), which in many cases implies non-normal residuals. Several approaches exist to respond to non-normality, including but not limited to reliance on asymptotic properties (Mehta et al. 2004), transformation of the data (Etzel et al. 2003; George and Elston 1987; Shete et al. 2004; Yang et al. 2006), and the use of nonparametric tests (Neave and Wothington 1989), which subsumes the analysis of rank data (e.g., Zak et al. 2007), permutation tests, and bootstrap approaches as special cases (Good 1999).

Recently, the rank-based inverse normal transformation (INT) has gained in popularity among genetic researchers. INTs have been applied in a variety of genetic research designs, which have included a wide range of species. For example, INTs have been applied to model the heritability of change in startle response after training among mice (Valdar et al. 2006), microarray data in a rodent model (Przybyla-Zawislak et al. 2005), SNP associations in Rheumatoid Arthritis (Kraja et al. 2007), the family transmission and heritability of externalizing disorders in humans (Hicks et al. 2004), linkage to human body mass index (BMI) (Wu et al. 2002), genome wide association studies with BMI (Scuteri et al. 2007), twin studies of psychopathic personality traits (Blonigen et al. 2003) and frontal brain functioning in humans (Anokhin et al. 2003), genome-wide linkage scans of musical aptitude (Pulli et al. 2008) and attention deficit disorder (Nanda et al. 2008), genome wide association and SNP interactions in asthma-related traits (Dixon et al. 2007), QTL analysis and gene

interactions in the cold tolerance of sorghum (Knoll and Ejeta 2008), and in multiple other genetic studies (e.g., Ashton and Borecki 1987; Hicks et al. 2007; Martin and Crawford 1998; Silverman et al. 1990; Tzou et al. 1991). INTs are implemented as an option in at least three statistical genetic software packages (e.g., Analysis System 130 (2003), POLY (2003); see Chen and Abecasis 2006) and SOLAR (2008); see Almasy and Blangero (1998)). Despite this recent widespread use, we are unaware of thorough discussions of this issue. Therefore, we offer this commentary to consider arguments and evidence regarding the benefits and disadvantages of INTs in modern genetic research.

Distinguishing among INTs

INTs are ways of transforming the sample distribution of a continuous variable to make it appear more normally distributed. There are several types of INTs. The first distinction we make is between *rank-based* and *non-rank-based* INTs. Non-rank-based INTs entail assuming a particular cumulative distribution function (CDF) for the observed data, estimating the parameters of that distribution, converting observed scores to estimated quantiles from the CDF, and then converting these quantiles to standard normal deviates using the inverse normal (or probit function). Such non-rank-based INTs are usually referred to as copulas (Basrak et al. 2004; Li et al. 2006) and will not be considered further. It is worth noting, however, that the rank-based INTs can be expressed as a special case of the copula method in which the empirical CDF is used instead of restricting the CDF to some family of distributions. That is, every moment is in effect estimated from the data and the quantiles become simple functions of the ranks.

Rank-based INTs involve a preliminary step of converting a variable to ranks and can be further subdivided into two classes: those that involve a stochastic element and those that are deterministic. We are aware of only one INT that involves a stochastic element and this approach has been referred to as the use of “random normal deviates” (Conover 1980). One deterrent to this approach is that each investigator applying the same method to the same dataset will obtain a slightly different answer, which might be unsatisfying to some. This approach has the theoretical advantage of avoiding the granularity in the distribution of *P* values, an issue which often plagues many nonparametric tests. Nevertheless, the stochastic nature of these INTs seems to discourage researchers and they are rarely, if ever, used.

Deterministic rank-based INTs can be classified into those that use expected normal scores (Fisher and Yates 1938) versus those that use back transformation of sample

quantile (or fractional rank) to approximate the expected normal scores. Using numerical integration, Harter (1961) has provided the most complete table of expected normal scores. INTs that involve back transformation of fractional ranks to approximate the expected normal scores of Fisher and Yates (Maritz 1982) appear to be the most commonly used in genetic research and will be the primary focus of attention. In back-transforming ranks, a fractional offset is needed to avoid having the minimum and maximum observations transformed to negative and positive infinity, respectively. Perhaps the most commonly used rank-based INT transformation entails creating a modified rank variable and then computing a new transformed value of the phenotype for the *i*th subject:

$$Y_i^t = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right); \quad (1)$$

where r_i is the ordinary rank of the *i*th case among the *N* observations and Φ^{-1} denotes the standard normal quantile (or probit) function. Blom (1958) recommended the value of $c = 3/8$. Other such INTs are minor variations, with c replaced by other values. The Blom transformation is available as an automated option in software such as SAS and SPSS, which includes three other INTs: *Tukey* (1962; $c = 1/3$), *Rankit* (Bliss 1967; $c = 1/2$), and *van der Waerden* (1952; $c = 0$). Because the Blom transformation with $c = 3/8$ appears to be the most commonly used, we will focus on it, but the other INTs are virtually linear transformations of the Blom and are so similar to the expected normal scores that it is unlikely to make any difference which one is used (Tukey 1962). Thus, our comments apply to all deterministic rank-based INTs.

Issues concerning INTs

Will use of an INT guarantee that the sample model residuals are normally distributed?

As stated above, most parametric tests assume that residuals from a model are normally distributed. In situations where two or more groups (e.g., genotypic groups or identity by descent classes) are being compared, this implies that within-group phenotypic distributions (conditional on any covariates included in the model) are normally distributed. In contrast, use of an INT assures that the marginal distribution of the phenotype is nearly normal, which is only appropriate under a complete null hypothesis (i.e., all effects in the model are null). Thus, approximate normality is assured for the wrong distribution (i.e., for the overall phenotype, not necessarily the residuals). This is in contrast to some other transformations, such as the Hodges and Lehmann (1962) method, which “aligns” the data (i.e.,

removes the effects of other variables) before rank transformation, or the Box–Cox transformation (Box and Cox 1964), which can be implemented such that it maximizes the normality of the sample residuals.

This limitation of INTs was recognized by Servin and Stephens (2007) who wrote: “Regarding the normality assumption, following a suggestion by Mathew Barber (personal communication), in practical applications, we are currently applying a normal quantile transform to phenotypes (replacing the r th biggest of N observations with the $(r - 0.5)/N$ th quantile of the standard normal distribution) before applying our methods. Imposing normality on our phenotype in this way is different from the normality assumption in our phenotype model, which states that the residuals are normally distributed. However, in this context, where effect sizes are expected to be generally rather small, normality of phenotype and normality of residuals are somewhat similar assumptions, *suggesting* that this transform may be effective.” [Emphasis added.] This also exposes a problem of INTs; they do not make any population distribution normal, they merely make particular sample distributions appear near-normal.

Has theoretical work shown that INTs have desirable statistical properties, especially in the context of genetic studies?

We have conducted a thorough search of the literature and contacted colleagues in the field, particularly those that use INTs in genetic studies, and have been unable to identify such theoretical work in the context of genetic studies. Few of the genetic studies using an INT that we read cited any evidence that it had any particular statistical properties or benefits. One of the few examples, Hicks et al. (2004) wrote “...we used a normalizing (Blom) transformation that has been shown to optimize model selection when analyzing psychiatric symptom count data” and cited van den Oord et al. (2000) in support. Yet a careful reading of van den Oord et al.’s paper and personal communication with van den Oord (8/23/2007) reveal that van den Oord et al. (2000) actually examined the performance of a simple rank transformation rather than an INT in comparison to other procedures for optimizing model selection and parameter estimation.

In the classic non-parametric statistics literature, however, there are many writings on the use of INTs and a few key points can be gleaned. First, it has been shown that among all possible rank-based INTs, those based on the Fisher and Yates expected normal scores are the most powerful (Barnard 1957; Bradley 1968). In practice, however, all deterministic rank-based INTs are so similar that it is unlikely to make any difference which one is used (Tukey 1962). Second, the original presentations of

tests based on INTs stress that the inference should be conducted via permutation if the test is to be exact (Sprent and Smeeton 2001), although in practice this is often eschewed in favor of parametric tests applied to INT scores.

Lastly, compared to the most powerful parametric tests, tests of group differences in location (e.g., mean) involving INTs have asymptotic relative efficiency (ARE) ≥ 1.0 (Bradley 1968; Chernoff and Savage 1958; Conover 1980). Unfortunately, ARE does not necessarily imply relative efficiency in any particular situation leaving open the question as to how well INTs will perform in specific situations. Moreover, while parametric tests with deterministic INTs have AREs ≥ 1.0 , so too do permutation tests with the untransformed scores (Stuart 1954).

Have extensive simulations shown that INTs have certain desirable properties in the context of genetic studies?

We have been able to identify only a few simulation studies in the genetics context that addresses the utility of rank-based INTs. Specifically, Wang and Huang (2002) developed a likelihood-based methodology in which a rank-based INT increased power. Their score-statistic approach to QTL mapping is asymptotically equivalent to the corresponding likelihood-ratio test and assumes normally distributed errors. In a series of simulations, the authors showed that with skewed ($\chi^2_{(1)}$) error distributions, the rank-based INT approach leads to noticeably greater power; however, Type 1 error control was inconsistent, especially with non-additive genetic models.

Peng et al. (2007) conducted a series of simulations of right-skewed (i.e., non-normal) phenotypes within pedigrees, performed Amos’ (1994) maximum likelihood (ML) variance components (VC) analysis on both untransformed and INT scores, specifically van der Waerden normal scores (Eq. 1 with $c = 0$), and compared them to a semiparametric QTL (SQTL) method (Diao and Lin 2005). VC analyses applied to INT scores maintained valid Type 1 error rates, while Type 1 error rates from the untransformed data resulted in inflated Type 1 errors. The SQTL method, on the other hand, appeared overly conservative at $\alpha = 0.05$ and less conservative at $\alpha = 0.001$ when compared with the INT-transformed results. In preliminary simulations, we found similar results in the context of testing VC in a twin study with an ACE model (i.e., model with additive genetic, common environmental, and unique environmental effects). Kraja et al. (2007) also found that the INT performed well in a large sample ML-based SNP association testing context. Therefore, the INT appears to perform nearly as well as if the true transformation to normality were known and applied, but it remains arguable whether the INT method

applied to ML estimation of VC is uniformly preferable to the SQTl method.

Our review of the general literature on non-parametric statistics shows that use of INT scores in parametric tests sometimes yields superior performance compared to parametric tests with untransformed data or other non-parametric alternatives when the normality assumption is not met, but in other circumstances INTs can yield poor performance (e.g., Pratt 1964; Keselman et al. 1977), especially when compared to other non-parametric alternatives (e.g., Knoke 1991). Because relative performance will vary by condition studied, it is important to study INTs in conditions that are similar to those used in modern complex trait genetic studies. Unfortunately, the specific circumstances under which INTs are currently being used often differ substantially from the conditions under which INTs and other rank-based approaches were studied in the past (in terms of both the tests that were studied and the α levels used) making the empirical work (e.g., simulations) of bygone days of lesser utility and suggesting the need to continue empirical evaluations in our present context.

Does use of INTs guarantee that tests assuming normality will have correct Type 1 error rates?

Because the marginal distribution of the sample phenotype is almost certain to be nearly normally distributed after an INT, it is tempting to conclude that this will guarantee a correct Type 1 error rate when the normality of residuals assumptions is violated. Indeed, in a recent *PLoS Genetics* article, Scuteri et al. wrote “To ensure adequate control of Type 1 error rates, we applied an inverse normal transformation to each trait prior to analysis.” [Emphasis added.] Yet, INTs do not necessarily fulfill the fundamental assumption of normally distributed residuals. It therefore seems a *fait accompli* that the use of INTs cannot guarantee that tests assuming normality will have correct Type 1 error rates on strictly theoretical grounds. Nevertheless, it might be conjectured that, in practice, INTs do offer approximately correct Type 1 error rates.

To explore this issue further, we conducted a number of simulations to illustrate key points. First, we investigate testing of the equality of two means for a continuous outcome (Y) based on two independent random samples of $n_1 = n_2 = 5$ subjects. These groups could be, for example, treated and untreated subjects in a microarray study or wild-type and knock-out mice in a gene manipulation study, where such small sample sizes are quite common. We began by simulating the situation in which all assumptions of the t -test were met and the null hypothesis was true (i.e., there was no mean group difference in the two populations sampled). As can be seen in the first two rows of Table 1, when the normality assumption is met and

no transformation (Y) is used, the correct Type 1 error rates are obtained as expected. In contrast, performing the parametric t -test on the INT (Blom) scores produced significant inflation of the Type 1 error rate (i.e., too many false positives) at the $\alpha = 0.05$ and 0.01 levels, whereas the Type 1 error rate at a nominal $\alpha = 0.001$ level was zero. These discrepancies from the expected Type 1 error rates are due to the fact that the t -test performed on INT scores produces P values that are smaller than the P values yielded by the permutation distribution of ranks. For $\alpha = 0.05$, the problem of granularity of the permutation distribution of ranks does not subside until the samples are greater than 10 cases per group (Bradley 1968).

The previous simulations clearly showed that INTs do not ensure adequate control of the Type 1 error rate in all situations. In response, one might argue that one would not use an INT when the normality assumption is met, so the aforementioned findings are irrelevant. However, one never knows unequivocally whether the normality assumption is met and assessing the fit of the sample data to the normality assumption is not very useful in small samples (Farrell and Rogers-Stewart 2006). Nevertheless, it could be conjectured that INTs will help control Type 1 error rates when the normality assumption is violated. Therefore, we followed the approach of Wang and Huang (2002) and simulated data from three distributions that were markedly non-normal. We simulated a one degree-of-freedom (df) chi-square distribution ($\chi^2_{(1)}$), which has a skewness of 2.83 and kurtosis of 12, and another highly-skewed heavy-tailed distribution, Weibull ($\lambda = 1$; $k = 0.5$), which has a skewness of 6.39 and kurtosis of 76. We also simulated a heavy-tailed (kurtosis of 3), but symmetric LaPlace distribution. Table 1 shows that the t -test does not make excess Type 1 errors with departures from the normality assumption (under the conditions we have simulated), but rather tends to be conservative. The INT approach as well as permutation tests performed on Y (Good 1999) and the ranks (Wilcoxon 1945; Mann and Whitney 1947) returns virtually the same Type 1 error rates as when the normality assumption was met. Again, the reason for these virtually identical results with the INT across distributions is due to the permutation distribution of ranks. If a permutation test were performed on the INT scores, it would have Type 1 error rates identical to the permutation test performed on ranks. Thus, even when the data are markedly non-normal, performing a parametric test on the INTs does not ensure correct Type 1 error rates, and thus provides no definitive advantage over the use of untransformed data or ranks in small samples.

Finally, one might argue that while the conclusions above may be relevant to small studies such as microarray or rodent knockout studies, however in larger studies, an INT will perform well. To address this question, we ran

Table 1 Empirical Type 1 error rates for tests of equality of two means

Test	Error distribution	<i>N</i> per group	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
<i>t</i> -test (<i>Y</i>)	Normal	5	0.05061	0.01013	0.00105
<i>t</i> -test (Blom)	Normal	5	0.05540 [‡]	0.01653 [‡]	0
Permutation test (<i>Y</i>)	Normal	5	0.04785	0.00807	0
Permutation test (Ranks)	Normal	5	0.03204 [*]	0.00807 [#]	0
<i>t</i> -test (<i>Y</i>)	LaPlace	5	0.04229	0.00719	0.00052
<i>t</i> -test (Blom)	LaPlace	5	0.05555 [‡]	0.01649 [‡]	0
Permutation test (<i>Y</i>)	LaPlace	5	0.04702	0.00809	0
Permutation test (Ranks)	LaPlace	5	0.03221 [*]	0.00809 [#]	0
<i>t</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	5	0.03102	0.00456	0.00051
<i>t</i> -test (Blom)	$\chi^2_{(1)}$	5	0.05548 [‡]	0.01582 [‡]	0
Permutation test (<i>Y</i>)	$\chi^2_{(1)}$	5	0.04755	0.00820	0
Permutation test (Ranks)	$\chi^2_{(1)}$	5	0.03222 [*]	0.00820 [#]	0
<i>t</i> -test (<i>Y</i>)	Weibull (1, 0.5)	5	0.02081	0.00217	0.00034
<i>t</i> -test (Blom)	Weibull (1, 0.5)	5	0.05473 [‡]	0.01540 [‡]	0
Permutation test (<i>Y</i>)	Weibull (1, 0.5)	5	0.04744	0.00770	0
Permutation test (Ranks)	Weibull (1, 0.5)	5	0.03108 [*]	0.00770 [#]	0
<i>t</i> -test (<i>Y</i>)	LaPlace	25	0.04808	0.00818	0.00054
<i>t</i> -test (Blom)	LaPlace	25	0.04882	0.00962	0.00094
Kruskal–Wallis (Ranks)	LaPlace	25	0.04884	0.00836	0.00056
<i>t</i> -test (<i>Y</i>)	LaPlace	50	0.04926	0.00948	0.00078
<i>t</i> -test (Blom)	LaPlace	50	0.04878	0.01048	0.00094
Kruskal–Wallis (Ranks)	LaPlace	50	0.04844	0.00984	0.00082
<i>t</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	25	0.04356	0.00632	0.00025
<i>t</i> -test (Blom)	$\chi^2_{(1)}$	25	0.04924	0.00996	0.00120
Kruskal–Wallis (Ranks)	$\chi^2_{(1)}$	25	0.04828	0.00880	0.00056
<i>t</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	50	0.04666	0.00822	0.00072
<i>t</i> -test (Blom)	$\chi^2_{(1)}$	50	0.04950	0.00976	0.00124
Kruskal–Wallis (Ranks)	$\chi^2_{(1)}$	50	0.04884	0.00928	0.00090
<i>t</i> -test (<i>Y</i>)	Weibull (1, 0.5)	25	0.03442	0.00304	0.00008
<i>t</i> -test (Blom)	Weibull (1, 0.5)	25	0.04982	0.00982	0.00102
Kruskal–Wallis (Ranks)	Weibull (1, 0.5)	25	0.04874	0.00908	0.00068
<i>t</i> -test (<i>Y</i>)	Weibull (1, 0.5)	50	0.04236	0.00542	0.00028
<i>t</i> -test (Blom)	Weibull (1, 0.5)	50	0.05094	0.01026	0.00116
Kruskal–Wallis (Ranks)	Weibull (1, 0.5)	50	0.05044	0.01022	0.00092

Note: 100,000 simulations

* Expected value of 0.031746

[‡] Indicates significantly greater than nominal α level at the 95% confidence level

[#] Expected value of 0.007937

simulations with larger (though still modest) sample sizes of $N = 50$ and 100 (25 and 50 subjects in each of two groups). With larger sample sizes, however, permutation testing becomes more difficult and one may assume that the *t*-test for the untransformed scores (*Y*) and approximate tests for ranks will yield valid Type 1 error rates. Table 1 shows that even with these larger sample sizes, the parametric *t*-test still has a suppressed Type 1 error rate, for

asymmetric error distributions. The Type 1 error inflation for the INT scores due to the granular nature of the permutation distribution is eliminated. Therefore, the *t*-test performed on INT scores has approximately the correct α level, but so too does the rank-based Kruskal and Wallis (1952) approximate test. Thus, while an INT may hold the Type 1 error rate under the nominal alpha asymptotically, so too will the use of untransformed and rank data. Thus,

there is no clear advantage to using the INT in terms of controlling the Type 1 error rate of OLS tests in one-way ANOVA-type designs.

To further examine this issue, we generalized our simulation to the one-way ANOVA with three groups. This analysis is extremely common in genetic association studies with a continuous phenotype (Y) as the dependent variable and genotype as the grouping variable. One notable issue with applying the one-way ANOVA to a genetic association studies is that the sample sizes per genotype group are never expected to be equal in the population. So if one randomly samples from the population, the genotype groups would not be of equal size. For example, with a di-allelic marker with a minor allele frequency of $P = 0.5$, the heterozygote genotype will have twice the sample size of the other two homozygote genotypes. This imbalance in the samples sizes worsens as the di-allelic frequencies deviate from 0.5. Unequal sample sizes in the one-way ANOVA are of particular importance because of the Behrens–Fisher problem. Specifically, the standard one-way ANOVA assumes that each group has the same population variance. The standard ANOVA F -test is robust to heterogeneous variances, as long as the sample sizes are equal (e.g., Kohr and Games 1974; Zimmerman 2004). It is well-known that the Type 1 error rate of the ANOVA F -test is spuriously inflated or suppressed by unequal variances combined with unequal sample sizes. The Welch (1947) separate-variances version of the F -test, which does not use a pooled variance estimate and has modified df s, usually eliminates these effects. In fact, Zimmerman (2004) concluded that optimum protection of the Type 1 error rate is assured by using the Welch test unconditionally whenever sample sizes are unequal. Although the exact distribution of the Welch statistic is known under normality (Ray and Pitman 1961), it remains an approximate solution to the Behrens–Fisher problem. Monte Carlo studies have shown that the Welch approximate solution is not robust to departures from normality (e.g., James 1959; Yuen 1974). Furthermore, solutions based on nonparametric or nonparametric-like procedures have been unsuccessful. For example, Pratt (1964) showed that the Mann–Whitney U and the expected normal scores test (Hájek and Sidák 1967) resulted in nonrobust Type I error rates. Feir-Walsh and Toothaker (1974) and Keselman, et al. (1977) found the Kruskal–Wallis test (Kruskal and Wallis 1952) and expected normal scores test (McSweeney and Penfield 1969) to be substantially affected by heterogeneity of variance.

To demonstrate these issues in the context of genetic association studies, suppose a continuous phenotype (Y) that under the null hypothesis of no mean differences among the three genotypes is either sampled from either a normal error distribution or from one of the three markedly non-normal error distributions: LaPlace; $\chi^2_{(1)}$; or Weibull. There is one marker of interest, G . For this simulation, allele frequencies

for G were set at $P = 0.5$ and 0.25 . Thus, for these two scenarios, the expected sample sizes for each genotype will differ substantially. For $P = 0.5$, the homozygote (gg and GG) groups would be expected to have 25% of the total sample size (N), while the heterozygote (Gg) group would be twice as large with 50% of N . For $P = 0.25$, the homozygote (gg) group would be expected to have 56.25% of N , the heterozygote (Gg) group would be expected to have 37.5% of N , and the homozygote (GG) group would be expected to have 6.25% of N . We investigated a fairly large total sample sizes of $N = 400$ and set the sample sizes for each genotype at the expected value (e.g., with $P = 0.25$, the sample size for the GG genotype was set at 25). In keeping with other statistical research on the Behrens–Fisher problem, we set two basic patterns of variance. One pattern is referred to as a positive pairing where the larger groups have a larger variance. In the case of $P = 0.5$, the smaller homozygote groups will have a smaller variance, $\sigma_{gg}^2 = \sigma_{GG}^2 = 1.0$ and the larger heterozygote group will have a larger variance, $\sigma_{Gg}^2 = 2.0$. With $P = 0.25$, the largest homozygote (gg) group will have the largest variance, $\sigma_{gg}^2 = 2.0$, and heterozygote group will have a variance of $\sigma_{Gg}^2 = 1.5$, and the smallest homozygote (GG) group will have the smallest variance, $\sigma_{GG}^2 = 1$. A second pattern is referred to as a negative pairing where the larger groups have a smaller variance. In the case of $P = 0.5$, the smaller homozygote groups will have a larger variance, $\sigma_{gg}^2 = \sigma_{GG}^2 = 2.0$ and the larger heterozygote group will have a smaller variance, $\sigma_{Gg}^2 = 1$. With $P = 0.25$, the largest homozygote (gg) group will have the smallest variance, $\sigma_{gg}^2 = 1$, and heterozygote group will have a variance, $\sigma_{Gg}^2 = 1.5$, and the smallest homozygote (GG) group will have the largest variance, $\sigma_{GG}^2 = 2.0$.

When there a positive pairing of sample sizes and variances (lower half of Tables 2, 3), the Type 1 error rate of the ANOVA F -test applied is generally suppressed if the error distributions are Normal or at least symmetric (LaPlace). Also, when there a negative pairing of sample sizes and variances (upper half of Tables 2, 3) and the error distributions are symmetric (Normal and LaPlace), the Type 1 error rate of the ANOVA F -test applied is generally inflated. In these situations, the Welch test applied to either the untransformed scores (Y) or to the INT scores (Blom) helps correct the Type 1 error rate. Therefore, in these circumstances, the INT may lead to a valid Type 1 error rate, but it does not demonstrate a clear advantage over using the untransformed data. Consistent with other studies (e.g., James, 1959; Yuen, 1974), if the error distribution are skewed the Welch test performed on the untransformed scores may not lead to a correct Type 1 error rate regardless of the sample size and variance pattern. What is most alarming, however, is that with skewed error distributions, the INT severely worsens the Type 1 error rate inflation of the ANOVA F -test and the Welch test. As can be seen in

Table 2 Empirical Type 1 error rates for tests of equality of three means with total sample size of $N = 400$, allele frequency of $P = 0.50$ ($n_{GG} = 100$, $n_{Gg} = 200$, $n_{gg} = 100$), negative (upper-half), and positive (lower-half) pairing of sample sizes and variances

Test	Error distribution	(GG Gg gg) σ^2 per group	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
<i>F</i> -test (<i>Y</i>)	Normal	2.0 1.0 2.0	0.07802 [¥]	0.01982 [¥]	0.00286 [¥]
<i>F</i> -test (Blom)	Normal		0.07189 [¥]	0.01763 [¥]	0.00249 [¥]
Welch test (<i>Y</i>)	Normal		0.05021	0.00982	0.00093
Welch test (Blom)	Normal		0.04709	0.00896	0.00083
<i>F</i> -test (<i>Y</i>)	LaPlace	2.0 1.0 2.0	0.07735 [¥]	0.01970 [¥]	0.00272 [¥]
<i>F</i> -test (Blom)	LaPlace		0.06774 [¥]	0.01587 [¥]	0.00196 [¥]
Welch test (<i>Y</i>)	LaPlace		0.04897	0.00951	0.00090
Welch test (Blom)	LaPlace		0.04918	0.00947	0.00087
<i>F</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	2.0 1.0 2.0	0.07641 [¥]	0.01890 [¥]	0.00251 [¥]
<i>F</i> -test (Blom)	$\chi^2_{(1)}$		0.97045 [¥]	0.89414 [¥]	0.70173 [¥]
Welch test (<i>Y</i>)	$\chi^2_{(1)}$		0.06122 [¥]	0.01718 [¥]	0.00358 [¥]
Welch test (Blom)	$\chi^2_{(1)}$		0.96537 [¥]	0.87953 [¥]	0.66590 [¥]
<i>F</i> -test (<i>Y</i>)	Weibull (1, 0.5)	2.0 1.0 2.0	0.07067 [¥]	0.01461 [¥]	0.00169 [¥]
<i>F</i> -test (Blom)	Weibull (1, 0.5)		0.99981 [¥]	0.99869 [¥]	0.98711 [¥]
Welch test (<i>Y</i>)	Weibull (1, 0.5)		0.08008 [¥]	0.02698 [¥]	0.00684 [¥]
Welch test (Blom)	Weibull (1, 0.5)		0.99976 [¥]	0.99820 [¥]	0.98406 [¥]
<i>F</i> -test (<i>Y</i>)	Normal	1.0 2.0 1.0	0.02975	0.00462	0.00042
<i>F</i> -test (Blom)	Normal		0.02721	0.00395	0.00028
Welch test (<i>Y</i>)	Normal		0.05043	0.01000	0.00096
Welch test (Blom)	Normal		0.04655	0.00881	0.00082
<i>F</i> -test (<i>Y</i>)	LaPlace	1.0 2.0 1.0	0.02943	0.00475	0.00030
<i>F</i> -test (Blom)	LaPlace		0.03297	0.00524	0.00038
Welch test (<i>Y</i>)	LaPlace		0.04889	0.00969	0.00086
Welch test (Blom)	LaPlace		0.04815	0.00936	0.00084
<i>F</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	1.0 2.0 1.0	0.03114	0.00568	0.00067
<i>F</i> -test (Blom)	$\chi^2_{(1)}$		0.96098 [¥]	0.87231 [¥]	0.65804 [¥]
Welch test (<i>Y</i>)	$\chi^2_{(1)}$		0.05264	0.01136	0.00145
Welch test (Blom)	$\chi^2_{(1)}$		0.96551 [¥]	0.88077 [¥]	0.66728 [¥]
<i>F</i> -test (<i>Y</i>)	Weibull (1, 0.5)	1.0 2.0 1.0	0.03460	0.00808	0.00100
<i>F</i> -test (Blom)	Weibull (1, 0.5)		0.99982 [¥]	0.99750 [¥]	0.98304 [¥]
Welch test (<i>Y</i>)	Weibull (1, 0.5)		0.05359	0.01076	0.00091
Welch test (Blom)	Weibull (1, 0.5)		0.99985 [¥]	0.99775 [¥]	0.98388 [¥]

Note: 100,000 simulations

[¥] Indicates significantly greater than nominal α level at the 95% confidence level

Tables 2 and 3, the Welch test applied to the INT (Blom) has rejection rates approaching 100% when the error distributions are skewed ($\chi^2_{(1)}$; Weibull), regardless of the sample size and variance pattern. This is to be expected to some extent because the INT is based on ranks and rank-based tests are known to be sensitive to heterogeneity of variance (e.g., Keselman, et al. 1977; Zimmerman, 1996; Zumbo and Coulombe 1997). As compared to other simulation studies on the Behrens–Fisher problem, we used a very realistic disparity among the sample sizes based expected frequencies of genotypes and we did not use

extreme disparities in the variances. Other studies have used variance ratios upto tenfold. In our simulations, the smallest variance was 1.0 and the largest was 2.0. In this situation, applied researchers, even if they is familiar with the Behrens–Fisher problem, might not be alarmed if the within-genotype distributions are skewed and one genotype has a standard deviation of 1 and another genotype has a standard deviation of 1.4. However, we show that in this case the Welch test may not maintain a valid Type 1 error rate, and furthermore, apply the Welch test to INT scores severely worsens the problem.

Table 3 Empirical Type 1 error rates for tests of equality of three means with total sample size of $n = 400$, allele frequency of $P = 0.25$ ($n_{GG} = 25$, $n_{Gg} = 150$, $n_{gg} = 225$), negative (upper-half), and positive (lower-half) pairing of sample sizes and variances

Test	Error distribution	(GG Gg gg) σ^2 per group	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
<i>F</i> -test (<i>Y</i>)	Normal	2.0 1.5 1.0	0.10030 [¥]	0.03034 [¥]	0.00533 [¥]
<i>F</i> -test (Blom)	Normal		0.09252 [¥]	0.02647 [¥]	0.00421 [¥]
Welch test (<i>Y</i>)	Normal		0.05001	0.01039	0.00116
Welch test (Blom)	Normal		0.04878	0.01048	0.00125
<i>F</i> -test (<i>Y</i>)	LaPlace	2.0 1.5 1.0	0.09985 [¥]	0.03036 [¥]	0.00556 [¥]
<i>F</i> -test (Blom)	LaPlace		0.08025 [¥]	0.02078 [¥]	0.00299 [¥]
Welch test (<i>Y</i>)	LaPlace		0.04827	0.00907	0.00077
Welch test (Blom)	LaPlace		0.04954	0.01068	0.00131
<i>F</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	2.0 1.5 1.0	0.09544 [¥]	0.02847 [¥]	0.00620 [¥]
<i>F</i> -test (Blom)	$\chi^2_{(1)}$		0.94547 [¥]	0.85131 [¥]	0.65849 [¥]
Welch test (<i>Y</i>)	$\chi^2_{(1)}$		0.08964 [¥]	0.03948 [¥]	0.01687 [¥]
Welch test (Blom)	$\chi^2_{(1)}$		0.88016 [¥]	0.68898 [¥]	0.37710 [¥]
<i>F</i> -test (<i>Y</i>)	Weibull (1, 0.5)	2.0 1.5 1.0	0.08504 [¥]	0.02861 [¥]	0.00867
<i>F</i> -test (Blom)	Weibull (1, 0.5)		0.99859 [¥]	0.99177 [¥]	0.96038
Welch test (<i>Y</i>)	Weibull (1, 0.5)		0.14157 [¥]	0.08047 [¥]	0.04306
Welch test (Blom)	Weibull (1, 0.5)		0.99407 [¥]	0.96715	0.84900
<i>F</i> -test (<i>Y</i>)	Normal	1.0 1.5 2.0	0.02446	0.00378 [¥]	0.00034 [¥]
<i>F</i> -test (Blom)	Normal		0.02388	0.00353 [¥]	0.00037 [¥]
Welch test (<i>Y</i>)	Normal		0.05037	0.00986 [¥]	0.00121 [¥]
Welch test (Blom)	Normal		0.04831	0.00935 [¥]	0.00108 [¥]
<i>F</i> -test (<i>Y</i>)	LaPlace	1.0 1.5 2.0	0.02492	0.00380	0.00034
<i>F</i> -test (Blom)	LaPlace		0.02883	0.00451	0.00042
Welch test (<i>Y</i>)	LaPlace		0.04814	0.00848	0.00084
Welch test (Blom)	LaPlace		0.04910	0.00943	0.00106
<i>F</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	1.0 1.5 2.0	0.02611	0.00508	0.00048
<i>F</i> -test (Blom)	$\chi^2_{(1)}$		0.80479 [¥]	0.57026 [¥]	0.27504 [¥]
Welch test (<i>Y</i>)	$\chi^2_{(1)}$		0.07594 [¥]	0.02862 [¥]	0.00938 [¥]
Welch test (Blom)	$\chi^2_{(1)}$		0.90466 [¥]	0.72989 [¥]	0.41380 [¥]
<i>F</i> -test (<i>Y</i>)	Weibull (1, 0.5)	1.0 1.5 2.0	0.03269	0.00801	0.00135
<i>F</i> -test (Blom)	Weibull (1, 0.5)		0.98888 [¥]	0.94116 [¥]	0.77915 [¥]
Welch test (<i>Y</i>)	Weibull (1, 0.5)		0.10358 [¥]	0.04449 [¥]	0.01593 [¥]
Welch test (Blom)	Weibull (1, 0.5)		0.99897 [¥]	0.99093 [¥]	0.93403 [¥]

Note: 100,000 simulations

[¥] Indicates significantly greater than nominal α level at the 95% confidence level

Statistical power

Wilcox (1995) persuasively argued that in many cases the compelling rationale for the use of non-parametric tests over parametric tests is not necessarily preservation of Type 1 error rate, but rather enhancement of power. To address this issue, we simulated data under an alternative hypothesis. Specifically, the variance of *Y* was 1.0 for each of the two groups and between-group mean difference of 0.5 and 0.125 was added. Table 4 shows that for smaller samples sizes the non-parametric approach of performing permutation tests on the untransformed scores (*Y*) and on the ranks does indeed yield more power than a parametric

approach, at least for the two larger α levels. At $\alpha = 0.05$, the permutation test performed on the untransformed scores has the most power. At $\alpha = 0.01$, permutation tests performed on *Y* and the ranks have identical power, which is due to the fact that in the tails of these permutation distributions are identical. At $\alpha = 0.001$, both permutation tests have zero power because they both have a minimum *P* value of 0.0079. With smaller sample sizes, the *t*-test performed on INT (Blom) scores failed to hold the Type 1 error rate (Table 1) and thus yields spuriously higher power; however, if a permutation test were performed on the INT scores, it would have power identical to the permutation test performed on ranks with $N = 5$ per group.

Table 4 Empirical power for tests of equality of two means

Test	Error distribution	<i>n</i> per group	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
<i>t</i> -test (<i>Y</i>)	Normal	5	0.10939	0.02733	0.00322
<i>t</i> -test (Blom)	Normal	5	0.11653 [‡]	0.04071 [‡]	0
Permutation test (<i>Y</i>)	Normal	5	0.10446	0.02162	0
Permutation test (Ranks)	Normal	5	0.03204	0.02162	0
<i>t</i> -test (<i>Y</i>)	LaPlace	5	0.12052	0.02851	0.00314
<i>t</i> -test (Blom)	LaPlace	5	0.14028 [‡]	0.05539 [‡]	0
Permutation test (<i>Y</i>)	LaPlace	5	0.12864	0.03139	0
Permutation test (Ranks)	LaPlace	5	0.09391	0.03139	0
<i>t</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	5	0.18244	0.05897	0.01084
<i>t</i> -test (Blom)	$\chi^2_{(1)}$	5	0.26666 [‡]	0.13369 [‡]	0
Permutation test (<i>Y</i>)	$\chi^2_{(1)}$	5	0.23624	0.09799	0
Permutation test (Ranks)	$\chi^2_{(1)}$	5	0.19199	0.09799	0
<i>t</i> -test (<i>Y</i>)	Weibull (1, 0.5)	5	0.34973	0.18931	0.07927
<i>t</i> -test (Blom)	Weibull (1, 0.5)	5	0.43828 [‡]	0.31499 [‡]	0
Permutation test (<i>Y</i>)	Weibull (1, 0.5)	5	0.44622	0.29085	0
Permutation test (Ranks)	Weibull (1, 0.5)	5	0.36267	0.29085	0
<i>t</i> -test (<i>Y</i>)	LaPlace	25	0.42706	0.20808	0.05924
<i>t</i> -test (Blom)	LaPlace	25	0.47360	0.24784	0.07972
Kruskal–Wallis (Ranks)	LaPlace	25	0.50998	0.26034	0.07448
<i>t</i> -test (<i>Y</i>)	$\chi^2_{(1)}$	25	0.46456	0.25234	0.08722
<i>t</i> -test (Blom)	$\chi^2_{(1)}$	25	0.87178	0.68580	0.38978
Kruskal–Wallis (Ranks)	$\chi^2_{(1)}$	25	0.86978	0.68242	0.37654
<i>t</i> -test (<i>Y</i>)	Weibull (1, 0.5)	25	0.58656	0.39310	0.20444
<i>t</i> -test (Blom)	Weibull (1, 0.5)	25	0.98630	0.93680	0.77968
Kruskal–Wallis (Ranks)	Weibull (1, 0.5)	25	0.99022	0.95642	0.82032
<i>t</i> -test (<i>Y</i>) [*]	Weibull (1, 0.5)	*25	0.08972	0.02034	0.00206
<i>t</i> -test (Blom) [*]	Weibull (1, 0.5)	*25	0.62910	0.37776	0.14214
Kruskal–Wallis (Ranks) [*]	Weibull (1, 0.5)	*25	0.60964	0.35880	0.12734

Note: 100,000 simulations

Population mean difference = 0.5 within group standard deviations, except * where Population mean difference = 0.125 within group standard deviations (last three rows). Kruskal–Wallis is the large-sample approximate test of the Wilcoxon–Mann–Whitney

[‡] Demonstrated significantly inflated Type 1 error rate (see Table 1)

With larger sample sizes, the Type 1 error rates for *t*-tests performed on *Y* and the Blom scores and the Kruskal–Wallis approximate rank test were not inflated, and thus, it is valid to compare their power rates, even though the parametric test with a suppressed false positive rate is at a disadvantage (Bradley 1978). With a larger sample size of 25 per group, the *t*-test performed on INT scores and the rank-based Kruskal–Wallis approximate test demonstrated more statistical power than the parametric test (see Table 4). With a larger effect size of 0.5 standard deviation units, the rank-based test had slightly more power than the INT approach. For the extremely non-normal Weibull distribution, we reduced the effect size to 0.125 and found that the INT approach had a slight power advantage over the rank-based procedure. Whether this constitutes a consistent power advantage is debatable.

Another concern that occurs frequently is an outcome measure (or phenotype) for which many subject have the same score. For example, when blood is drawn, it is very common for many subjects to have zero level of some measures. This creates a problem because it violates the normality assumption of the *t*-test. For an INT this would result in many subjects having the same transformed score and would not remedy the issue of ties. In terms of rank-based tests, the permutation distribution of ranks is affected by ties. Thus, for smaller sample sizes we suggest using a permutation test or an exact test for ranks; performing the parametric *t*-test on INT scores does not seem appropriate. For larger samples, the Kruskal–Wallis test has a well-known correction for ties. The *t*-test is likely to be robust but have low power.

In a small scale simulation, we generated normally distributed data (*Y*) with a mean of 0 and variance of 1 for

both groups and set all values below zero equal to zero, thus creating a distribution for Y that has 50% of the values equaling zero and the rest of values are positive (i.e., a positively skewed distribution with many ties). We generated data for two samples sizes that could make permutation tests very time consuming (two groups with equal sample sizes on $n_1 = n_2 = 12$ and 20), although one may not be convinced that the asymptotic properties of the t -test and the WMW test apply at these sample sizes. We found that the t -test, WMW and the t -test applied to the INT scores maintained a valid Type 1 error rate at $\alpha = 0.05$, 0.01 and 0.001. The WMW and the t -test applied to INT scores had very similar statistical power, whereas, the t -test performed on the original data had substantially less power (results not tabled).

The intricacies of the differences among the power functions of the t -test, WMW, and the t -test performed on INTs with different sample sizes, effect sizes, and error distributions need further investigation. However, this does suggest some circumstances in which performing the t -test on INTs could be useful (i.e., analysis of extremely non-normal data from simple research designs with sample sizes large enough to make permutation testing intractable, especially when smaller effect sizes are suspected).

How do rank-based INTs fare in complex models?

One might also wonder how the INT approach will perform in the more complex and sophisticated tests that geneticists often employ. The use of rank transformed data in more complicated models has many statistical issues. For example, when other blocking factors are present, Hodges and Lehmann (1962) illustrated the necessity of “aligning” the data before rank transformation. This is because in a factorial design the expected value of ranks for an observation in one cell has a non-linear dependence on the original means of the other cells (Headrick and Sawilowsky 2000; Thompson 1991). Consequently, interaction and main effect relationships are not maintained after rank transformations are performed (Blair et al. 1987). Parametric tests for interaction applied to ranks lack an invariance property, which produces distorted Type 1 and Type 2 error rates and have performed poorly compared with their normal theory counterparts (e.g., Salter and Fawcett 1993; Mansouri and Chang 1995; Toothaker and Newman 1994). Interaction tests for the rank transform (Conover and Iman 1981) have also performed poorly for a variety of other designs (Akritas 1990; Thompson 1993) including polynomial and response surface regression (Headrick and Rotou 2001), analysis of covariance (Headrick and Sawilowsky 2000; Headrick and Vineyard 2001) and repeated measures designs (Beasley 2002). These findings support Hora and Conover’s (1984) warning

that simply ranking the data does not result in an adequate test for non-additivity (i.e., interaction), which has implications for using rank-based INTs when evaluating epistasis and gene \times environment interactions.

In a genome wide association study of asthma-related phenotypes, Dixon et al. (2007) applied an INT to each trait and subsequently examined SNP interactions, some of which were on the same chromosome, and thus, potentially in linkage disequilibrium (LD) to some degree. In a QTL, study of cold tolerance in sorghum, Knoll and Ejeta (2008) tested two- and three-way gene interactions after applying the van der Waerden (1952) INT to their dependent variables. In both of these studies the researchers applied an INT to the dependent variables, which would normalize the marginal distribution of the phenotypes, not the residuals. Knoll and Ejeta (2008) cited Mansouri and Chang’s (1995) suggestion that the use of normal scores allows for the analysis of interactions. However, Mansouri and Chang (1995) only show that this practice is valid in balanced designs with normal error distributions, in which case the INT would not be needed. Furthermore, Mansouri and Chang showed the INT approach to be a conservative test for interaction.

To illustrate this issue in the context of genetic studies of epistasis, we conducted a simulation study for tests of gene \times gene interactions in association studies. Salter and Fawcett (1993) showed that rank-based test applied to testing interactions in the presence of main effects resulted in inflated Type 1 error rates. Therefore, we generated data for two models with genotypic main effects, but no interaction (epistasis):

$$\begin{aligned} \text{Additive Main Effect Pattern: } Y \\ = \beta_0 + \beta_1 G_1 - \beta_2 G_2 + \varepsilon; \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Dominant Main Effect Pattern: } Y \\ = \beta_0 + \beta_1 G_1 - \beta_2 G_2 + \beta_3 G_2^2 + \varepsilon; \end{aligned} \quad (3)$$

where, G_1 and G_2 are variables representing diallelic markers coded as: -1 for the homozygote with two copies of the minor allele; 0 for the heterozygote; and 1 for the homozygote with two copies of the major allele. Allele frequencies for G_1 and G_2 (P_1 and P_2 , respectively) were set at 0.5 and 0.25. LD between the G_1 and G_2 markers was defined by a correlation coefficient (r) and set at three levels $r = 0$ (No LD), 0.3 (moderate LD) and 0.7 (strong LD). The coefficients β_1 , β_2 , and β_3 were all set at one; β_0 was set at zero. Thus, Model 2 has only additive genetic effects, whereas Model 3 has an additive effect for the first marker and a dominance effect for the second marker. Similar to our other simulations, we generated four error terms (ε): Normal, LaPlace, $\chi_{(1)}^2$, and Weibull ($\lambda = 1$; $k = 0.5$).

The untransformed data and INT scores were analyzed with two statistical models. One model assumes that only

Table 5 Type 1 error rates for tests of epistasis (Additive main effect pattern, Eq. 2)

Test	Error distribution	LD	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Additive only model (4)—1 <i>df</i> (Y)	Normal	0	0.04832	0.00936	0.00100
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0	0.04880	0.00976	0.00098
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0	0.02864	0.00454	0.00038
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0	0.03532	0.00632	0.00074
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0	0.04982	0.01018	0.00100
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0	0.05116	0.01100	0.00124
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0	0.03536	0.00544	0.00044
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0	0.03970	0.00758	0.00076
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05114	0.01026	0.00108
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05132	0.01248	0.00194
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.19156 [‡]	0.06888 [‡]	0.01274 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.11214 [‡]	0.03144 [‡]	0.00476 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.05502	0.01198	0.00136
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.05376	0.01426	0.00270
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.20128 [‡]	0.07208 [‡]	0.01546 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.10312 [‡]	0.02878 [‡]	0.00554 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Normal	0.3	0.05034	0.01052	0.00098
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0.3	0.05158	0.01034	0.00106
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0.3	0.03740	0.00636	0.00034
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0.3	0.03492	0.00596	0.00050
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0.3	0.05000	0.01040	0.00108
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0.3	0.05238	0.01220	0.00198
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0.3	0.04008	0.00720	0.00070
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0.3	0.03564	0.00676	0.00108
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.04994	0.00980	0.00118
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.05576 [‡]	0.01684 [‡]	0.00428 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.15634 [‡]	0.05010 [‡]	0.00858 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.12646 [‡]	0.03384 [‡]	0.00618 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.05292	0.01158	0.00094
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.06164 [‡]	0.02498 [‡]	0.00992 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.16540 [‡]	0.05860 [‡]	0.01208 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.11816 [‡]	0.03416 [‡]	0.00702 [‡]

Note: 100,000 simulations

$N = 200$, $P_1 = 0.50$, $P_2 = 0.50$

[‡] Demonstrated significantly inflated Type 1 error rate

additive effects exist and the interaction is therefore analyzed as the cross-product of G_1 and G_2 :

$$Y = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1 G_2 + \varepsilon_j \quad (4)$$

It should be noted that even for data meeting the normality assumption, this model is misspecified for data generated with the Dominant Main Effect Pattern (3).

The second approach uses Cockerham's (1954) approach to model both the additive and non-additive effects of G_1 and G_2 :

$$Y = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1^2 + \beta_4 G_2^2 + \beta_5 G_1 G_2 + \beta_6 G_1 G_2^2 + \beta_7 G_1^2 G_2 + \beta_8 G_1^2 G_2^2 + \varepsilon_j, \quad (5)$$

yielding tests with 2 *df* for each main effect. The interaction is therefore analyzed as the cross-product of the G_1 and G_2 , main effect terms, yielding a 4 *df* interaction test ($H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$). It should be noted that this model is appropriately specified for data generated with both Additive (2) and Dominant Main Effect (3) Patterns.

Table 5 shows that when the INT is applied to data with additive genotypic main effects (2) and skewed ($\chi^2_{(1)}$; Weibull) error distributions, both tests for epistasis (gene–gene interaction) have drastically inflated Type 1 error rates. Table 6 reports the Type 1 error rates for data with additive and dominant genotypic main effects with no

Table 6 Type 1 error rates for tests of epistasis (Dominant Main Effect Pattern, Eq. 3)

Test	Error distribution	LD	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Additive only model (4)—1 <i>df</i> (Y)	Normal	0	0.05166	0.01046	0.00108
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0	0.04880	0.00976	0.00098
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0	0.04180	0.00724	0.00072
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0	0.04060	0.00754	0.00080
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0	0.05218	0.01102	0.00094
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0	0.05116	0.01100	0.00124
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0	0.05658 ^Y	0.01190	0.00110
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0	0.05244	0.01090	0.00102
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05124	0.01074	0.00098
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05132	0.01248	0.00194
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.22728 ^Y	0.08028 ^Y	0.01454 ^Y
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.20990 ^Y	0.07384 ^Y	0.01552 ^Y
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.05274	0.01210	0.00160
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.05376 ^Y	0.01426 ^Y	0.00270 ^Y
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.36864 ^Y	0.16258 ^Y	0.04032 ^Y
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.43908 ^Y	0.22136 ^Y	0.07180 ^Y
Additive only model (4)—1 <i>df</i> (Y)	Normal	0.3	0.48438 [*]	0.25542 [*]	0.08614 [*]
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0.3	0.05158	0.01034	0.00106
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0.3	0.55036 [*]	0.29514 [*]	0.09588 [*]
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0.3	0.03980	0.00712	0.00060
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0.3	0.48280 [*]	0.25694 [*]	0.08556 [*]
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0.3	0.05238	0.01220	0.00198
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0.3	0.61430 [*]	0.36058 [*]	0.13364 [*]
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0.3	0.04266	0.00848	0.00118
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.30734 [*]	0.13412 [*]	0.03490 [*]
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.05576 ^Y	0.01684 ^Y	0.00428 ^Y
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.80786 [*]	0.58828 [*]	0.29482 [*]
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.21216 ^Y	0.07300 ^Y	0.01480 ^Y
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.52270 [*]	0.30720 [*]	0.12274 [*]
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.06164 ^Y	0.02498 ^Y	0.00992 ^Y
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.96588 [*]	0.88262 [*]	0.67264 [*]
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.39810 ^Y	0.18200 ^Y	0.05258 ^Y

Note: 100,000 simulations

$N = 200$, $P_1 = 0.50$, $P_2 = 0.50$

* Inflated Type 1 error rate due to misspecified model

^Y Demonstrated significantly inflated Type 1 error rate

epistasis (3). These results show that even with normally distributed errors using a misspecified model (i.e., using the Additive Effects Only Model 4 when Dominant Main Effects exist) will lead to severely inflated Type 1 error when the markers are in LD, and this inflation worsens as the degree of LD increases (results not tabled). Applying the INT to the data does not alleviate this problem and as was the case in the condition presented in Table 5, the Type 1 error rates are inflated even when the properly specified model (5) is applied to INT transformed data with skewed errors. More subtly, these results show that INTs

applied to data with symmetric error distributions (Normal; LaPlace) tends to suppress the Type 1 error rate, which will lead to a reduction in statistical power to detect epistasis. Table 7 demonstrates that the Type 1 error rates for the INT transformed data are problematic even with different allele frequencies. Consistent with Wang and Huang (2002) who showed that the using INTs in QTL testing can inflate Type 1 error rate with non-additive genetic models, these simulations show that after applying an INT, tests for epistasis (i.e., gene–gene interaction, which is also a non-additive effect) can have inflated Type 1 error when error

Table 7 Type 1 error rates for tests of epistasis (Additive Main Effect Pattern, Eq. 2)

Test	Error distribution	LD	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Additive only model (4)—1 <i>df</i> (Y)	Normal	0	0.04900	0.00916	0.00086
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0	0.03816	0.00730	0.00076
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0	0.03372	0.00506	0.00036
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0	0.02964	0.00510	0.00050
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0	0.04962	0.01058	0.00152
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0	0.04464	0.01212	0.00262
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0	0.03700	0.00656	0.00046
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0	0.03298	0.00768	0.00072
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05198	0.01138	0.00146
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0	0.05528 [‡]	0.02212 [‡]	0.00788 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.16610 [‡]	0.05670 [‡]	0.01100 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0	0.09524 [‡]	0.03300 [‡]	0.00720 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.05018	0.01286	0.00274
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0	0.06398 [‡]	0.03184 [‡]	0.01646 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.17104 [‡]	0.06120 [‡]	0.01366 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0	0.09026 [‡]	0.03556 [‡]	0.01140 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Normal	0.3	0.05004	0.01028	0.00106
Cockerham model (5)—4 <i>df</i> (Y)	Normal	0.3	0.04644	0.00920	0.00104
Additive only model (4)—1 <i>df</i> (Blom)	Normal	0.3	0.03842	0.00702	0.00050
Cockerham model (5)—4 <i>df</i> (Blom)	Normal	0.3	0.03214	0.00564	0.00066
Additive only model (4)—1 <i>df</i> (Y)	LaPlace	0.3	0.05128	0.01014	0.00102
Cockerham model (5)—4 <i>df</i> (Y)	LaPlace	0.3	0.05556 [‡]	0.01576 [‡]	0.00360 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	LaPlace	0.3	0.03904	0.00642	0.00044
Cockerham model (5)—4 <i>df</i> (Blom)	LaPlace	0.3	0.03458	0.00790	0.00120
Additive only model (4)—1 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.05234	0.01206	0.00204 [‡]
Cockerham model (5)—4 <i>df</i> (Y)	$\chi^2_{(1)}$	0.3	0.06736 [‡]	0.02744 [‡]	0.01028 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.15276 [‡]	0.05064 [‡]	0.01006 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	$\chi^2_{(1)}$	0.3	0.12418 [‡]	0.04250 [‡]	0.01046 [‡]
Additive only model (4)—1 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.04984	0.01440	0.00350 [‡]
Cockerham model (5)—4 <i>df</i> (Y)	Weibull (1, 0.5)	0.3	0.08024 [‡]	0.04280 [‡]	0.02216 [‡]
Additive only model (4)—1 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.15784 [‡]	0.05298 [‡]	0.01066 [‡]
Cockerham model (5)—4 <i>df</i> (Blom)	Weibull (1, 0.5)	0.3	0.11670 [‡]	0.04138 [‡]	0.01288 [‡]

Note: 100,000 simulations

$N = 200$, $P_1 = 0.25$, $P_2 = 0.25$

[‡] Demonstrated significantly inflated Type I error rate

distributions are skewed and reduced power as compared to the untransformed data when error distributions are symmetric.

Several studies have shown that aligning the data before ranking yields valid tests of the interactions in factorial designs (Beasley 2002; Higgins and Tashtoush 1994). In the studies we reviewed, however, an INT was applied to the phenotype, which would affect the marginal distributions (main effects), not the residuals. The phenotypic data were not aligned before the application of the INT. Given the functional relationship of ranks and INTs and the issues we have noted with INTs, after the data are aligned and

ranked, we do not anticipate any particular advantage to using an INT over the ranks themselves. Hettmansperger and McKean (1977) and Jaeckel (1972) have developed approaches based on the rank of residuals that are robust and efficient relative to parametric methods for linear models when the normality assumption is not met. Conover (1973) has shown that functions of ranks can be derived for any alternative hypothesis, resulting in locally most powerful tests.

Finally, if one wants a non-parametric test in more complex situations, a legitimate permutation test can be constructed in the overwhelming majority of cases.

However, see Allison et al. (2006) and references therein for nuances and caveats. For example, one caveat is that the analysis of INTs, which have ranks as their basis, fundamentally changes the null hypothesis being tested, even if one performs a parametric *t*-test on INTs in a simple two-group design (Bradley 1968). The permutation distribution of ranks does not change due to factors such as within-group variance or distributional shape. Thus, without assuming that the error distributions have identical variance and shape, the *t*-test performed on INTs evaluates a null hypothesis of identical distributions, not necessarily identical location parameters (means). Thus, a statistical significant *t*-test performed on an INT indicates that the groups are not from identical populations, but that does not necessarily indicate that the difference is strictly due to mean differences (Vargha and Delaney 1998). Further, in more complicated statistical models, the null hypothesis being tested may become even less obvious (Beasley and Zumbo 2003).

Summary and conclusions

We demonstrated that INTs are not *necessarily* helpful and can in some cases make things worse. Our results indicate that there may be a few circumstances in which INTs will improve performance of testing procedures that assume normality; however, to our knowledge, data to support this proposition are largely wanting. Furthermore, it seems implausible that use of INTs could be superior to certain types of non-parametric testing (Good 2004).

Contrary to statements in the genetics literature, INTs do not necessarily maintain proper control over Type 1 error rates relative to the use of untransformed data unless they are coupled with permutation testing. Alternatives exist in the form of monotonic non-rank-based transformations, classic rank-based non-parametric tests, and computer intensive resampling methods. In larger sample sizes, the use of INTs offers one the opportunity to conduct valid inference using methods in which one must assume that a specified likelihood is correct even in the absence of permutation. Our results also indicate that INT may be useful for the analysis of extremely non-normal data from simple research designs with sample sizes large enough to make permutation testing intractable, especially when smaller effect sizes are suspected. However, the analysis of INTs does fundamentally change the null hypothesis being tested, which becomes an increasingly complicated issue with increases in the complexity of the statistical model. Furthermore, in more complex models linear models that have interaction terms, rank-based INTs can inflate the Type 1 error in some circumstances and reduce statistical power in other situations. Whether INTs are useful in other genetic research contexts relative to other available testing

approaches, remains to be demonstrated. We do not imply that such utility does not exist; however, rigorous research assessing the performance of INTs in situations germane to modern genetic research is warranted.

Acknowledgments We thank Christian Dina for asking cogent questions that inspired this commentary and for making useful comments on an earlier draft and also thank Jay Conover, Roger Berger, Brian Hicks, Rui Feng, Michael C. Neale, Goncalo Abecasis, Bernard S. Gorman and Alfred A. Bartolucci for their helpful advice or comments on earlier drafts. This article is supported in part by NIH grants P30DK056336, U54CA100949, R01ES09912, and T32HL072757.

References

- Akritis MG (1990) The rank transform method on some two factor designs. *J Am Stat Assoc* 85:73–78. doi:10.2307/2289527
- Allison DB, Neale MC, Zannolli RZ, Schork NJ, Amos CI, Blangero J (1999) Testing the robustness of the likelihood ratio test in a variance-component quantitative trait loci (QTL) mapping procedure. *Am J Hum Genet* 65:531–544. doi:10.1086/302487
- Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7(1):55–65. doi:10.1038/nrg1749
- Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–1211. doi:10.1086/301844
- Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 54:535–543
- Analysis System 130 (2003) Method and apparatus for analysis of data from biomolecular arrays, US Patent 6516276, (<http://www.patentstorm.us/patents/6516276-description.html>)
- Anokhin AP, Heath AC, Ralano A (2003) Genetic influences on frontal brain function: WCST performance in twins. *NeuroReport* 14(15):1975–1978. doi:10.1097/00001756-200310270-00019
- Ashton GC, Borecki IB (1987) Further evidence for a gene influencing spatial ability. *Behav Genet* 17(3):243–256. doi:10.1007/BF01065504
- Barnard GA (1957) *Mathematical gazette*, 41(338), 298–300. Review of: Tafeln zum Vergleich Zweier Stichproben mittels X-Test und Zeichentest tables for comparing two samples by X-test and sign test by B. L. van der Waerden; E. Nievergelt. doi:10.2307/3610142
- Basrak B, Klaassen CA, Beekman M, Martin NG, Boomsma DI (2004) Copulas in QTL mapping. *Behav Genet* 34(2):161–171. doi:10.1023/B:BEGE.0000013730.63991.ba
- Beasley TM (2002) Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multiv Behav Res* 37:197–226. doi:10.1207/S15327906MBR3702_02
- Beasley TM, Zumbo BD (2003) Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Comput Stat Data Anal* 42(4):569–593
- Berry WD (1993) *Understanding regression assumptions*. Sage, Newbury Park
- Blair RC, Sawilowsky SS, Higgins JJ (1987) Limitations of the rank transform statistic in test for interactions. *Comm Stat-Simul Comp* 16(113):3–1145
- Bliss CI (1967) *Statistics in biology*. McGraw-Hill, New York
- Blom G (1958) *Statistical estimates and transformed beta-variables*. Wiley, New York
- Blonigen DM, Carlson SR, Krueger RF, Patrick CJ (2003) A twin study of self-reported psychopathic personality traits. *Pers Individ Dif* 35:179–197. doi:10.1016/S0191-8869(02)00184-8

- Box GEP, Cox DR (1964) An analysis of transformations. *J R Stat Soc B* 26:211–252
- Bradley JV (1968) *Distribution-free statistical tests*. Prentice-Hall, New York
- Bradley JV (1978) Robustness? *Br J Math Stat Psychol* 31:144–152
- Chen WM, Abecasis GR (2006) Estimating the power of variance component linkage analysis in large pedigrees. *Genet Epidemiol* 30:471–484. doi:10.1002/gepi.20160
- Chernoff H, Savage IR (1958) Asymptotic normality and efficiency of certain nonparametric tests. *Ann Math Stat* 29:972–994. doi:10.1214/aoms/1177706436
- Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39:859–882
- Conover WJ (1973) Rank tests for one sample, two samples, and k samples without the assumption of a continuous distribution function. *Ann Stat* 1(6):1105–1125. doi:10.1214/aos/1176342560
- Conover WJ (1980) *Practical nonparametric statistics*, 2nd edn. Wiley, New York
- Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. *Am Stat* 35:124–133. doi:10.2307/2683975
- Diao G, Lin DY (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. *Am J Hum Genet* 77:97–111. doi:10.1086/431683
- Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. *Nat Genet* 39(10):1202–1207. doi:10.1038/ng2109
- Etzel CJ, Shete S, Beasley TM, Fernandez JR, Allison DB, Amos CI (2003) Effect of box–cox transformation on power of Haseman–Elston and maximum-likelihood variance components tests to detect quantitative trait loci. *Hum Hered* 55:108–116. doi:10.1159/000072315
- Farrell P, Rogers-Stewart K (2006) Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *J Stat Comp Simul* 76(9):803–816. doi:10.1080/10629360500109023
- Feir-Walsh BJ, Toothaker LE (1974) An empirical comparison of the ANOVA *F*-test, normal scores test and Kruskal–Wallis test under violation of assumptions. *Educ Psychol Measur* 34:789–799. doi:10.1177/001316447403400406
- Fisher RA, Yates F (1938) *Statistical tables for biological, agricultural, and medical research*, 1st edn. Oliver & Boyd, Edinburgh
- George VT, Elston RC (1987) Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiol* 4(3):193–201. doi:10.1002/gepi.1370040304
- Good PI (1999) *Resampling methods. A practical guide to data analysis*. Birkhauser, Boston
- Good PI (2004) Efficiency comparisons of rank and permutation tests by statistics in medicine 2001; 20:705–731. *Statistics in Medicine*, 23(5), 857. doi:10.1002/sim.1738
- Hájek J, Sidák F (1967) *Theory of rank tests*. Academic Press and Academia, Prague
- Harter HL (1961) Expected values of normal order statistics. *Biometrika* 48:151–165
- Headrick TC, Rotou O (2001) An investigation of the rank transformation in multiple regression. *Comput Stat Data Anal* 38:203–215. doi:10.1016/S0167-9473(01)00034-2
- Headrick TC, Sawilowsky SS (2000) Properties of the rank transformation in factorial analysis of covariance. *Comm Stat-Simul Comp* 29:1059–1087. doi:10.1080/03610910008813654
- Headrick TC, Vineyard G (2001) An empirical investigation of four tests of interaction in the context of factorial analysis of covariance. *Mult Linear Regress View* 27:3–15
- Hettmansperger TP, McKean JW (1977) A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* 19:275–284. doi:10.2307/1267697
- Hicks BM, Krueger RF, Iacono WG, McGue M, Patrick CJ (2004) Family transmission and heritability of externalizing disorders: a twin-family study. *Arch Gen Psychiatry* 61:922–928. doi:10.1001/archpsyc.61.9.922
- Hicks BM, Bernat E, Malone SM, Iacono WG, Patrick CJ, Krueger RF, McGue M (2007) Genes mediate the association between P3 amplitude and externalizing disorders. *Psychophysiology* 44(1):98–105. doi:10.1111/j.1469-8986.2006.00471.x
- Higgins JJ, Tashtoush S (1994) An aligned rank transform test for interaction. *Nonlinear World* 1:201–211
- Hodges JL, Lehmann EL (1962) Rank methods for combination of independent experiments in analysis of variance. *Ann Math Stat* 33:482–497. doi:10.1214/aoms/1177704575
- Hora SC, Conover WJ (1984) The *F*-statistic in the two-way layout with rank-score transformed data. *J Am Stat Assoc* 79:668–673. doi:10.2307/2288415
- Jaekel LA (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann Math Stat* 43:1449–1458. doi:10.1214/aoms/1177692377
- James GS (1959) The Behrens–Fisher distribution and weighted means. *J R Stat Soc [Ser A]* 21:73–80
- Keselman HJ, Rogan JC, Feir-Walsh BJ (1977) An evaluation of some nonparametric and parametric tests for location equality. *Br J Math Stat Psychol* 30:213–221
- Knoke JD (1991) Nonparametric analysis of covariance for comparing change in randomized studies with baseline values subject to error. *Biometrics* 47(2):523–533. doi:10.2307/2532143
- Knoll J, Ejeta G (2008) Marker-assisted selection for early-season cold tolerance in sorghum: QTL validation across populations and environments. *Theor Appl Genet* 116(4):541–553. doi:10.1007/s00122-007-0689-8
- Kohr RL, Games PA (1974) Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. *J Exp Educ* 43:61–69
- Kraja AT, Corbett J, Ping A, Lin RS, Jacobsen PA, Crosswhite M, Borecki IB, Province MA (2007) Rheumatoid arthritis, item response theory, Blom transformation, and mixed models. *BMC Proc* 1(Suppl. 1):S116
- Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 47:583–621. doi:10.2307/2280779
- Li M, Boehnke M, Abecasis GR, Song PX (2006) Quantitative trait linkage analysis using Gaussian copulas. *Genetics* 173(4):2317–2327. doi:10.1534/genetics.105.054650
- Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 18:50–60. doi:10.1214/aoms/1177730491
- Mansouri H, Chang G-H (1995) A comparative study of some rank tests for interaction. *Comput Stat Data Anal* 19:85–96. doi:10.1016/0167-9473(93)E0045-6
- Maritz JS (1982) *Distribution-free statistical methods*. Chapman and Hall, London
- Martin LJ, Crawford MH (1998) Genetic and environmental components of thyroxine variation in Mennonites from Kansas and Nebraska. *Hum Biol* 70(4):745–760
- McSweeney M, Penfield D (1969) The normal scores test for the c-sample problem. *Br J Math Stat Psychol* 20:187–204
- Mehta T, Tanik M, Allison DB (2004) Toward sound epistemological foundations of statistical methods for high dimensional biology. *Nat Genet* 36:943–947. doi:10.1038/ng1422
- Miccieri T (1989) The unicorn, the normal curve, and other improbable creatures. *Psychol Bull* 105:156–166. doi:10.1037/0033-2909.105.1.156

- Nanda NJ, Rommelse Arias-Vásquez A, Altink ME, Buschgens CJM, Fliers E, Asherson P, Faraone SV, Buitelaar JK, Sergeant JA, Oosterlaan J, Franke B (2008) Neuropsychological endophenotype approach to genome-wide linkage analysis identifies susceptibility loci for ADHD on 2q21.1 and 13q12.11. *Am J Hum Gen* 9:9–105
- Neave HR, Wothington PL (1989) *Distribution-free tests*. Routledge, New York
- Peng B, Yu RK, DeHoff KL, Amos CI (2007) Normalizing a large number of quantitative traits using empirical normal quantile transformation. *BMC Proc*, (Suppl 1), p S156
- POLY: Computer program for polygenic analysis and power analysis (2003) [<http://www.sph.umich.edu/csg/chen/public/software/poly/>]
- Pratt JW (1964) Robustness of some procedures for the two-sample location problem. *J Am Stat Assoc* 59:665–680. doi:10.2307/2283092
- Przybyla-Zawislak BD, Thorn BT, Alia SF et al (2005) Identification of rat hippocampal mRNAs altered by the mitochondrial toxicant, 3-NPA. *Ann N Y Acad Sci* 1053:162–173. doi:10.1196/annals.1344.014
- Pulli K, Karma K, Norio R, Sistonen P, Göring HH, Järvelä I (2008) Genome-wide linkage scan for loci of musical aptitude in Finnish families: evidence for a major locus at 4q22. *J Med Genet* 45(7):451–456. doi:10.1136/jmg.2007.056366
- Ray WD, Pitman A (1961) An exact distribution of the Fisher–Behrens–Welch statistic for testing the difference between the means of two normal populations with unknown variances. *J R Stat Soc [Ser A]* 23:377–384
- Salter KC, Fawcett RF (1993) The ART test of interaction: a robust and powerful test of interaction in factorial models. *Comm Stat-Simul Comp* 22:137–153
- Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orru M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, Cao A, Lakatta E, Abecasis GR (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLOS Genetics* 3(7):e115. doi:10.1371/journal.pgen.0030115
- Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genetics* 3(7):e114. doi:10.1371/journal.pgen.0030114
- Shete S, Beasley TM, Etzel CJ, Fernández JR, Chen J, Allison DB, Amos CI (2004) Effect of Winsorization on power and type I error of variance components and related methods of QTL detection. *Behav Genet* 34:153–159. doi:10.1023/B:BEGE.0000013729.26354.da
- Silverman EK, Province MA, Campbell EJ, Pierce JA, Rao DC (1990) Biochemical intermediates in alpha 1-antitrypsin deficiency: residual family resemblance for total alpha 1-antitrypsin, oxidized alpha 1-antitrypsin, and immunoglobulin E after adjustment for the effect of the Pi locus. *Genet Epidemiol* 7(2):137–149. doi:10.1002/gepi.1370070204
- SOLAR: Sequential Oligogenic Linkage Analysis Routines (2008) [<http://www.sfbr.org/solar/>]
- Spren P, Smeeton NC (2001) *Applied nonparametric statistical methods*, 3rd edn. Chapman & Hall, London
- Stuart A (1954) Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *J Am Stat Assoc* 49:147–157. doi:10.2307/2281041
- Thompson GL (1991) A note on the rank transform for interactions. *Biometrika* 78:697–701. doi:10.1093/biomet/78.3.697
- Thompson GL (1993) A correction note on the rank transform for interactions. *Biometrika* 80:711
- Toothaker LE, Newman D (1994) Nonparametric competitors to the two way ANOVA. *J Educ Behav Stat* 19:237–273
- Tukey JW (1962) The future of data analysis. *Ann Math Stat* 33:1–67. doi:10.1214/aoms/1177704711
- Tzou GG, Everson DO, Bulls RC, Olson DP (1991) Classification of beef calves as protein-deficient or thermally stressed by discriminant analysis of blood constituents. *J Anim Sci* 69:864–873
- Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JN, Mott R, Flint J (2006) Genetic and environmental effects on complex traits in mice. *Genetics* 174(2):959–984. doi:10.1534/genetics.106.060004
- van den Oord EJ, Simonoff E, Eaves LJ, Pickles A, Silberg J, Maes H (2000) An evaluation of different approaches for behavior genetic analyses with psychiatric symptom scores. *Behav Genet* 30(1):1–18. doi:10.1023/A:1002095608946
- van der Waerden BL (1952) Order tests for the two-sample problem and their power. *Proc Koninklijke Nederlandse Akademie van Wetenschappen. Ser A* 55:453–458
- Vargha A, Delaney HD (1998) The Kruskal-Wallis test and stochastic homogeneity. *J Educ Behav Stat* 23:170–192
- Wang K, Huang J (2002) A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *Am J Hum Genet* 70(2):412–424. doi:10.1086/338659
- Welch BL (1947) The generalization of Student's problem when several different population variances are involved. *Biometrika* 34:28–35
- Wilcox RR (1995) ANOVA: a paradigm for low power and misleading measures of effect size? *Rev Educ Res* 65:51–77
- Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics* 1:80–83. doi:10.2307/3001968
- Wu X, Cooper RS, Borecki I, Hanis C, Bray M, Lewis CE, Zhu X, Kan D, Luke A, Curb D (2002) A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *Am J Hum Genet* 70(5):1247–1256. doi:10.1086/340362
- Yang R, Yi N, Xu S (2006) Box-Cox transformation for QTL mapping. *Genetica* 128(1–3):133–143. doi:10.1007/s10709-005-5577-z
- Yuen KK (1974) The two-sample trimmed t for unequal population variances. *Biometrika* 61:165–170
- Zak M, Baierl A, Bogdan M, Futschik A (2007) Locating multiple interacting quantitative trait Loci using rank-based model selection. *Genetics* 176(3):1845–1854. doi:10.1534/genetics.106.068031
- Zimmerman DW (1996) A note on homogeneity of variance of scores and ranks. *J Exp Educ* 64:351–362
- Zimmerman DW (2004) A note on preliminary tests of equality of variances. *Br J Math Stat Psychol* 57(1):173–181
- Zumbo BD, Coulombe D (1997) Investigation of the robust rank-order test for non-normal populations with unequal variances: the case of reaction time. *Can J Exp Psychol* 51:139–150. doi:10.1037/1196-1961.51.2.139