Beyond the traditional simulation design for evaluating type 1 error rate: from 'theoretical' to 'empirical' null

Ting Zhang [1], Lei Sun [2,1]

[1] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, ON, Canada

[2] Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, ON, Canada

Corresponding Author:

Lei Sun (sun@utstat.toronto.edu)

100 St. George Street

Department of Statistical Sciences

University of Toronto

Toronto, ON, M5S 3G3

Abstract

When evaluating a newly developed statistical test, the first step is to check its type 1 error (T1E) control using simulations. This is often achieved by the standard simulation design S0 under the so-called 'theoretical' null of no association. In practice, whole-genome association analyses scan through a large number of genetic markers (Gs) for the ones associated with an outcome of interest (Y), where Y comes from an unknown *alternative* while the majority of Gs are *not* associated with Y, that is under the 'empirical' null. This reality can be better represented by two other simulation designs, where design S1.1 simulates Y from an alternative model based on G then evaluates its association with independently generated G^{new}, while design S1.2 evaluates the association between permutated Y^{perm} and G. More than a decade ago, Efron (2004) has noted the important distinction between the 'theoretical' and 'empirical' null in false discovery rate control. Using scale tests for variance heterogeneity and location tests of interaction effect as two examples, here we show that not all null simulation designs are equal. In examining the accuracy of a likelihood ratio test, while simulation design S0 shows the method has the correct T1E control, designs S1.1 and S1.2 suggest otherwise with empirical T1E values of 0.07 for the 0.05 nominal level. And the inflation becomes more severe at the tail and does not diminish as sample size increases. This is an important observation that calls for new practices for methods evaluation and interpretation of T1E control.

Key words: type 1 error, simulation, whole-genome association scans, variance heterogeneity, interaction

1 Introduction

Type 1 error (T1E) control evaluation using simulations is always the first step in understanding the performance of any newly developed statistical test. To formulate the problem more precisely, let us consider the current large-scale genome-wide association studies (GWAS) or next-generation sequencing (NGS) studies of complex and heritable traits. These studies scan through millions or more genetic markers (Gs) across the genome for the ones associated with a trait of interest (Y), while accounting for environmental effects. Many Y-G association tests have been developed, and they often require the assumption of (approximately) normally distributed errors to maintain T1E accuracy, with some being more robust than others. For example, Bartlett test for variance heterogeneity has been shown to have large inflated T1E rates when the error term e follows a t- or χ^2-distribution (Struchalin et al. 2010), and the likelihood ratio test (LRT) is similarly sensitive (Cao et al. 2014), while Levene's test appears to be more robust (Soave et al. 2015; Soave and Sun 2017).

Standard T1E simulation design, denoted as S0, generates phenotype data $Y_0 \sim e$ under the 'theoretical' null model of no association, then independently generates genotype data G and estimates the empirical T1E rate from $Y_0 \sim G + \epsilon$; for notation simplicity and without loss of generality, intercept and additional covariates Zs are omitted from the conceptual expression of the regression model. A method is generally considered sound if T1E is well controlled under the $e \sim N(0, \sigma^2)$ assumption, and robustness is then evaluated by assuming other distribution forms for e. Given statistical accuracy of T1E control, statistical efficiency in terms of power will be studied by generating phenotype under an alternative, $Y_1 \sim G + e$, and often it is assumed that $e \sim N(0, \sigma^2)$.

In practice, GWAS and NGS receive an *empirical* Y that comes from an unknown *alternative*, and a large number of Gs of which the majority are *not* associated with Y. That is, most Y-G association pairs are in fact under the 'empirical' null. Now consider two alternative simulation designs to evaluate T1E control. Design S1.1 simulates $Y_1 \sim G + e$ from an alternative, then it independently generates G^{new} and evaluates T1E from $Y_1 \sim G^{new} + \epsilon$. Design S1.2 permutes the simulated Y_1 and evaluates T1E from $Y_1^{perm} \sim G + \epsilon$. A important question can then be asked as to whether the S1.1 and S1.2 designs lead to similar T1E conclusion as the S0 design. In particular,

even if the $e \sim N(0, \sigma^2)$ assumption was true and a test appeared to be accurate based on the S0 evaluation, do we expect it to perform well in real data which are better represented by the S1.1 and S1.2 simulation designs; note that Y_1 is in contrast to Y_0 and ϵ may or may be normally distributed. The answer would depend on the type of test statistics used.

Efron (2004) has brought up the discussion of the 'theoretical' vs. 'empirical' null more than a decade ago. Focusing on controlling the false discovery rate (FDR), Efron (2004) outlined several possible sources of non-normality including unobserved covariates and hidden correlation, and he proposed an empirical Bayes approach to the problem. Here, we study the practical implications of T1E evaluation based on the the commonly used 'theoretical' null simulation design S0 in the context of whole-genome scans. We show that while a method may appear to be accurate under S0 and assuming normality, it can have incorrect T1E rates under the 'empirical' null of S1.1 or S1.2 and also 'assuming normality'. The fundamental cause of the discrenpancy is that, in evaluating $Y_1 \sim G^{new} + \epsilon$ (or $Y_1^{perm} \sim G + \epsilon$), the marginal distribution of Y_1 may not be normal even if it was generated assuming normality ($Y_1 \sim G + e$), conditional on the true causal G and other covariates.

As a proof-of-principle, we will focus on scale tests for variance heterogeneity, recently proposed to identify Gs associated with *variance* of a quantitative trait Y (Pare et al. 2010; Aschard et al. 2013; Cao et al. 2014; Soave et al. 2015). Traditional Y-G association tests focus on location parameters, studying changes in mean of Y across different genotype groups. Gene-environment (GxE) and gene-gene (GxG) are expected for complex traits. However, in practice, incomplete E data may preclude straightforward GxE interaction analyses, and computational or multiple hypothesis testing concerns can make whole-genome exhaustive GxG interaction searches undesirable. It was then recognized that because un-modelled interactions induce variance heterogeneity in Y when conditional only on G, scale tests such as Levene's test, originally developed for model diagnostics, can be used to indirectly test for the interaction effects; it is worth noting that the causes of variance heterogeneity are multifaceted beyond potential interactions (Sun et al. 2013; Dudbridge and Fletcher 2014; Wood et al. 2014).

Inference of scale parameters is generally more sensitive than that of location parameters (Khan and Rayner, 2003). Thus, the distinction between the 'theoretical' and 'empirical' null can be particularly consequential for these emerging association tests that are designed to improve power

by going beyond the first moment. In this work, we reveal the existing problems in T1E evaluation based on the 'theoretical' null simulation design S0. We show that (1) a T1E conclusion drawn from S0 could be different from the two alternative 'empirical' null simulation designs S1.1 and S1.2; (2) The T1E discrepancy can remain as sample size increases; (3) The T1E issue may be more severe at the tail.

In some settings, the 'theoretical' vs. 'empirical' null can also affect inference of location parameters in a regression, in addition to the better known cause of mean or variance model mis-specification. Assume E was available for direct modelling of the GxE interaction effect, Voorman et al. (2011) and Rao and Province (2016) showed that T1E rate of testing GxE or Gx$G_{non-repeating}$ in a whole-genome interaction scan can be sensitive to reasons beyond model mis-specifications. $G_{non-repeating}$ represents a fixed SNP G and we are testing its interaction with other SNPs, and Gx$G_{non-repeating}$ is statistically similar to GxE which we use, hereinafter, to refer to both. Focusing on inflated or deflated genomic inflation factor λ_{GC} (Devlin and Roeder 1999), Rao and Province (2016) demonstrated a larger variation in λ_{GC} (similar to a larger variation in T1E rates between different whole-genome association scans), when testing the interaction effect as compared to the main effect under the 'theoretical' null. They attributed this to dependence between the interaction test statistics, because E (or $G_{no-repeating}$) is fixed between tests. And they noted that increasing sample size mitigates the problem. Here, we use this opportunity to revisit location testing of interaction effect. We show that, under the conventional 'theoretical' null, while T1E rates are indeed variable between simulation replicates, the average T1E rate is correct regardless of the sample size. In contrast, under the 'empirical' null, a different picture emerges as in the scale test setting above. In what follows, we first describe in Section 2 the scale tests to be investigated and the three simulation designs, S0, S1.1 and S1.2. We then provide numerical results from extensive simulation studies in Section 3, together with direct location tests for main and interaction effects. Importantly, we note that even if the departure from normality is generally minor in practice and appears to pass standard diagnostic tests for non-normality, the different null simulation designs can still noticeably affect conclusion regarding T1E control for some tests. Finally, in Section 4, we remark that future T1E evaluation and interpretation should go beyond the traditional 'theoretical' null and adopt the alternative 'empirical' null simulation designs.

2 Methods

For association study of a complex trait Y using a sample of size n, we first define genotype data G_i for individual i at each SNP under the study. As in tradition, G_i denotes the number of copies of the minor allele, coded additively as $G_i = 0$, 1 and 2. And G_i is assumed to come from a multinomial distribution, $G_i \sim \text{multinomial}(1, ((1-f)^2, 2f(1-f), f^2))$, where f is minor allele frequency, MAF.

The analytical context of using scale tests to detect SNP G that influences variance of trait Y is the following. Suppose the true generating model is

$$Y = \beta_G G + \beta_E E + \beta_{GE} G \times E + e, \text{ where } e \sim N(0, \sigma^2), \tag{1}$$

and suppose information regarding E was not collected, then the working model can only account for the main effect of G. However, it is straightforward to show that variances of Y stratified by the three genotype groups of G differ if $\beta_{GE} \neq 0$,

$$Var(Y|G) = (\beta_E + \beta_{GE} G)^2 Var(E) + \sigma^2 = \sigma^2_G. \tag{2}$$

Thus, when E is missing and direct interaction modelling is not feasible, scale tests can be utilized to identify G associated with variance of Y (Pare et al. 2010). A joint location-scale testing framework can provide robustness against either $\beta_G = 0$ or $\beta_{GE} = 0$, and it can improve power if both main and interaction effects are present (Soave et al. 2015). Here we focus on studying the more sensitive scale tests, because the power of the joint test depends on the individual components.

Different scale tests have been studied in this context, and chief among them are the Levene's test (Levene et al. 1960) considered by Pare et al. (2010) and Soave et al. (2015), and the LRT considered by Cao et al. (2014). Levene's test for variance heterogeneity between k groups is an ANOVA of the absolute deviation of each observation y_i from its group mean or median. The resulting test statistic *Levene* follows a $\text{F}(k-1, n-k)$ distribution under normality, and it is asymptotically $\chi^2_{k-1}/(k-1)$ distributed; $k = 3$ in our case. Using median instead of mean to measure the spread within each group has been proved to be more robust to non-normality, particularly for t-distributed or skewed data (Brown et al. 1974; Soave and Sun 2017). And we will be using the median version of *Levene*

in the remaining paper.

The variance likelihood ratio test considered by Cao et al. (2014) contrasts the null model of no variance difference with the alternative model,

$$Y = \beta_G G + e, e \sim N(0, \sigma^2) \;\; vs. \;\; Y = \beta_G G + e, e \sim N(0, \sigma_G^2), \tag{3}$$

and conduct the corresponding LRT for $H_0 : \sigma^2_{G=0} = \sigma^2_{G=1} = \sigma^2_{G=2}$. The corresponding test statistic LRT_v is asymptotically χ^2_2 distributed; joint mean-variance LRT considering both $\beta_G = 0$ and $\sigma^2_G \equiv \sigma^2$ can be readily conducted. Cao et al. (2014) has pointed out that LRT_v is sensitive to the normality assumption, but under normality they have demonstrated that LRT_v has the correct T1E control. However, we show in the following that although this conclusion is analytically correct under the 'theoretical' null, it can be invalid when the method is applied to whole-genome scans which are better represented by the 'empirical' null.

[Table 1 here]

Table 1 outlines the different null simulation designs, where S0 is the 'theoretical' null considered by Cao et al. (2014) as in convention, while S1.1 and S1.2 are the 'empirical' null designs that better represent the condition of real data. Under the S1.1 and S1.2 designs, the marginal distribution of phenotype Y is a weighted linear combination of normal distributions (Supplementary Materials). Thus, tests thought to be accurate based on S0 may have T1E issues based on S1.1 and S1.2, depending on the weighting factors and the means and variances of individual normal distributions. For example, LRT_v, the LRT statistics for variance heterogeneity can be shown to be asymptotically equal to the weighted sum of independent $\chi^2_{(1)}$ (Supplementary Materials and Theorem 3.4.1(1) of Yanagihara et al. 2005). Thus, before the simulation study in the next section, we shall expect that LRT_v will have T1E issue when the simulated data is not normally distributed marginally.

Assume that E was known, we can then directly test the interaction effect β_{GE} using classical likelihood ratio test ($LRT_{\beta_{GE}}$) or the score test ($Score_{\beta_{GE}}$) based on model (1). In that case, it is straightforward to define the 'empirical' null design. That is, we first simulate $Y_1 = \beta_G G + \beta_E E + \beta_{GE} G \times E + e$ using model (1) based on the true G and E. We then independently simulate G^{new} and test β_{GE} from $Y_1 = \beta_G G^{new} + \beta_E E + \beta_{GE} G^{new} \times E + \epsilon$; similarly for S1.2.

There are a number of 'theoretical' null designs possible. For example, we can simulate $Y_0 = \beta_E E + e$ without the main G effect (S0.1, Model I of Rao and Province 2016), or $Y_0 = \beta_G G + \beta_E E + e$ with the main effect (S0.2, Model II of Rao and Province 2016). In addition, we can also implement each 'theoretical' null model in two ways. Consider $Y_0 = \beta_G G + \beta_E E + e$, we can simply simulate *nrep* sets of G and E to generate Y_0. Alternatively, within each of *nrep.out* replicates (e.g. 100) of E in an outer simulation loop, we can simulate *nrep.in* replicates (e.g. 10^5) of G and use them combined with the fixed E to simulate *nrep.in* replicates of Y_0. We can then test β_{GE} and estimate the T1E rate using the *nrep.in* replicates, similar to a whole-genome scan. Finally, we can average the T1E rate over *nrep.out* replicates to account for sampling variation inherent in simulation of E or one scan. We will be examining all four combinations for the 'theoretical' null. For the 'empirical' null, although the single-loop approach is possible, the *nrep.in* $\times$ *nrep.out* double-loop is more intuitive. That is, within each of *nrep.out* replicates of G and E, and Y_1 simulated based on model (1), we simulate *nrep.in* replicates of G^{new} for testing and T1E rate estimation. We then average across the *nrep.out* replicates.

3 Simulations

For evaluating scale tests for variance heterogeneity, we considered two modelling frameworks adopted, respectively, by Cao et al. (2014) and Aschard et al. (2013) (Tables 2). Cao et al. (2014) used model (3) to directly simulate variance heterogeneity in Y stratified by G. In contrast, Aschard et al. (2013) used model (1) to indirectly simulate variance heterogeneity that has better genetic epidemiology interpretation, because the size of β_{GE} corresponds to power of scale tests under alternatives. Assume that E was known, model (1) also allows us to evaluate T1E control for our second study of directly testing for the interaction effect β_{GE}. Conveniently, the corresponding 'empirical' null model S0 in Table 2, $Y_0 = \beta_E E + e$, is conceptually the same as the simulation model I of Rao and Province (2016), except E was $G_{non-repeating}$.

[Table 2 here]

For each parameter value combination in Table 2, instead of studying power, we focused on evaluating T1E control of the LRT and Levene's scale tests for variance heterogeneity, and location

tests for interaction effect, by contrasting the proposed 'empirical' null with the previously considered 'theoretical' null. We first generated genotype and phenotype data for G, G^{new}, (and E if needed), Y_0, Y_1 and Y_1^{perm} as described in Tables 1 and 2. We focus on the $nrep.in \times nrep.out$ double-loop implementation, but we note that the single loop design leads to the same conclusion as long as the total number of replicates is large (results not shown).

First, assume that information regarding E was not collected in practice, we applied the scale tests, LRT_v and Levene, using the following working models,

- S0: $Y_0 \sim G$
- S1.1, an alternative 'empirical' null: $Y_1 \sim G^{new}$
- S1.2, another alternative 'empirical' null: $Y_1^{perm} \sim G$

That is, we tested $Var(Y_0|G)$ across G under the 'theoretical' null of no association of S0, and $Var(Y_1|G^{new})$ across G^{new} and $Var(Y_1^{perm}|G)$ across G under the 'empirical' null of no association of, respectively, S1.1 and S1.2. We recorded the empirical T1E rates for each setting and bolded in red colour the ones that exceed the $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ range, where α is the nominal T1E rate and $rep.in$ is the number of simulation replicates used to estimate the empirical T1E rate for each of the $rep.out$ replicates. Thus, $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ is a conservative interval. For completeness, we also kept the results of location tests (LRT_m and $Score_m$) for testing mean differences in Y across G, similarly contrasting the 'theoretical' null design of S0 with the alternative 'empirical' null designs of S1.1 and S1.2.

Revisiting the subtle dependency issue between interaction tests examined by Rao and Province (2016), we then assumed that E was available. That is, we tested β_{GE} in $Y = \beta_G G + \beta_E E + \beta_{GE} G \times E + e$ using the likelihood ratio test ($LRT_{\beta_{GE}}$) and the score test ($Score_{\beta_{GE}}$). However, S0 evaluated the association between Y_0 and $G \times E$ under the conventional 'theoretical' null design, while S1.1 examined Y_1 and $G^{new} \times E$, and S1.2 studied Y_1^{perm} and $G \times E^{perm}$ under the alternative 'empirical' null designs.

4 Results

As expected from the analytical insights, results in Table 3 show that while location tests for phenotypical mean differences (LRT_m and $Score_m$) are generally robust to the choice of 'theoretical' (S0) vs. 'empirical' (S1.1 or S1.2) null, it is not the case for the LRT scale test (LRT_v) for variance heterogeneity; the empirical T1E rates of Levene's test were slightly deflated but not significantly. Different choice of the null lead to different conclusions regarding the accuracy of LRT_v. For example, simulation design S0 shows LRT_v has the correct T1E control across the parameter values considered, but designs S1.1 and S1.2 suggest otherwise with empirical T1E values of 0.07 for the nominal $\alpha = 0.05$ level for some settings. While the increased T1E rates under the S1.1 and S1.2 'empirical' null designs appear to be mild and occur in extreme models (i.e. large un-modelled β_{GE} GxE interaction effect), results in Table 4 demonstrate that the T1E issue under the 'empirical' null simulation designs of S1.1 and S1.2 can be more severe at the tail. For example, for the nominal $\alpha = 1 \times 10^{-5}$ level, the empirical T1E rate can be as high as 11.5×10^{-5}. Because the genome-wide significance level for GWAS is $\alpha = 5 \times 10^{-8}$ (Dudbridge and Gusnanto 2008), an inflation of false positive findings can be of a real problem in practice. Further, results in Table 5 confirm that increasing sample size n (from 10^3 to 10^4) does not mitigate the discrepancy in T1E conclusion drawn from the 'theoretical' vs. 'empirical' null. The root cause is that Y_1 marginally is not normally distributed, even if it was generated (conditional on the true G) using a normally distributed error e term.

[Table 3 here]

[Table 4 here]

[Table 5 here]

In practice, it is routine (and recommended) to display and examine the empirical distribution of a trait under the study. However, Figure 1 shows that even under the most extreme setting where $\beta_{GE} = 1$, the marginal histogram of Y appears to be approximately normal visually, unless a formal diagnostic test for normality was conducted. The slightly right-skewed empirical distribution of Y is the result of mixing six conditional distributions of Y, each perfectly normally distributed conditional on the causal G and E; this is the key difference between the 'theoretical' and 'empirical'

null simulation designs, regardless of the sample size. For a less extreme case where $\beta_{GE} = 0.2$, although both the histogram and Q-Q plot (Figure S1) suggest that normal distribution is a good fit (passing the Shapiro-Wilk normality test), the T1E discrepancy between the 'theoretical' and 'empirical' null remains albeit less severe as shown in Table 3 and Figure S2.

[Figure 1 here]

[Figure 2 here]

The asymptotic distribution of LRT_v under the 'empirical' null is a weighted sum of χ^2_1 (Supplementary Materials). Figure 2 compares the asymptotic distribution (black solid curve) with the finite-sample distribution (red dashed curve) of LRT_v under the 'empirical' null, as well as with χ^2_2 (blue dot-dashed curve), which is the asymptotic distribution of LRT_v under the 'theoretical' null. While the asymptotic distribution derived under the 'empirical' null well approximates the finite-sample one, it is clear that the distributions of LRT_v differ between the 'empirical' and 'theoretical' null; the difference is more visible on the scale of critical value for statistical significance (the vertical lines). Thus, applying LRT_v to empirical GWAS or NGS while using the significance threshold derived from χ^2_2 can lead to T1E problem.

Tables 3, 4 and 5 also included T1E results for testing phenotypic mean (as opposed to variance) difference across the genotype groups. Although location testing for the main effects are generally quite robust to the assumption of normality, problem can arise when testing for interaction effects beyond model mis-specification (Rao and Province 2016).

In testing the interaction effect β_{GE} ($\beta_{GG_{non-repeating}}$ to be more precise), Rao and Province (2016) used the classical 'theoretical' null simulation design considering both S0.1 (without the main G effect) and S0.2 (with the main G effect). Regardless, Figures 1B-1C of Rao and Province (2016) showed that the variation in the resulting λ_{GC} was substantially bigger when testing β_{GE} than testing β_G. And their Figures 1D and 1E demonstrated that the variation diminishes as sample size increases. However, we note that this observation was made before averaging across the 414 simulated interaction scans/datasets; each scan contained 20,000 SNPs from which a λ_{GC} value was estimated.

[Table 6 here]

The results of Rao and Province (2016) are consistent with ours shown in Figure S5. Figure

S5 showed that scan-specific estimated T1E rates are indeed variable and become less so as sample size increases; 100 GxE interaction scans of 10^5 SNPs each. However, it is important to note that the average T1E rate across *nrep.out* simulated scans reflects better the long-run behaviour of a method. Alternatively, assume $nrep.out = 1$, increasing the number of SNPs (i.e. the size of *nrep.in*) will decrease the sampling variation inherent in estimating T1E based on simulation studies. Indeed, results in Table 6 show that the T1E rate of testing β_{GE}, estimated from $10^5 \times 100$ ($nrep.in \times nrep.out$) simulated replicates, is well controlled under the conventional 'theoretical' (S0) null simulation design. But, this is not the case for the 'empirical' (S1.1 or S1.2) null simulation designs. Similar to the LRT_v scale test for variance heterogeneity, the discrepancy between two types of designs becomes more prominent at the tail and persists as sample increases (Table 6).

5 Discussion

In this article, we highlight the importance of distinguishing the 'theoretical' and 'empirical' null distributions, first noted by Efron (2004), in a different application context. Focusing on scale tests for variance heterogeneity and through simulation studies, we showed that conclusions of type 1 error control of a statistical test could differ depending on the choice of the null. For example, the LRT variance test appears to be accurate under the 'theoretical' null but invalid under the 'empirical' null (Tables 3, 4 and 5, and Figure S2). Although the error term for generating the phenotype or outcome data was assumed to be normally distributed, the increased T1E rates under the 'empirical' null are, fundamentally, attributed to sensitivity of LRT_v to departure from normality, because the marginal distribution of the empirical outcome data was not normal (Figures 1 and S1). Thus, tests shown to be sensitive to the assumption of normality are particularly vulnerable when applied to real data that are better represented by the 'empirical' null than the 'theoretical' null.

In practice, investigators often rely on visual inspection of histograms of outcome data as illustrated in Figures 1 and S1. And we have noted that the departure from normality does not have to be severe to have an effect on tests such as LRT_v. For example, Soave et al. (2015) applied the LRT_v test of Cao et al. (2014) to a GWAS of lung function measures in cystic fibrosis subjects. Despite the fact that the lung measures were approximately normally distributed and permuted prior to the

variance association analysis, the histogram of GWAS p-values clearly showed an increased T1E rate (Figure S2.G of Soave et al. 2015); the actual application was a joint LRT_m and LRT_v test but the T1E issue was due to the LRT_v component. Furthermore, for data appear to deviate from normal such as that in Figure 1, even if investigators chose to perform some standard normal transformations, the T1E issue can persist. For example, let us consider the phenotype data simulated based on Aschard's genetic model, as described in Table 2 where $\beta_{GE} = 1$ (Figure 1). After square-root or log transformations (Goh and Yap 2009), although the empirical marginal distribution of the phenotype improved as expected (Figure S3), the severity of T1E inflation of LRT_v in fact worsened under the 'empirical' S1.1 and S1.2 null (Figure S4).

Beyond scale test of variance heterogeneity, Voorman et al. (2011) showed that spurious false positives can occur in genome-wide scans for GxE interactions, particularly in the presence of model mis-specification. And Rao and Province (2016) also presented inflated/deflated genomic inflation factors in a GxG interaction scan when one SNP is anchored (i.e. Gx$G_{non-repeating}$), using the conventional 'theoretical' null simulation design without any apparent model mis-specification. In our simulation studies, the situation when E was assumed available for direct modelling of the interaction term is similar to the dependency case examined previously. We note that the large variation in λ_{GC} estimate demonstrated by Rao and Province (2016) corresponds to the sampling variation inherent in estimating T1E rate from *nrep.in* replicates/SNPs across *nrep.out* replicates. This, however, does not translate to T1E issue based on the classical frequentist interpretation. Results in Table 6 show that, similar to scale test of variance, T1E conclusion for location test of interaction effect β_{GE} is sensitive to the choice of 'theoretical' S0 vs. 'empirical' S1.1 or S1.2 null simulation designs. Theoretical justifications are provided in Section 3 of the Supplementary Materials.

In practice, permutation-based method must be carried out carefully, for example, in the presence of sample correlation (Abney 2015). Thus, the 'empirical' S1.1 design is perhaps earlier to implement than S1.2. For direct testing of the interaction effect β_{GE}, the different 'theoretical' null designs (i.e. S0.1 without vs. S0.2 with main G effect) did not lead to different T1E conclusions.

To conclude, although we only presented two examples (i.e. scale tests for variance heterogeneity and location tests of interaction effects), the findings here have important implications for future

evaluation of type 1 error control and interpretation. The newer test statistics being developed are increasingly complex, often going beyond the first moment such as the scale tests studied here, or beyond single variant approaches such as pathway and data integration analyses that have yet to be examined. Conventional simulation design S0 under the 'theoretical' null can lead to misleading conclusion regarding the accuracy of a test. The alternative simulation designs S1.1 and S1.2 under the 'empirical' null, on the other hand, can reveal the true behaviour of a test when applied to real data.

Acknowledgements

The authors have no conflict of interest to declare. The authors would like to thank Dr. David Soave and Dr. Jerry Lawless for helpful discussions. This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, 250053-2013), and the Canadian Institutes of Health Research (CIHR, 201309MOP-310732-G-CEAA-117978) and to LS.

References

[1] Mark Abney. "Permutation Testing in the Presence of Polygenic Variation". In: *Genetic Epidemiology* 39.4 (2015), pp. 249–258.

[2] Hugues Aschard et al. "A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes". In: *Genetic epidemiology* 37.4 (2013), pp. 323–333.

[3] Morton B Brown and Alan B Forsythe. "Robust tests for the equality of variances". In: *Journal of the American Statistical Association* 69.346 (1974), pp. 364–367.

[4] Ying Cao et al. "A versatile omnibus test for detecting mean and variance heterogeneity". In: *Genetic epidemiology* 38.1 (2014), pp. 51–59.

[5] B Devlin and K Roeder. "Genomic control for association studies." In: *Biometrics* 55.4 (1999), pp. 997–1004.

[6] Frank Dudbridge and Olivia Fletcher. "Gene-environment dependence creates spurious gene-environment interaction". In: *The American Journal of Human Genetics* 95.3 (2014), pp. 301–307.

[7] Frank Dudbridge and Arief Gusnanto. "Estimation of significance thresholds for genomewide association scans". In: *Genetic epidemiology* 32.3 (2008), pp. 227–234.

[8] Bradley Efron. "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis". In: *Journal of the American Statistical Association* 99.465 (2004), pp. 96–104.

[9] Friedhelm Eicker. "Limit theorems for regressions with unequal and dependent errors". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1. 1967, pp. 59–82.

[10] Liang Goh and Von Bing Yap. "Effects of normalization on quantitative traits in association test". In: *BMC bioinformatics* 10.1 (2009), p. 415.

[11] E Gómez-Sánchez-Manzano, MA Gómez-Villegas, and JM Marín. "Sequences of elliptical distributions and mixtures of normal distributions". In: *Journal of multivariate analysis* 97.2 (2006), pp. 295–310.

[12] Azmeri Khan and Glen D Rayner. "Robustness to non-normality of common tests for the many-sample location problem". In: *Journal of Applied Mathematics & Decision Sciences* 7.4 (2003), pp. 187–206.

[13] Julia Kozlitina and William R Schucany. "A robust distribution-free test for genetic association studies of quantitative traits". In: *Statistical applications in genetics and molecular biology* 14.5 (2015), pp. 443–464.

[14] Howard Levene et al. "Robust tests for equality of variances". In: *Contributions to probability and statistics* 1 (1960), pp. 278–292.

[15] Guillaume Paré et al. "On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study". In: *PLoS genetics* 6.6 (2010), e1000981.

[16] Tara J Rao and Michael A Province. "A Framework for Interpreting Type I Error Rates from a Product-Term Model of Interaction Applied to Quantitative Traits". In: *Genetic epidemiology* 40.2 (2016), pp. 144–153.

[17] David Soave and Lei Sun. "A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty". In: *Biometrics* (2017).

[18] David Soave et al. "A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways". In: *The American Journal of Human Genetics* 97.1 (2015), pp. 125–138.

[19] Maksim V Struchalin et al. "Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations". In: *BMC genetics* 11.1 (2010), p. 92.

[20] Xiangqing Sun et al. "What is the significance of difference in phenotypic variability across SNP genotypes?" In: *The American Journal of Human Genetics* 93.2 (2013), pp. 390–397.

[21] Tetsuji Tonda, Hirofumi Wakaki, et al. "Asymptotic expansion of the null distribution of the likelihood ratio statistic for testing the equality of variances in a nonnormal one-way {ANOVA} model". In: *Hiroshima Mathematical Journal* 33.1 (2003), pp. 113–126.

[22] Arend Voorman et al. "Behavior of QQ-plots and genomic control in studies of gene-environment interaction". In: *PloS one* 6.5 (2011), e19416.

[23] Andrew R Wood et al. "Another explanation for apparent epistasis". In: *Nature* 514.7520 (2014), E3.

[24] Hirokazu Yanagihara, Tetsuji Tonda, and Chieko Matsumoto. "The effects of nonnormality on asymptotic distributions of some likelihood ratio criteria for testing covariance structures under normal assumption". In: *Journal of Multivariate Analysis* 96.2 (2005), pp. 237–264.

Tables and Figures

Table 1: Summary of the three simulation designs, the 'theoretical' null S0, and the two 'empirical' null S1.1 and S1.2 for evaluating scale (or location) tests for *variance* (or mean) heterogeneity in phenotype Y across the three genotype G groups.

Design	Genotype	Phenotype (conditionally normal)	Null of No Association	Marginal Normality
S0: 'Theoretical' Null	G	$Y_0 \sim N(\mu, \sigma^2)$	$Y_0 \perp\!\!\!\perp G$	Y_0 normal
S1.1: 'Empirical' Null	G $G^{new} \perp\!\!\!\perp G$	$Y_1 \vert G \sim N(\mu, \sigma^2_G)$	$Y_1 \perp\!\!\!\perp G^{new}$	Y_1 not normal
S1.2: 'Empirical' Null	G	$Y_1 \vert G \sim N(\mu, \sigma^2_G)$ permute Y_1: Y_1^{perm}	$Y_1^{perm} \perp\!\!\!\perp G$	Y_1^{perm} not normal

Table 2: Summary of the two modelling approaches (Cao et al. 2014 and Aschard et al. 2013) used here to generate phenotypic variable Y_0 under the 'theoretical' null simulation designs S0, and Y_1 under the 'empirical' null simulation designs S1.1 and S1.2 as detailed in Table 1. Assume E was available for direct modelling and testing β_{GE}, the Aschard et al. (2013) model coincides with Model I of Rao and Province (2016), except E was $G_{non-repeating}$. T1E rate is first estimated from *nrep.in* simulation replicates in an inner loop (similar to one whole-genome scan), then averaged over *nrep.out* simulation replicates in an outer loop.

	Directly introduce variance heterogeneity by σ^2_G (Cao et al. 2014)	Indirectly introduce variance heterogeneity by $G \times E$ (Aschard et al. 2013) Or, Directly test β_{GE} (assuming E was available)
Null Model for S0	$Y_0 = e$, $e \sim N(0, \sigma^2)$	$Y_0 = \beta_E E + e$, $e \sim N(0, \sigma^2)$
Alternative Models for S1.1 and S1.2	$Y_1 = \beta_G G + e$, $e \sim N(0, \sigma^2_G)$	$Y_1 = \beta_G G + \beta_E E + \beta_{GE} G \times E + e$, $e \sim N(0, \sigma^2)$
Parameters	MAF=0.4 β_G=0.3 σ^2_0=0.23, σ^2_1=0.25, σ^2_2=0.29	MAF=0.4; $\mathbb{P}$(E=1)=0.3 β_G=0.01, β_E=0.3, β_{GE}=0.1, 0.2,$\cdots$,1 $\sigma^2 = 1$
Sample size	n=10^3 or 10^4	n=10^3 or 10^4
Nominal T1E	α=0.05	α=0.05, 0.01, 0.001, 10^{-5}
Replications	*nrep.in*=10^5 *nrep.out*=100	*nrep.in*=10^5, or 10^7 for α=10^{-5} *nrep.out*=100

Table 3: Empirical T1E rates of LRT_m and $Score_m$ location tests for mean difference in Y across the three G groups, and of LRT_v and *Levene* scale tests for variance difference in Y, based on the 'theoretical' null design of S0 and the alternative 'empirical' null designs of S1.1 and S1.2. Alternative empirical Y_1 data were generated using the *Aschard's genetic model* as described in Table 2. Empirical T1E rates outside $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ are bolded in red.

			$\alpha = 5 \times 10^{-2}$, $n = 10^3$					
		β_{GE}	0.0	0.2	0.4	0.6	0.8	1
Location	LRT_m	S0	5.029	5.027	5.026	5.027	5.027	5.026
		S1.1	5.039	5.021	5.021	5.019	5.014	5.010
		S1.2	4.997	5.023	5.022	5.020	5.017	5.011
	$Score_m$	S0	5.002	4.998	4.997	4.998	4.998	4.997
		S1.1	5.014	4.992	4.993	4.991	4.985	4.981
		S1.2	4.974	4.994	4.993	4.990	4.988	4.983
Scale	LRT_v	S0	5.035	5.083	5.081	5.081	5.081	5.079
		S1.1	5.029	5.188	**5.262**	**5.507**	**5.979**	**6.757**
		S1.2	5.031	5.198	**5.274**	**5.519**	**5.994**	**6.756**
	Levene	S0	4.956	4.898	4.898	4.898	4.898	4.896
		S1.1	4.989	4.901	4.904	4.905	4.909	4.907
		S1.2	4.906	4.922	4.912	4.911	4.915	4.908

Table 4: Empirical T1E rates of LRT_m and $Score_m$ location tests for mean difference in Y across the three G groups, and of LRT_v and *Levene* scale tests for variance difference in Y, based on the 'theoretical' null design of S0 and the alternative 'empirical' null designs of S1.1 and S1.2. Alternative empirical Y_1 data were generated using the *Aschard's genetic model* as described in Table 2, focusing on the extreme case of large interaction effect, $\beta_{GE} = 1$. Empirical T1E rates outside $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ are boded in red.

	$n = 10^3$	α	5×10^{-2}	1×10^{-2}	1×10^{-3}	1×10^{-5}
Location	LRT_m	S0	5.016	0.999	0.985	0.988
		S1.1	5.009	1.007	0.988	0.990
		S1.2	5.011	1.009	1.033	1.013
	$Score_m$	S0	5.008	0.998	0.989	0.990
		S1.1	4.982	0.998	0.982	0.991
		S1.2	4.982	0.999	0.983	0.997
Scale	LRT_v	S0	5.009	1.002	1.024	1.033
		S1.1	**6.923**	**1.636**	**2.059**	**11.599**
		S1.2	**6.920**	**1.639**	**2.042**	**11.624**
	Levene	S0	4.955	0.961	0.938	0.971
		S1.1	4.964	0.978	0.932	0.952
		S1.2	4.962	0.970	0.953	0.958

Table 5: Empirical T1E rates of LRT_m and $Score_m$ location tests for mean difference in Y across the three G groups, and of LRT_v and *Levene* scale tests for variance difference in Y, based on the 'theoretical' null design of S0 and the alternative 'empirical' null designs of S1.1 and S1.2. Alternative empirical Y_1 data were generated using the *Cao's genetic model* as described in Table 2, and using two difference sample sizes of $n = 10^3$ and 10^4. Empirical T1E rates outside $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ are bolded in red.

			$\alpha = 5 \times 10^{-2}$	
			$n = 10^3$	$n = 10^4$
Location	LRT_m	S0	5.011	5.012
		S1.1	4.992	5.003
		S1.2	4.934	4.989
	$Score_m$	S0	5.011	5.012
		S1.1	4.993	5.003
		S1.2	4.934	4.989
Scale	LRT_v	S0	5.103	5.165
		S1.1	**7.034**	**7.125**
		S1.2	**7.007**	**7.020**
	Levene	S0	4.924	4.945
		S1.1	4.905	4.965
		S1.2	4.874	4.825

Table 6: Empirical T1E rates of $LRT_{\beta_{GE}}$ and $Score_{\beta_{GE}}$ location tests of the interaction coefficient β_{GE} *for the* $G \times E$ *interaction term in a regression*, based on the 'theoretical' null design of S0 and the alternative 'empirical' null designs of S1.1 and S1.2. Alternative empirical Y_1 data were generated using the Aschard's genetic model as described in Table 2 when $\beta_{GE} = 1$, but E was assumed to be known in this case and direction interaction modelling was possible. Empirical T1E rates outside $\alpha \pm 3\sqrt{\alpha \times (1-\alpha)/nrep.in}$ are bolded in red.

		$n = 10^3$			$n = 10^4$		
	$\alpha =$	5×10^{-2}	1×10^{-2}	1×10^{-3}	5×10^{-2}	1×10^{-2}	1×10^{-3}
$LRT(\beta_{GE})$	S0	5.034	1.017	1.029	5.033	1.007	0.979
	S1.1	**7.091**	**1.763**	**2.422**	**6.886**	**1.709**	**2.410**
	S1.2	**7.100**	**1.771**	**2.389**	**6.972**	**1.707**	**2.339**
$Score(\beta_{GE})$	S0	4.982	1.003	1.002	5.021	1.004	0.972
	S1.1	**7.028**	**1.738**	**2.363**	**6.874**	**1.705**	**2.400**
	S1.2	**7.040**	**1.747**	**2.339**	**6.960**	**1.703**	**2.326**

Figure 1: Marginal (mixture) and conditional (stratified by G and a binary E) histograms of empirical phenotype data Y_1, based on the *Aschard's model* as described in Table 2 when $\beta_{GE} = 1$.

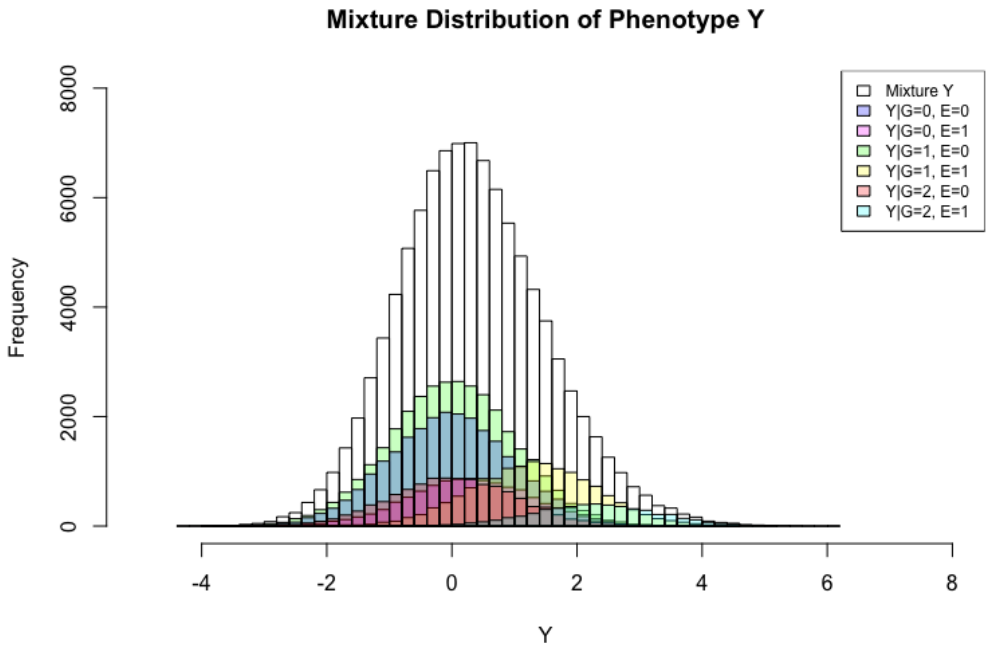

Figure 2: Comparison of the asymptotic distribution (black solid) and finite-sample distribution (red dashed) of LRT_v under the 'empirical' null, with χ^2_2 (blue dot-dashed) which is the asymptotic distribution of LRT_v under the 'theoretical' null. Vertical lines correspond the 99.9% quantile cutoffs for $\alpha = 0.001$.