Performance of MAX Test and Degree of Dominance Index in Predicting the Mode of Inheritance

Elias Zintzaras, University of Thessaly School of Medicine and Tufts University School of Medicine
Mauro Santos, Universitat Autònoma de Barcelona
Performance of MAX Test and Degree of Dominance Index in Predicting the Mode of Inheritance

Elias Zintzaras and Mauro Santos

Abstract

We evaluate power performance to detect the correct mode of inheritance in gene-disease associations of two different approaches: the MAX test and the degree of dominance index or h-index. The MAX test is a special case of the conditional independence tests that simultaneously test for association and select the most likely genetic model based on a three-dimensional normal distribution. The h-index is based on the philosophy of using orthogonal contrasts to infer the mode of inheritance quantitatively. A population genetic model is developed where the real mode of inheritance is known a priori and power performance can be accurately determined. The simulations showed that none of the two approaches generally outperforms the other, nor each of them provides a panacea to estimate efficiently the mode of inheritance in all parameter space. However, the simultaneous application of both approaches can provide insights in determining the underlying mode of inheritance.

KEYWORDS: genetic association, mode of inheritance, degree of dominance, h-index, MAX test

Author Notes: EZ is supported by the University of Thessaly Research Committee and the Biometrical, Epidemiological and Clinical Research Organization (BECRO). MS is supported by grants 2009SGR 636 from Generalitat de Catalunya to the Grup de Biologia Evolutiva and CGL2010-15395 from the Ministerio de Ciencia e Innovación (Spain), and by the ICREA Acadèmia program.
Address for correspondence: Elias Zintzaras, Department of Biomathematics, University of Thessaly School of Medicine, 2 Panepistimiou Str, Larissa 41110, Greece.
1 Introduction

The association between a disease and a bi-allelic gene with wild-type (wt) and mutant-type (mt) alleles in a case-control study is usually assessed using a chi-squared test for trend in a $3 \times 2$ contingency table with entries ‘genotypes’ (mtmt, wtmt and wtwt) and ‘disease status’ [Armitage (1955), Lewis (2010)]. When a significant association is revealed, the following genetic contrasts (or models) are examined by merging genotypes [Zintzaras and Lau (2008)]: (i) the additive contrast that compares genotypes mtmt vs. wtwt; (ii) the recessive contrast that compares mtmt vs. wtwt + wtmt; (iii) the dominant contrast that compares mtmt + wtmt vs. wtwt; and (iv) the co-dominant contrast that compares wtmt vs. mtmt + wtwt.

However, lack of prior knowledge of the real mode of inheritance hampers the biological justification for choosing a specific contrast and all possible contrasts are usually tested, making the interpretation of results complicated [Zintzaras and Lau (2008)]. This is because there are only 2 degrees of freedom for testing association in a $3 \times 2$ contingency table and the examined contrasts can not be interpreted independently. To estimate the unknown underlying genetic model, Hothorn and Hothorn (2009) proposed an extended MAX test, which is based on multiplicity-adjusted $P$-values for the Cochran-Armitage (CA) trend statistics [Armitage (1955)] of the dominant, additive and recessive models. The MAX test is a special case of the conditional independence tests [Freidlin et al. (2002), Hothorn and Hothorn (2009)] that simultaneously test for association and select the most likely genetic model based on a three-dimensional normal distribution.

We have recently proposed and alternative approach that uses the degree of dominance index $h$-index to infer the mode of inheritance in genetic association studies (GAS) in a continuous scale [Zintzaras and Santos (2011)]. The $h$-index is a function of two orthogonal genetic contrasts (the co-dominant and the additive contrasts) and its statistical significance is assessed from a logit model of the effects of the two contrasts. The performance of the $h$-index was investigated against simulated populations with an a priori defined mode of inheritance using principles of population genetics [Crow and Kimura (1970), Falconer and Mackay (1996)]. However, no analogous investigation has been carried out in the case of the MAX test [Hothorn and Hothorn (2009)]. Our aim here is to compare the MAX test as described by the Hothorn and Hothorn (2009) and the $h$-index in terms of power performance in detecting the correct mode of inheritance when it is actually known in advance.
2 MAX test

Consider a GAS with a bi-allelic polymorphism that evaluates the risk associated with allele $mt$. The genotype frequencies are given in a $3 \times 2$ contingency table with counts:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$mtmt$</td>
<td>$w_1$</td>
<td>$w_4$</td>
</tr>
<tr>
<td>$wtmt$</td>
<td>$w_2$</td>
<td>$w_5$</td>
</tr>
<tr>
<td>$wtwt$</td>
<td>$w_3$</td>
<td>$w_6$</td>
</tr>
</tbody>
</table>

To test for the association between case-control status $y$ and genotype $x$, the null hypothesis of independence $H_0: D(y/x) = D(y)$ can be tested against alternative genetic models ($x^{\text{model}}$) using the MAX test following Hothorn and Hothorn (2009). The MAX test is based on a three-dimensional test statistic $T$ (vectors are denoted in bold) where each dimension presents a specific genetic model: dominant, additive and recessive. The test statistic for the three genetic models is given by:

$$T = \begin{pmatrix} T_{\text{dom}} \\ T_{\text{add}} \\ T_{\text{rec}} \end{pmatrix} = \sum_{i=1}^{n} w_i y_i x_i,$$

where $n = 6$ (i.e., the six different genotype groups), $w_i$ is the frequency of the $i$th genotype group ($i=1$ to 6), $y_i$ is the case-control status ($y_i = 1$ for the cases and $y_i = 0$ for the controls), and $x_i = (x_{i}^{\text{dom}}, x_{i}^{\text{add}}, x_{i}^{\text{rec}})'$ is the corresponding scoring vector as follows: $x_{i}^{\text{dom}} = (0, 1, 1, 0, 1, 1)'$ for the dominant model (i.e., $x_{i}^{\text{dom}} = 0, x_{i}^{\text{dom}} = 1, x_{i}^{\text{dom}} = 1, x_{i}^{\text{dom}} = 0, x_{i}^{\text{dom}} = 1, x_{i}^{\text{dom}} = 1$), $x_{i}^{\text{add}} = (0, 1, 2, 0, 1, 2)'$ for the additive model, and $x_{i}^{\text{rec}} = (0, 0, 1, 0, 0, 1)'$ for the recessive model.

Under the null hypothesis the expectation and variance of $T$ are given by (see Appendix):
\[ E(T) = \frac{1}{n} \sum_{i=1}^{n} w_i y_i \sum_{i=1}^{n} w_i x_i, \]

\[ Var(T) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i) \right) \left( \sum_{i=1}^{n} w_i x_i \right) \]

\[ = -\frac{1}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i) \right) \left( \sum_{i=1}^{n} w_i x_i \right) \right). \]

Then, the distribution of \( T \) is asymptotically a three-dimensional normal.

Now, the MAX test is defined as:

\[ T_{\text{max}} = \max \left| \frac{T - E(T)}{\sqrt{\text{diag}(Var(T))}} \right|, \]

where the maximization is over the three models (dominant, additive and recessive) and \( \sqrt{\text{diag}(Var(T))} \) are the standard deviations of the elements of \( T \).

The significance of \( T_{\text{max}} \) is evaluated using three-dimensional normal probabilities. Thus, \( T_{\text{max}} \) provides a global testing of association and multiplicity adjusted \( P \)-values for each alternative genetic model, and points that the most likely genetic model is the one with the smallest \( P \)-value. Note that the multiplicity adjusted \( P \)-values are only for the number of genetic models tested and not for the number of loci tested when genome-wide association studies (GWAS) are considered.

### 3 The \( h \)-index

The \( h \)-index is based on the philosophy of analyzing a GAS using only orthogonal contrasts [Zintzaras and Santos (2011)]. Consider the GAS with genotype frequencies given in (1). Then, it can be analyzed using the following logit model:

\[ \log \left( \frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 \times g \]

where \( \pi \) is the probability of being diseased, the betas are constants that need to be estimated from the data themselves, and \( g \) is the genotype effect with 2 degrees of freedom (\( df_g = 2 \)). The deviance due to \( g \) is defined as

\[ D_g = -2 \ln \left( \frac{I_g}{I_f} \right) \] where \( I_g \) is the likelihood of the current model and \( I_f \) is the
likelihood of the saturated model. The deviance $D_g$ determines significance of $g$, and it is tested against the $\chi^2$ distribution with $df_g=2$.

Since the significance of $g$ (i.e., the association) is based on 2 df, there are two orthogonal contrasts with 1 df each, which can be interpreted separately: the co-dominant ($df_{co}=1$) and the additive ($df_a=1$) contrasts. The co-dominant contrast ($L_{co}$) is the contrast between the heterozygote and the average of the two homozygotes:

$$ L_{co} = \sum_i c_i^{(co)} g_i ; c^{(co)} = [-0.5 \ 1 \ -0.5] ; \sum_i c_i^{(co)} = 0. $$ (6)

The additive contrast ($L_a$) presents the contrast between the two extreme homozygotes:

$$ L_a = \sum_i c_i^{(a)} g_i ; c^{(a)} = [1 \ 0 \ -1] ; \sum_i c_i^{(a)} = 0. $$ (7)

These two contrasts can be interpreted separately since they are orthogonal: $[1 \ 0 \ -1][-0.5 \ 1 \ -0.5]=0$. Thus, testing the significance of the $g$ is equivalent to simultaneously testing the significance of the two contrasts.

Then, model (5) is equivalent to the following model:

$$ \log \left( \frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 \times L_{co} + \beta_1 \times L_a. $$ (8)

The significance of the co-dominant and additive contrasts are examined by testing the deviances $D_{co}$ and $D_a$ against the $\chi^2$ distribution with 1 df, respectively. Now, the orthogonality implies that $D_g = D_{co} + D_a$, where $D_{co}$ is the change in deviance due to adding in the null model the co-dominant effect, and $df_g = df_{co} + df_a$.

The additive and co-dominant contrasts are estimated by the natural logarithm ($\ln$) of the following odds ratios: $\theta_a = \left( w_1 \times w_6 \right) / \left( w_3 \times w_4 \right)$ and $\theta_{co} = \left[ w_2 \times \left( w_4 + w_6 \right) \right] / \left[ w_5 \times \left( w_1 + w_3 \right) \right]$, respectively. Thus, the significance of $\ln(\theta_a)$ and $\ln(\theta_{co})$ are tested using $D_a$ and $D_{co}$, respectively.

The degree of dominance is defined as the ratio of the natural logarithms ($\ln$) of the odds ratios ($\theta$) of the orthogonal contrasts:

$$ h = \frac{\ln(\theta_{co})}{\ln(\theta_a)}. $$ (9)

The sign of $\ln(\theta_{co})$ determines the direction of dominance, and the value of $\ln(\theta_{co})$ relative to the absolute value of $\ln(\theta_a)$ the magnitude of dominance deviation $h$, which can take any value from negative infinity to positive infinity.
The mode of inheritance is inferred from the $h$-index as follows (Figure 1) [Zintzaras and Santos (2011)]:

i. When the deviance $D_a$ of the additive contrast and the deviance $D_{co}$ of the co-dominant contrast are both non-significant, the association is not considered significant and there is no meaning to assess the value of $h$-index.

ii. When the deviance of the additive contrast $D_a$ is significant and the deviance of the co-dominant contrast is not significant, there is association and non-dominance (or additivity) is assumed ($h = 0$). This implies that the heterozygote $wmt$ “lies” in the middle of the two homozygotes, with the mutant homozygote $mmt$ having the maximum susceptibility of being diseased and the wild-type homozygote $wtt$ having the least.

iii. When the deviance of the co-dominant contrast $D_{co}$ is significant (irrespective of the significant level of the additive contrast), there is association and dominance is assumed ($h > 0$ or $h < 0$). When $h > 0$ the heterozygote $wmt$ lies towards $mmt$ (i.e., dominance of the allele $mt$) and, therefore, $wmt$ is expected to have a risk of being diseased somewhere in between the middle of the two homozygotes and towards $mmt$. When $h < 0$ the heterozygote $wmt$ lies towards $wtt$ (i.e., dominance of the allele $wt$) and the heterozygote $wmt$ is expected to have a risk of being diseased somewhere in between the middle of the two homozygotes and towards $wtt$.

Figure 1: Different scenarios of dominance based on the degree of dominance $h$-index.
4 Simulation model

Initially (at time \( t_0 \)) a cohort of \( N = 10,000 \) healthy subjects is assumed. This cohort is polymorphic for a target gene with two alleles, wild type-\( wt \) and mutant type-\( mt \), where the allele \( mt \) is associated with an increased probability of being diseased. The genotype frequencies at time \( t_0 \) are defined as follows:

\[
\begin{align*}
\text{Genotype} & \quad \text{Frequencies} \\
mtmt & \quad w_4 = q^2N \\
wtmt & \quad w_5 = 2pqN \\
wtwt & \quad w_6 = p^2N
\end{align*}
\]

(10)

where \( p \) is the frequency of allele \( wt \) and \( q = 1 - p \) is the frequency of allele \( mt \).

To explore how the different approaches employed to estimate the underlying mode of inheritance perform we can use the standard population genetics model that allows expressing a priori any degree of dominance associated to allele \( mt \) [Crow and Kimura (1970), Falconer and Mackay (1996)]. This model assumes that the genotype dependent per capita probability of being disease relative to the wild type genotype \( wtwt \) is \( s (s > 0) \) for the mutant homozygote \( mtmt \) and \( Hs \) for the heterozygote \( wtmt \), where \( H \) (\( 0 \leq H \leq 1 \)) is the degree of dominance. Thus, if \( H = 0 \) the allele \( mt \) is fully recessive, if \( H = 1 \) \( mt \) is fully dominant, and non-dominance (additiveness or co-dominance) results when \( H = 0.5 \). Then, from (10) the genotype distribution of diseased subjects (cases) at time \( t_1 \) \( (t_1 > t_0) \) is as follows:

\[
\begin{align*}
\text{Genotype} & \quad \text{Cases} \\
mtmt & \quad w_1 = q^2N \times (s + P_d) \\
wtmt & \quad w_2 = 2pqN \times (Hs + P_d) \\
wtwt & \quad w_3 = p^2N \times P_d
\end{align*}
\]

(11)

where \( P_d \) is the per capita genotype-independent probability of being disease.

Therefore, we can define parametric controls and cases for any model of inheritance using (10) and (11). Here we simulated studies with \( 0.05 \leq s \leq 0.10 \) and the following models of inheritance: (i) dominance, \( H \) takes values from a uniform distribution in the range \( 0.75 \leq H \leq 1 \); (ii) additiveness, \( 0.375 \leq H \leq 0.625 \); and (iii) recessiveness, \( 0 \leq H \leq 0.25 \).

To simulate a “sampling” case-control study, \( n = 1,000/1,000 \) cases/controls subjects were sampled from the parametric cases in (11) and parametric controls in (10) based on the spaces defined by the cumulative
genotype frequencies; e.g., if a control subject is randomly sampled in the space from \( w_5/(w_4 + w_5 + w_6) \) to \( (w_5 + w_6)/(w_4 + w_5 + w_6) \) then it is assigned as heterozygous \( \text{wtmt} \). The same procedure was applied to randomly sampling case subjects. The simulations were performed using VCF90 with the IMSL library.

5 Results

We first assessed type I error by simply setting \( s = 0 \) in the different models of inheritance analyzed (dominance, additiveness and recessiveness). Type I error rates were close to their nominal value 0.05 although generally in some excess for the MAX test. For instance, when the frequency of the mutant allele is \( q = 0.05 \) type I error rates for the MAX test were 0.0596, 0.0601 and 0.0612 for the dominance, additiveness and recessiveness situations, respectively. The type I error rates for the \( h \)-index in these situations were, respectively, 0.0508, 0.0538 and 0.0518. (Note: henceforth \( q \) refers to the frequency of the mutant allele in the controls, which increases in the case subjects as a function of \( H \) and \( s \) according to (11)).

For each model of inheritance power was estimated as:

\[
\text{Power} = \frac{\# \text{ simulations that detected the correct model of inheritance}}{\# \text{ total simulations}}.
\] (12)

Figure 2 shows a summary of power results for different allele frequencies \( q \) (each point is based on 10,000 simulations). We now discuss these results according to the true mode of inheritance. In all situations we assumed that \( q \) ranged from 0.05 to 0.30 [Manolio (2010), Dickson et Al. (2010)] in steps of 0.05, \( P_d = 0.20 \), and two values of \( s \): \( s = 0.05 \) and \( s = 0.10 \).

Dominance \((0.75 \leq H \leq 1)\). For \( s = 0.05 \) the power of \( h \)-index to detect the correct model of inheritance decreased monotonically from 98% to 68% as the allele frequency \( q \) increased, whereas the power of MAX test remained fairly constant (ranging from 54% to 63%) with increasing \( q \). When \( s = 0.10 \) power was improved in both instances at similar rate (Figure 2).

Additiveness \((0.375 \leq H \leq 0.625)\). When \( s = 0.05 \) the power of the MAX test remains constant (about 37%) and it is higher than the power of \( h \)-index when \( q < 0.15 \); however, the power of \( h \)-index monotonically increases with \( q \) and reaches the value of 61% when \( q = 0.30 \). On the other hand, when \( s = 0.10 \) power for the MAX test is generally higher than that for the \( h \)-index except for \( q = 0.30 \) (Figure 2).
Figure 2: Power results for different frequencies of \( mt (q) \), for \( s = 0.05 \) or \( 0.10 \) and real mode of inheritance \((H)\) as dominant, additive, recessive.
Recessiveness \((0 \leq H \leq 0.25)\). The situation in terms of power is in favor of the MAX test since it performs better than \(h\)-index and reaches power around 60% when \(q \geq 0.15\).

To understand the behavior of \(T_{\text{max}}\) we expressed \(w_i\) in terms of the parameters of the population genetics model (i.e., \(p, q, P_d, H\) and \(s\)) using the formulas (10) and (11). Then, \(T_{\text{dom}}, T_{\text{add}}\) and \(T_{\text{rec}}\) were plotted against \(H\) and \(q\) assuming \(s = 0.05\) (Figure 3). \(T_{\text{dom}}\) and \(T_{\text{add}}\) values produced a similar and largely overlapping pattern, especially for larger values of \(H\). This explains the weakness of \(T_{\text{max}}\) to detect dominance or non-dominance sufficiently. Conversely, \(T_{\text{rec}}\) gives the highest values as \(H\) decreases, which explains the good performance of the MAX test when the underlying mode of inheritance is recessiveness.

It is also easy to explain why power for the \(h\)-index is maximum with dominance and decreases when the mutant allele tends towards the additiveness or recessiveness situations. We can write the additive and co-dominant contrasts on which the \(h\)-index is based as a function of the parametric frequencies in (10) and (11) as follows:

\[
\begin{align*}
\theta_a &= 1 + \frac{s}{P_d}, \\
\theta_{co} &= \frac{P_d \left( p^2 + q^2 \right) + Hs \left( p^2 + q^2 \right)}{P_d \left( p^2 + q^2 \right) + q^2 s}.
\end{align*}
\]

From (13) it is clear that the additive contrast \(\theta_a > 1\) whenever \(s > 0\) (i.e., the additive contrast will be statistically significant assuming there is enough power), but the co-dominant contrast is a more complicated function of the underlying degree of dominance \(H\) and the per capita probability of being disease \(s\) for the mutant homozygote \(mmt\). This function is plotted in Figure 4 assuming \(s = 0.05\). Because the degree of dominance \(h\) is essentially inferred from the statistical significance of the co-dominant contrast (see above), a visual inspection of Figure 4 clarifies inferences of the underlying model of inheritance based on the \(h\)-index. Thus, the value for the co-dominant contrast is always higher than one when \(H \geq 0.75\) (i.e., when the underlying model of inheritance of allele \(mt\) tends to dominance): statistical significance of \(\theta_{co}\) will be detected and the real model of inheritance correctly inferred. When \(H \approx 0.5\) (i.e., the underlying model of inheritance is non-dominance or additiveness) there are situations where \(\theta_{co} \neq 1\); in these circumstances statistical significance of the co-dominant contrast can be detected and additivity will be wrongly rejected. Finally,
when \( H \leq 0.25 \) (i.e., when allele \( mt \) tends to be recessive) there is a range of the parameter space where \( \theta_{co} \approx 1 \) and recessivity will be missed.

Figure 3: The three dimensions of \( T_{max} \) (\( T_{dom} \), \( T_{add} \) and \( T_{rec} \)) as a function of the real mode of inheritance (\( H \)) and the frequency of \( mt \) (\( q \)).
Figure 4: Effect size of co-dominant contrast as a function of the real mode of inheritance (H) and the frequency of mt (q).
6 Discussion and conclusion

Hereby, we performed a comparative study to explore the power performance of two recently proposed approaches for identifying the genetic mode of inheritance: the MAX test [Hothorn and Hothorn (2009)] and the $h$-index [Zintzaras and Santos (2011)]. The philosophy of the two approaches is quite different: the MAX test predicts the genetic model choosing from a set of three pre-specified models (dominant, additive and recessive) whereas the $h$-index determines the genetic model quantitatively based on the estimated degree of dominance; that is, the magnitude of the deviation of the heterozygote form the middle of the two homozygotes. Though the two approaches are not directly equivalent we have made an effort to compare their ability to detect the true mode of inheritance in the presence of association.

The novelty of the present work is that a population genetic model was developed where the mode of inheritance is known *a priori*. The model uses the concept of dominance of a mutant allele based on the risk of being disease for the homozygote $mtmt$ relative to wild-type homozygote $wtwt$. Parametric controls and cases with a known mode of inheritance can be easily derived from this model. Note that this is the first study to examine the power performance of MAX test in a setting where the true underlying genetic model is known *a priori*.

The simulations showed that none of the two approaches consistently outperforms the other nor each of them provides a panacea to estimate efficiently the mode of inheritance through all parameter space. If the true mode of inheritance is recessiveness (i.e., the heterozygote $wtmt$ has a risk of being disease closer to the wild-type homozygote $wtwt$) MAX test is generally far better than the $h$-index. Conversely, if the true mode of inheritance is dominance ($wtmt$ has a risk of being disease closer to the mutant homozygote $mtmt$) the $h$-index outperforms MAX test. The real hurdle is when the true mode of inheritance is additiveness ($H \approx 0.5$); the MAX test is better than the $h$-index at relatively low frequencies of allele $mt$, while the reverse is true at relatively high frequencies of this allele.

Since the $h$-index quantifies the mode of inheritance in terms of degree of dominance and the MAX test categorizes the mode of inheritance qualitatively into one of the three genetic models, the two approaches may be combined to provide a better estimate of the underlying mode of inheritance. The MAX test may be applied first to reveal one of the three genetic models, and then the $h$-index may provide an estimate of the degree of dominance of the mutant allele subjected, of course, to the condition that both approaches detect the same direction of dominance.

Most of the test statistics used to test for association, such the CA trend test or the Pearson $\chi^2$-test, assume a specific genetic model such as dominant,
additive and recessive [Zheng et al. (2009)]. However, the false specification of the genetic model may lead to loss of power [Gonzalez et al. (2008)]. Also, association may be tested by performing multiple tests for every genetic model and then to adjust for multiple testing [Ziegler and König (2010)] and to choose the most significant test [Gonzalez et al. (2008)]. A more efficient method is the MAX test [Hothorn and Hothorn (2009)] where it tests simultaneously for association and selects the most likely model. However, the MAX test is a special case of conditional independence tests [Hothorn and Hothorn (2009)] and thus, the distribution is approximated using a three-dimensional normal distribution where each dimension corresponds to a test for each genetic model. Then, multiplicity adjusted $P$-values for each test are used to indicate the most likely genetic model. Although simulation studies have shown that MAX test performs well in testing association when the genetic model is assumed to be dominant, additive or recessive [Gonzalez et al. (2008), Li et al. (2009), Zheng et al. (2009)], the power performance of MAX test was never really examined against a population (consisted of cases and diseased) when the real mode of inheritance is known. Thus, the genetic model is typically selected based on tests performing for these specific models.

In conclusion, the two approaches perform quite differently according to the degree of dominance, selection coefficient and frequency of the mutant allele. However, the application of both tests may provide insights in determining the underlying mode of inheritance.

**Appendix**

Let define the test statistic $T$ as $T = \sum_{i=1}^{n} w_i y_i x_i$. Then, the three-dimensional expectation of $T$ under the null hypothesis of independence is given by

$$E(T) = E(y) \sum_{i=1}^{n} w_i x_i,$$

where $E(y) = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$. Thus,

$$E(T) = \frac{1}{n} \sum_{i=1}^{n} w_i y_i \sum_{i=1}^{n} w_i x_i.$$

The corresponding covariance matrix of $T$ is given by

$$Var(T) = \frac{n}{n-1} Var(y) \left( \sum_{i=1}^{n} w_i x_i \right)^2 - \frac{1}{n-1} var(y) \left( \sum_{i=1}^{n} w_i x_i \right) \left( \sum_{i=1}^{n} w_i x_i \right).$$
where \( \text{Var}(y) = \left( \frac{1}{n} \sum_{i=1}^{n} w_i(y_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i) \right) \). Thus,

\[
\text{Var}(T) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} w_i(y_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i) \right) \left( \sum_{i=1}^{n} w_i x_i \right)^2
\]

\[
- \frac{1}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} w_i(y_i - \frac{1}{n} \sum_{i=1}^{n} w_i y_i) \right) \left( \sum_{i=1}^{n} w_i x_i \right) \left( \sum_{i=1}^{n} w_i x_i \right).
\]

References


