Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 15 Issue 4 Version 1.0 Year 2015
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Direct Estimation of Effects and Tests of their Statistical Significance for the Case of One Autosomal Locus with Two Alleles

By Charles J. Mode
Drexel University, United States

Abstract- In previous papers, see the references, the author introduced methods for estimating effects directly in samples of individuals whose genomes had be sequenced for the cases of one and two or more quantitative traits. In these papers, no attention was given to developing procedures of testing the statistical significance of the estimated effects. This paper is devoted to the development of statistical tests of significance of estimated effects for the simple case of one autosomal locus with two alleles, using Monte Carlo simulation methods. Because no real data was available to the author, artificial data for the three genotypes was simulated by using a Monte Carlo simulation procedures with fixed sample size for each genotypes as well as expectations and variances. In all cases considered, the null hypothesis was described in detail so as to inform a reader on the basic concepts underlying the proposed tests of statistical significance.

Keywords: estimating effects, estimated heritability, tests of statistical significance, definitions of null hypotheses, Monte Carlo simulations methods, simulated data, absolute normal distribution.

GJSFR-F Classification : FOR Code : MSC 2010: 37A60

Strictly as per the compliance and regulations of :

© 2015. Charles J. Mode. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ref

3. Mode, C. J. (2014-a) Estimating Effects and Variance Components in Models of Quantitative Genetics in an Era of Sequenced Genomes. Global Journal of Science Frontier Research-Mathematics and Decisions Sciences. Issue 5 Version 1.0 Online ISSN 2249-4626.

Direct Estimation of Effects and Tests of their Statistical Significance for the Case of One Autosomal Locus with Two Alleles

Charles J. Mode

Abstract- In previous papers, see the references, the author introduced methods for estimating effects directly in samples of individuals whose genomes had be sequenced for the cases of one and two or more quantitative traits. In these papers, no attention was given to developing procedures of testing the statistical significance of the estimated effects. This paper is devoted to the development of statistical tests of significance of estimated effects for the simple case of one autosomal locus with two alleles, using Monte Carlo simulation methods. Because no real data was available to the author, artificial data for the three genotypes was simulated by using a Monte Carlo simulation procedures with fixed sample size for each genotypes as well as expectations and variances. In all cases considered, the null hypothesis was described in detail so as to inform a reader on the basic concepts underlying the proposed tests of statistical significance. For class of statistical tests described in this paper, two types of p-values may be distinguished. One type of p-values consists of those for each of the estimated effects. A second type of p-values consist concerns the joint statistical significance of two or more estimated effects. The consideration of the simple case of two autosomal loci is useful, because it provides insights into how the Monte Carlo simulation procedures used in this paper may be extended to cases of two or more autosomal loci with two or more alleles at each locus.

Keywords: estimating effects, estimated heritability, tests of statistical significance, definitions of null hypotheses, monte carlo simulations methods, simulated data, absolute normal distribution.

I. Introduction

As was suggested in previous papers, when an investigator is dealing with a sample of individuals whose genomes have been sequenced and a set of regions of the genome have been identified that affect the expression of a quantitative trait or traits, then it becomes possible to provide a working definition of set of loci in each individual at the genomic level, see Mode [3] and [4]. Moreover, if it is also possible to use makers in the DNA of each individual to provide working definitions of at least two alleles at each locus, then an investigator can develop a concrete working definition of the set of loci with two alleles at each that have shown to have an effect on the expression of a quantitative trait of traits as expressed in a numerical measurement or measurements on each individual in the sample.

Author: Professor Emeritus, Department of Mathematics, Drexel University, Philadelphia, PA. e-mail: cjmode@comcast.net

© 2015 Global Journals Inc. (US)

In principle, a quantitative trait or traits may be analyzed statistically for any combination of the set of loci under consideration, but because the number of genotypes that can be identified increase at a fast rate as the number of loci under consideration increase, the sample of individuals may not be sufficiently large to assure that the number of individuals of each genotype is large enough to obtain statistically significant results as discussed in chapters 3 and 4. Such a situation will usually arise whenever the number of loci under consideration is greater than 5 or 6. Therefore, if an investigator has 6 loci or more loci under consideration in a sample of data, it would be prudent to perform a preliminary analysis of the data for each locus under consideration in order to develop an understanding as to which combination of loci would be most informative and fruitful to explore.

To execute such an experiment, an investigator would need software to estimate each effect whose square is a component of the genetic variance for each locus under consideration. In the papers presented in chapters 3 and 4, it was assumed that any allele in a genotype could be identified as to whether it was contributed by the father or mother of any individual in the sample. In many data sets, however, this assumption is not valid so that an investigator could identify only three genotypes per locus for the case of two alleles per locus. These three genotypes consist of two homozygotes and a heterozygote for which it was not possible to identify whether each allele was of maternal or paternal in origin. The purpose of this chapter is to provide an overview of the software necessary to carry out a preliminary exploration of a data set, such that the genome of each individual in the set has been sequenced, for each locus under consideration. The main focus of this chapter is to implement software to do the necessary computations for the simplest case of one autosomal locus with two alleles with a view towards extending the software to cases of two or more autosomal loci. A mathematical description of this software is provided for those investigators who write code using a programing language based on the manipulation of arrays such as APL or MATLAB.

A Monte Carlo simulation procedure was used to provide data for illustrating how the software may be used to analyze real data when it is available. Contained in this software are procedures of estimating the squares of effects and a description of a procedures to formulate null hypotheses to test the statistical significance of the estimated squares of effects. After a null hypothesis is defined, a Monte Carlo simulation procedure was used to estimate p-values to judge whether each estimated square of an effect was statistically significant, given some null hypothesis. Detailed technical descriptions of null hypotheses that were used in tests of statistical significance are also provided for each reported experiment.

II. Non-Identifiabilty of Heterozygotes in the Case of one Autosomal Locus

In previous work on defining effects in connection with components of variance models in quantitative genetics, it was assumed that two kinds of heterozygotes could be identified when working with a sample of individuals whose genomes have been sequenced. For example, let the symbol 1 denote the presence of a marker on a haplotype with respect to some locus, and let the symbol 0 denote the absence of this marker. Then, if it is assumed that it is possible to detect in each individual whether each of the two alleles in a genotype that was contributed by the maternal or paternal parent, then a genotype could to represented by the symbol (x, y), where x and y denote, respectively, the maternal and paternal allele. Thus, if both x and y are assigned the symbols 1 or 0, then four genotypes in the set

$$\mathbb{G}_1 = \{(1,1),(1,0),(0,1),(0,0)\} \tag{2.1}$$

© 2015 Global Journals Inc. (US)

can be identified in a sample of individuals. But, if it is not possible to identify the maternal and paternal alleles in an individual, then the two possible heterozygotes, $(1,0)$ and $(0,1)$ would be lumped into a single category called heterozygotes.

But, in such samples, the two homozygotes, $(1,1)$ and $(0,0)$, could be identified unambiguously. In what follows, the genotypes in the three categories will be denoted by the symbols $(1,1),(x \neq y)$ and $(0,0)$, where the symbol $(x \neq y)$ stands for heterozygotes in the non-identifiable case. In this case,

$$\mathbb{G}_2 = \{(1,1),(x \neq y),(0,0)\} = \{g_1, g_2, g_3\} \tag{2.2}$$

is the set of genotypes of recognizable genotypes. To simplify the notation in this case, the symbols g_1, g_2, g_3 will stand for three genotypes under consideration as indicated in (2.2)

For the case of non-identifible heterozygotes, let $n > 1$ denote the number of individuals in a sample, and let $n(1,1)$, $n(x \neq y)$ and $n(0,0)$ denote, respectively, the number of individuals of each of the three genotypes. To simply the notation, the numbers of each of these three genotypes is indicated in the set

$$\{n(g_1), n(g_2), n(g_3)\} \tag{2.3}$$

denote the set of genotypes and let g, with or without subscripts, denote an element in the set $\mathbb{G}_2$ as indicated on the right in (2.2). Then,

$$n = \sum\nolimits_{g \in \mathbb{G}_2} n(g) \ , \tag{2.4}$$

is the number of individuals in the sample, and

$$p_\nu = \frac{n(g_\nu)}{n} \tag{2.5}$$

is the estimated frequency of genotype g_ν in the population for $\nu = 1,2,3$. This collection of estimates will be referred to as the estimated genotypic distribution.

To take into account the problem setting up a structure that incorporates phenotypic variation among individuals of the same genotype in sample with respect to some quantitative trait, let W denote and random variable taking values in the set of $\mathbb{R}_W$ of possible values of the phenotype. Usually, $\mathbb{R}_W$ is a set of rational real numbers. Given genotype $g \in \mathbb{G}_2$, let $f(w \mid g)$ denote the conditional probability density function of the random phenotypic variable W, and suppose there are $n(g)$ observed realizations of the random variable W denoted by the symbols W_i for $i = 1,2,\cdots,n(g)$. The conditional expectation of the random variable W, given the genotype $g \in \mathbb{G}$, is

$$E[W \mid g] = \int_{\mathbb{R}_W} f(w \mid g)\, dw \tag{2.6}$$

for every genotype $g \in \mathbb{G}_2$.

Therefore, an estimator of the conditional expectation $E[W \mid g]$ is

$$\mu(g) = \frac{1}{n(g)} \sum\nolimits_{i=1}^{n(g)} W_i \tag{2.7}$$

for all $g \in \mathbb{G}$. In what follows, these estimates will be referred to as the genetic values. Let μ denote the unconditional expectation of the genetic values. Then,

$$\mu = \sum\nolimits_{g \in \mathbb{G}_2} p(g)\, \mu(g) \tag{2.8}$$

© 2015 Global Journals Inc. (US)

Observe that this formula of this type would also be valid if the set $\mathbb{G}_1$ of genotypes were under consideration, in this case there would be 4 genotypes to take into account.

One objectives of this chapter is to suggest ways of write software to implement the formulas above, using a array manipulating programing language. Included in the class of array manipulating programing languages are APL and MATLAB. For such languages, a good starting point is to represent the genotypic distribution in (2.5) and the set of estimated expectations in (2.7) in matrix forms. To cast the genotypic distribution in matrix form, suppose the set of genotypes under consideration is $\mathbb{G}_2$. Then, the matrix of estimated of genetic values will have the form

$$\boldsymbol{\mu}_{\mathbb{G}_2} = \begin{bmatrix} \mu(g_1) & \mu(g_2) \\ 0 & \mu(g_3) \end{bmatrix} . \tag{2.9}$$

If both of the genotypes $(1,0)$ and $(0,1)$, i.e., both maternal and paternal alleles in a sample of individuals can be identified, then the set of genotypes under consideration would be the set $\mathbb{G}_1$ as defined above. In this case, the matrix of genetic values has the form

$$\boldsymbol{\mu}_{\mathbb{G}_1} = \begin{bmatrix} \mu(1,1) & \mu(1,0) \\ \mu(0,1) & \mu(0,0) \end{bmatrix} = [\mu(i,j)] \ , \tag{2.10}$$

where in this case $\mu(0,1\ )$ may be positive.

Let the array

$$\boldsymbol{p}_{\mathbb{G}_1} = \begin{bmatrix} p(1,1) & p(1,0) \\ p(0,1) & p(0,0) \end{bmatrix} = [p(i,j)] \tag{2.11}$$

represent the matrix form of the genotypic distribution for case 1. For the case the set of genotypes $\mathbb{G}_1$ is under consideration, let a matrix operation of element by element multiplication be denoted by

$$\boldsymbol{p} \times \mu = [p(i,j)\,\mu(i,j)] . \tag{2.12}$$

The symbol on the left in (2.12) would be the format for doing the operation of matrix of element by element implication in APL. Then, for case 1, the unconditional expectation of the matrix of genetic values may be written in the form

$$\mu_{\mathbb{G}_1} = \sum_{(i,j)\in\mathbb{G}_1} [p(i,j)\,\mu(i,j)] \ . \tag{2.13}$$

It is important to observe that if the set $\mathbb{G}_2$ of genotypes were under consideration, then formula of type (2.13) could also be used to compute unconditional expectation $\mu_{\mathbb{G}_2}$. From the point of view of writing software with an array processing programming language, it is important to observe that the same program could be used for these cases, except case 2, were the matrix of expected genetic values would have the form in (2.9) and the matrix form of the genotypic distribution would be represented in the form

$$\boldsymbol{p}_{\mathbb{G}_2} = \begin{bmatrix} p(g_1) & p(g_2) \\ 0 & p(g_3) \end{bmatrix} . \tag{2.14}$$

In what follows in this section, a general notation will be used to partition the phenotypic variance of a trait into the genetic and environmental variances, using a general notation that includes cases 1 and 2 described above. Let $\mathbb{A}$ denote the set of alleles at some autosomal locus, let (x,y) any genotype, where $x \in \mathbb{A}$ and $y \in \mathbb{A}$ and let

$$\mathbb{G} = \{(x,y) \mid x \in \mathbb{A} \text{ and } y \in \mathbb{A}\} \tag{2.15}$$

Notes

© 2015 Global Journals Inc. (US)

be the set of genotypes under consideration. Given genotype $(x,y) \in \mathbb{G}$, a set of genotypes under consideration, let $W(x,y)$ denote a realization of the phenotypic random variable W, given (x,y). Then observe that the equation

$$W(x,y) - \mu = (\mu(x,y) - \mu) + (W(x,y) - \mu(x,y)) \tag{2.16}$$

is valid for all genotypes $(x,y) \in \mathbb{G}$.

The phenotypic variance, genetic and environmental variances are defined as

$$var_P[W] = \sum_{(x,y)} p(x,y) (W(x,y) - \mu)^2 \tag{2.17}$$

$$var_G[W] = \sum_{(x,y)} p(x,y) (\mu(x,y) - \mu)^2 \tag{2.18}$$

and

$$var_E[W] = \sum_{(x,y)} p(x,y) (W(x,y) - \mu(x,y))^2 \tag{2.19}$$

respectively.

By using a conditioning argument given (x,y) and (2.16), it can be shown that the equation

$$var_P[W] = var_G[W] + var_E[W] \tag{2.20}$$

is valid. If a reader is interested in a detailed derivation of the formula in (2.20), consult chapter 3, which contains a detailed account for the case of one quantitative trait. Observe that if the variances on right of equation (2.20) are estimates based on data, then the sum on the right in (2.20) is an estimator of $var_P[W]$, the phenotypic variance. By definition, H, the heritability of a trait, is

$$H = \frac{var_G[W]}{var_G[W] + var_E[W]} = \frac{var_G[W]}{var_P[W]} . \tag{2.21}$$

From this formula, it can be seen that to estimate H, it suffices to compute $var_G[W]$ and $var_E[W]$, and then use equation (2.20) to check the validity of formula (2.20) numerically. Note that any estimate of H will satisfy the condition $0 \leq H \leq 1$.It should also be observed that from formula (2.17) that the phenotypic variance $var_P[W]$ may also be computed directly.

The formulas just derived are theoretical and when a sample of data is available to an investigator, the genotypic distribution may be estimated as well as means and variances. For example, for each genotype $(i,j) \in \mathbb{G}_2$, let $W_\nu(i,j)$ for $\nu = 1, 2 \cdot \cdot, n(i,j)$ be a set of quantitative observations on some trait. Then, for every genotype $(i,j) \in \mathbb{G}_2$,

$$\widehat{\mu}(i,j) = \frac{1}{n(i,j)} \sum_{\nu=1}^{n(i,i)} W(i,j) \tag{2.22}$$

is an estimate of the expectation $\mu(i,j)$, and an estimate of $\sigma^2(i,j)$ is

$$\widehat{\sigma^2}(i,j) = \frac{1}{n(i,j) - 1} \sum_{\nu=1}^{n(i,j)} (W(i,j) - \widehat{\mu}(i,j))^2 \tag{2.23}$$

is an estimate of $\sigma^2(i,j)$ for $n(i,j) > 1$, and

$$\widehat{p}(i,j) = \frac{n(i,j)}{n} \tag{2.24}$$

is an estimate of the genotypic distribution. In the sections that follow the symbol $\widehat{\circ}$ will be dropped to lighten the notation, but it will be tacitly understood that when a set of data is under consideration, the symbols $\mu(i,j), \sigma^2(i,j)$ and $p(i,j)$ are estimates of parameters.

© 2015 Global Journals Inc. (US)

III. Estimating Effects Directly for Case of One Autosomal Locus

In classical quantitative genetics, the variances in variance component models were usually estimated indirectly by using analysis of variance-covariance procedures, but within this framework it was not possible to estimate the effects, because they were squared terms in the weighted sums that were, by definition, variance components. See chapter 1 for a specific example of an analysis of variance-covarince procedure applied to data. But, as will be shown in this section, when the genotype of each individual in a sample is known at the DNA level, it is possible to estimate effects directly from the data for cases of at least two alleles at the locus as was suggested in chapters 3 and 4.

Suppose that the maternal and paternal alleles are not identified in a sample, and let $n(1,1)$, $n(0,0)$ and $n(het)$ denote, respectively, the total number of individuals that were homozygous for the alleles 1 or 0 and heterozygous for these alleles. Then, the total number of individuals in the sample is

$$n = n(1,1) + n(0,0) + n(het) \ . \tag{3.1}$$

Let $p(1,1,)\,, p(0,0)$ and $p(het)$ denote, respectively, the frequencies of the three genotypes under consideration. Then, $p(1,1) = n(1,1)/n$, $p(0.0) = n(0,0)/n$ and $p(het) = n(het)/n.$ are estimators of these frequencies. It is also essential to have estimates of the frequencies of alleles in the sample. Let $p(1)$ and $p(0)$ denote the estimated frequencies of alleles 1 and 0 in the sample. The number of copies of allele 1 in individuals of genotype $(1,1)$ is 2, and the number of copies of this allele in each heterozygote is one. Therefore, the estimated frequency of allele 1 in the sample is

$$p(1) = \frac{2n(1,1) + n(het)}{2n} = p(1,1) + \frac{1}{2}p(het) \ . \tag{3.2}$$

Similarly, the estimate of allele 0 in the sample is

$$p(0) = p(0,0) + \frac{1}{2}p(het) \tag{3.3}$$

Observe that

$$p(1) + p(0) = 1 \tag{3.4}$$

as they should.

In order to define all effects for the case of one autosomal locus, it will be necessary to define conditional expectations of the estimated means defined in equation (2.7) in section 2 with respect to the genotypic distribution defined below. When programming in an array processing language, it is convenient to represent the data and estimates in a matrix form. Let $p(1,0) = p(0,1) = 0.5p(het)$. Then, the matrix form of the genotypic distribution is

$$\boldsymbol{p}_{\mathbb{G}_2} = \begin{bmatrix} p(1,1) & p(1,0) \\ p(0,1) & p(0,0) \end{bmatrix} \ . \tag{3.5}$$

The subscript $\mathbb{G}_2$ denotes the set of genotypes defined in (2.2) in section 2. Given this matrix, the frequency of allele 1 has the form

$$p(1) = p(1,1) + p(1,0) \ . \tag{3.6}$$

Similarly, the frequency of allele 0 has the form

$$p(0) = p(0,1) + p(0,0) \tag{3.7}$$

From the point of view of writing computer code in an array processing language, one could use a single command to compute the sum of the rows of the matrix $\boldsymbol{p}_{\mathbb{G}_2}$ that would result in an array with the elements $p(1)$ and $p(0)$.

© 2015 Global Journals Inc. (US)

To expedite the writing of code in an array processing programing language, let $\mu(1,1), \mu(0,0)$ and $\mu(het)$ denote the genetic means for three genotypes under consideration, and, to cast these means in matrix form as in (3.5), let $\mu(0,1) = \mu(1,0) = \mu(het)$. Then, the matrix of these means may be represented in the form

$$\boldsymbol{\mu}_{\mathbb{G}_2} = \begin{bmatrix} \mu(1,1) & \mu(1,0) \\ \mu(0,1) & \mu(0,0) \end{bmatrix} . \tag{3.8}$$

An essential step in defining the effects in what follows is to estimate the mean μ defined in (2.8) of section 2. To this end consider the matrix product

$$\boldsymbol{p}_{\mathbb{G}_2} \times \boldsymbol{\mu}_{\mathbb{G}_2} , \tag{3.9}$$

where the symbol $\times$ stands for element by element multiplication. Then, in this notation, the mean μ has the form

$$\mu = \sum \boldsymbol{p}_{\mathbb{G}_2} \times \boldsymbol{\mu}_{\mathbb{G}_2} = \sum_{(i,,j)\in\mathbb{G}_2} p(i,j)\,\mu(i,j) , \tag{3.10}$$

see (2.8) of section 2. When writing code in an array processing programing language, only a few symbols would be required to write the code to do the operations defined in (3.10).

Another essential step in defining effects is to define the conditional distributions based in terms of the elements of the matrix $\boldsymbol{p}_{G_2}$in (3.5). By definition, the conditional distribution of the genotypic distribution, given allele 1, is

$$\frac{1}{p(1)} (p(1,1), p(1,0)) . \tag{3.11}$$

Let $\mu(1)$ denote the conditional expectation of the means $(\mu(1,1)), \mu(1,0)$, given allele 1. Then, by definition

$$\mu(1) = \frac{1}{p(1)} (p(1,1)\,\mu(1,1)) + p(1,0)\,\mu(1,0) . \tag{3.12}$$

The conditional expectation $\mu(0)$ is defined similarly. Observe that if the sample is in a Hardy-Weinberg equilibrium so that $p(i,j) = p(i)\,p(j)$ for all genotypes $(i,j) \in \mathbb{G}_2$,then $\mu(1)$ has the form

$$\mu(1) = \frac{1}{p(1)} p^2(1)\,\mu(1,1) + p(1)\,p(0)\,\mu(1,0) = p(1)\,\mu(1,1) + p(0)\,\mu(1,0) \tag{3.13}$$

The conditional expectation $\mu(0)$ has a similar form when the sample is in a Hardy-Weinberg equilibrium.

Given the above definitions, the effect of allele 1 is defined by

$$\alpha(1) = \mu(1) - \mu . \tag{3.14}$$

Similarly, the effect of allele 0 is defined by

$$\alpha(0) = \mu(0) - \mu . \tag{3.15}$$

In what follows, the effects defined in (3.14) and (3.15) will be referred to as first order effects. Observe that the expectation of these effects with respect to the genotypic distribution is

$$E_{\mathbb{G}_2}[\alpha] = p(1)\,\alpha(1) + p(0)\,\alpha(0) = 0 . \tag{3.16}$$

It is clear that the effects just defined can be estimated directly from the data, given that the genotype of every individual in the sample can be identified.

© 2015 Global Journals Inc. (US)

For reasons that will be made clear subsequently, the additive variance is defined by

$$var_A[W] = p(1)\alpha^2(1) + p(0)\alpha^2(0) \tag{3.17}$$

The symbol W has been included in left the side of this equation as a reminder that effects on the right side of the equation have been estimated from data. From a perspective of using an analysis of variance procedure to estimate the additive variance in (3.17), it would be impossible to estimate the effects defined in (3.14) and 3. from an estimate of $var_A[W]$. This observation clearly differentiates classical methods of estimating variance components, based on some analysis of variance procedure, from the direct method of estimating effects from the data as outlined above.

Additional effects can also be defined as measures of the presence of interactions among the two alleles in the genotype of any individual in the sample. For any genotype (i,j), a second order effect $\alpha(i,j)$ is defined as

$$\alpha(i,j) = \mu(i,j) - \mu - \alpha(i) - \alpha(j) \tag{3.18}$$

for all genotypes $(i,j) \in \mathbb{G}_2$. From this equation, it follows that

$$\mu(i,j) = \mu + \alpha(i) + \alpha(j) + \alpha(i,j) \tag{3.19}$$

for all genotypes $(i,j) \in \mathbb{G}_2$. If there are no interactions among the alleles i and j, then $\alpha(i,j) = 0$ and equation (3.19) reduces to

$$\mu(i,j) = \mu + \alpha(i) + \alpha(j) \tag{3.20}$$

If this equation holds, then it is said that the alleles i and j act additively, but if $\alpha(i,j) \neq 0$, then there is some interaction among the alleles i and j.

If a sample of individuals is in a Hardy-Weinberg equilibrium,see definition below, then in the additive case, the genetic variance has the form

$$var_G[W] = \sum_{(i.j)\in\mathbb{G}_2} p(i,j)(\mu(i,j)-\mu)^2 = p(1)\alpha^2(1) + p(0)\alpha^2(0) = var_A[W] \tag{3.21}$$

which justifies equation (3.17). By definition

$$var_{IAI}[W] = \sum_{(i,j)\in G_2} p(i,j)\alpha^2(i,j) \tag{3.22}$$

is the intra-allelic interaction variance. If the sample is in a Hardy-Weinberg equilibrium, then, by definition $p(i,j) = p(i)p(j)$ for all alleles i and j, and it can be shown that

$$var_G[W] = var_A[W] + var_{IAI}[W] \ . \tag{3.23}$$

A proof of these results may be found in chapter 3.

If the population is not in a Hardy-Weinberg equilibrium, then from equation it can be shown that $var_G[W]$ would also contain covariance terms, but the details will be omitted, but if a reader is interested in more details, chapter 3 may again be consulted. In classical quantitative genetics, the objective of an experiment would be the estimation of the variance components on the left side of equation (3.23), using some analysis of variance procedure. But, as was shown above, when the genotype of every individual in the sample is known, then all second order effects defined above can be estimated directly from the data.

For example, from equation (3.19), it follows that the formula for estimating the second order effect for genotype $(1,1)$ is

$$\alpha(1,1) = \mu(1,1) - \mu - 2\alpha(1) \ . \tag{3.24}$$

© 2015 Global Journals Inc. (US)

A similar formula for an estimator of the effect $\alpha(0,0)$ would also have the form of equation (3.24). The formula for estimating the second order effect for genotype (i,j) such that $(i \neq j)$ directly has the form

$$\alpha(1,2) = \mu(1,2) - \mu - \alpha(1) - \alpha(2) \ . \tag{3.25}$$

As can be seen from the symmetric matrices in (3.5) and (3.8), it follows that $\alpha(2,1) = \alpha(1,2)$.

Notes

For the case in which maternal and paternal alleles can be identified in any genotype, the set of genotypes $\mathbb{G}_1 = \{(1,1),(1,0),(0,1),(0,0)\}$, see section 2 for more detailed comments, would be under consideration. In this case the 2×2 matrix $\boldsymbol{P}_{\mathbb{G}_1}$ would not be symmetric, because the relation $p(1,2) \neq p(2,1)$ would hold except for rare coincidences. Similarly, the 2×2 matrix $\boldsymbol{\mu}_{G_1}$ genetic values, conditional means, would also be non-symmetric. In this case it would also be necessary to distinguish the frequencies of maternal and paternal alleles. For example, if the rows of the matrix $\boldsymbol{P}_{\mathbb{G}_1}$ were summed, the result would be the distribution of maternal alleles $(p_{mat}(1), p_{mat}(0))$. Similarly, if the columns of this matrix were summed, the result would be the distribution of $(p_{pat}(1), p_{pat}(0))$ of paternal alleles. Given these allelic distributions, to write the computer code to compute the effects in this case would require only a few changes in the code for the case in which the maternal and paternal cannot be distinguished as outlined above.

A question that naturally arises at this point formulating the model under consideration is how can we construct procedures to test the statistical significance of the estimated effects? Because the squares of the effects are summed when defining variance components, it seem fitting to use squared effect in designing tests of statistical significance. Another advantage of using squared effects is that their signs are always non-negative. For the case in which the maternal and paternal alleles cannot be distinguished, let the 5×1 column vector

$$\boldsymbol{E} = \begin{bmatrix} \alpha^2(1) \\ \alpha^2(0) \\ \alpha^2(1,1) \\ \alpha^2(0,0) \\ \alpha^2(1,0) \end{bmatrix} \tag{3.26}$$

denote the squared effects. For the case the maternal and paternal alleles can be distinguished, this vector would contain 8 elements. In the next section, procedures for testing the statistical significance of the elements in the vector in (3.26) will be presented.

IV. Permutation Tests for Statistical Significance of Estimated Effects

In this section, a class of tests of statistical significance based on computer intensive methods will be discussed. In the statistical literature, this class of tests described in this section are often referred to as permutation tests. For example, consider the case in which the maternal and paternal alleles cannot be distinguished. Then, the set of genotypes in the model would be $\mathbb{G}_2 = \{(1,1),(0,0),(1,0)\}$, where the symbol $(1,0)$ represents heterozygotes. Let $n(1,1), n(0,0)$ and $n(1,0)$ denote the number of individuals of the three genotypes in a sample of

$$n = n(1,1) + n(0,0) + n(1,0) \tag{4.1}$$

individuals whose genomes have been sequenced. For each genotype $(i,j) \in \mathbb{G}_2$, let $\boldsymbol{W}(i,j) = \{W_\nu(i,j) \mid \nu = 1,2,\cdots,n(i,j)\}$ denote the the sample of $n(i,j)$ realizations of the phenotypic random variable W describing the variability in

© 2015 Global Journals Inc. (US)

the expression of some quantitative trait for individuals of genotype (i,j).The combined data set under consideration may be represented as

$$DATA = \boldsymbol{W}(1,1), \boldsymbol{W}(0,0), \boldsymbol{W}(1,0) \tag{4.2}$$

and consists of n observations.

The first step in setting up a permutation test of the data is to compute a random permutation denoted by $PERM$ of the data set in (4.2). Given this random permutation of the data, the next step consists of choosing the first $n(1,1)$ elements of $PERM$ as a sample of the quasi observations of the $n(1,1)$ individuals of genotype $(1,1)$. Similarly, the next $n(0,0)$ elements of $PERM$ would represent quasi observations on the $n(0,0)$ of genotype $(0,0)$. Finally, the last $n(1,0)$ elements of $PERM$ would represent the $n(1,0)$ quasi observations on individuals of genotype $n(1,0)$. Then suppose these operations are repeated N times to generate a set of random mutations of the data in (4.2).

As is well known, the set $\mathbb{S}$ of all permutations of the data chosen in this manner contains

$$M = \frac{n!}{n(1,1)!n(0,0)!n(1,0)!} \tag{4.3}$$

elements. For example, for the case $n(1,1) = 33, n(0,0) = 33$ and $n(1,0) = 34$, this number is

$$M = \frac{100!}{33!33!34!} = 4.1924 \times 10^{45}, \tag{4.4}$$

which is a very large number. Indeed it is so large that most current computers would be unable compute this many permutations of the data in an acceptably short time span. Consequently, when doing a permutation test of the data, an investigator would need to compute some number N of permutations that is much less than M.Most programming languages contain programs to compute random numbers, which can be used to write code for computing a sample of random permutations of a data set.

When doing a permutation test, the null hypothesis H_0 is that the observed data set is a random sample from the set $\mathbb{S}$ of all possible permutations of the data. To carry out such a test, a computer would need to be programmed in such a way that some number N of random permutations of the data would be computed and for each permutation of the data estimates of the effects of the vector $\boldsymbol{E}$ in equation (3.26) of section 3 would be computed. Let $\boldsymbol{SIM}$ denote a $5 \times N$ matrix of simulated estimates of the five effects in the vector $\boldsymbol{E}$ such that each column of this matrix is an estimated realization of the observed vector $\boldsymbol{E}$ based on a random permutation of the data. Similarly, let $\boldsymbol{ALPHA}$ denote a $5 \times N$ matrix such that each column is a copy of the vector $\boldsymbol{E}$. Then consider the relationship $\geq$ and a $5 \times N$ matrix $\boldsymbol{R}$ defined by

$$\boldsymbol{R} = \boldsymbol{SIM} \geq \boldsymbol{ALPHA}\,. \tag{4.5}$$

Let $\boldsymbol{SIM}(i,j)$ denote the element from the i-th row and j-th column of the matrix $\boldsymbol{SIM}$, and define the element $\boldsymbol{ALPHA}(i,j)$ analogously. Each elements of the matrix $\boldsymbol{R}$ is 0 or 1. If the relation

$$\boldsymbol{SIM}(i,j) \geq \boldsymbol{ALPHA}(i,j) \tag{4.6}$$

is true, then the element $\boldsymbol{R}(i,j) = 1$, and if this relation is false, $\boldsymbol{R}(i,j) = 0$. As will be demonstrated in what follows, by using the matrix $\boldsymbol{R}$ various types of p-values may used to judge statistical significance of each of the estimated effects.

For example, let the $SUMROWS$ denote the array with 5 elements that results from summing the rows of the matrix $\boldsymbol{R}$, and let $r(\nu)$ denote the ν-th element in this array, where $\nu = 1, 2, \cdots, 5$. For example, according the ordering

© 2015 Global Journals Inc. (US)

used in defining the 5×1 vector $\boldsymbol{E}$ in section 3, the number $r(\nu)$ denotes the number of times among the N sample values that the inequality

$$\boldsymbol{SIM}(1,j) \geq \boldsymbol{ALPHA}(1,j) \tag{4.7}$$

was satisfied for effect 1.

Observe that for every $\nu = 1.2.\cdots, 5$ the value of $r(\nu)$ will be a member of the set $\{x \mid x = 0, 1, 2 \cdot \cdot, N\}$. Consequently, $p(\nu) = r(\nu)/N$ is the estimated p-value for the estimated effect $\alpha^2(\nu)$ for $\nu = 1, 2, \cdots, 5$, see the vector $\boldsymbol{E}$ in (3.26) in section 3 for details. Note that according to the elements in this vector, $p(1)$ is the p-value for the estimated effect $\alpha^2(1)$ of the marker allele 1. If for any $\nu = 1, 2, \cdots, 5$, $p(\nu) < 0.05$, then the null hypothesis H_0 that the data are a sample from the set $\mathbb{S}$ of all permutations of the data will be rejected and the estimate of the effect $\alpha(\nu)$ will be said to be statistically significant. In this example, the probability 0.05 was chosen arbitrarily, but an investigator would be free to choose any other small probability as a bench mark for declaring statistical significance. Alternatively, an investigator may want to observe each p-value, and make a judgement a to whether an estimated effect was statistically significant. At this point in the discussion, note that each of the probabilities estimated from the data, pertains to only the statistical significance of one of the five estimated effects under consideration. But, as will be demonstrated below, it will also be possible, to obtain joint probabilities that some sets of effects are jointly significant that will be illustrated in the next paragraph.

This other set of interesting joint p-values may be computed by summing the columns of the matrix $\boldsymbol{R}$. Let $SUMCOLUMNS$ denote an array with N elements that are sums of the columns of $\boldsymbol{R}$, and let $c(\nu)$ denote the sum of column ν. Then, the value of each $c(\nu)$ is an integer in the set $\{y \mid y = 0, 1, 2, 3, 4, 5\}$ for $\nu = 1, 2, \cdots, N$. For example, if for some column ν, $c(\nu) = 0$, then all the numbers in column ν of $\boldsymbol{R}$ would be 0, indicating that for every row i the inequality (4.6) was not satisfied for column $\nu = j$. But, if $c(\nu) = 5$ for some column $\nu = j$, then the inequality in (4.6) would be satisfied for all rows $i = 1, 2, 3, 4, 5$. Given the array $SUMCOLUMNS$, it would be possible to estimate a distribution that would provide insights into the joint statistical significance of the 5 estimated effects in the column $\boldsymbol{E}$ in (3.26) in section 3. For any fixed $\nu = 0, 1, 2, 3, 4, 5$, let $m(\nu)$ be the number of elements of the array $SUMCOLUMNS$ has the value ν. Let $p_{joint}(\nu) = m(\nu)/N$ denote the estimated probability for the values $\nu = 0, 1, 2, 3, 4, 5$. By viewing these joint probabilities, an investigator would be in a position to judge whether all the five estimated effects were jointly statistically significant. If, for example, this distribution were skewed to the left so that the probabilities $p_{joint}(\nu)$ for $\nu = 0, 1$ were larger than the probabilities $p_{joint}(\nu)$ for $\nu = 4, 5$, then an investigator could make a judgement as to whether all estimates of the five effects were jointly statistically significant.

Some investigators may wish to carry out a permutation test, but in this chapter the focus of attention will be focused on another class of tests of statistical significance based on Monte Carlo simulation methods that will be formulated in the next section. However, in practice, an investigator may want to carry out statistical test of significance belonging to different classes of tests to get some idea as to whether estimated effects are statistically significant for at least two classes of tests of statistical significance

V. Testing the Statistical Significance of Estimated Effects Based on Monte Carlo Simulation Methods

There is also another approach to judging whether estimates of the five effects are statistically significant by using Monte Carlo simulation methods. Suppose, for example, that there are $n(i,j)$ non-negative simulated realizations of the random variable $\boldsymbol{W}(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$. The rationale

© 2015 Global Journals Inc. (US)

under lying the choice of non-negative random variable is that most measures with respect to some quantitative trait are non-negative numbers. One of the simplest approaches to simulation realizations of non-negative random variables is to use the absolute or folded normal distribution. Suppose a random $X(i,j)$ has a normal distribution with an expectation $\mu(i,j)$ and variance $\sigma^2(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$. Let Z denote a standard normal random variable with expectation 0 and variance 1. Then, as is well known, if Z is a simulated realization from a standard normal distribution, then $X(i,j) = \mu(i,j) + \sigma(i,j) Z$ is a simulated realization of the random variable $X(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$. Given a simulated realization of a random variable $X(i,j)$, $W(i,j) =\mid X(i,j) \mid$ is a simulated realization of a random variable $W(i,j)$ with an absolute normal distribution for every genotype $(i,j) \in \mathbb{G}_2$. A more detailed description of the folded normal distribution will be given in a appendix.

The next step in setting up a Monte Carlo simulation experiment is to formulate a procedure for testing a null hypothesis. Suppose, for example, that the $n(i,j)$ observations of the random variable $\boldsymbol{W}(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$ are a random sample from a folded normal distribution. Moreover, suppose there are real data consisting of a sample of size $n(i,j)$ for each genotype $(i,j) \in \mathbb{G}_2$. To simplify the notation, from now on the symbols $\mu(i,j)$ and $\sigma^2(i,j)$ will denote estimates of the corresponding expectation and variance for each genotype $(i,j) \in \mathbb{G}_2$ based on the data. Given the data, an investigator could also estimate the genotypic distribution $\{p(i,j) \mid (i,j) \in \mathbb{G}_2\}$ as well as the 5×1 vector $\boldsymbol{E}$ of effects. In this situation, an investigator may wish to entertain the null hypothesis H_0 of homogeneity and suppose that there are positive numbers μ and σ^2 such that

$$\mu(i,j) = \mu_{uc} \text{ and } \sigma^2(i,j) = \sigma^2_{nc} \tag{5.1}$$

for all genotypes $(i,j) \in \mathbb{G}_2$, the subscript uc stands for unconditional.

At this point, to simulate samples from a normal distribution with expectation μ_{uc} and variance σ^2_{uc}, an investigator may decide to choose μ_{uc} and σ^2_{uc} as

$$\mu_{uc} = \sum_{(i,j)\in\mathbb{G}_2} p(i,j)\,\mu(i,j) \tag{5.2}$$

and

$$\sigma^2_{uc} = \sum_{(i,j)\in\mathbb{G}_2} p(i,j)\,\sigma^2(i,j) \tag{5.3}$$

as the parameters in the normal distribution to be used to simulate $n(i,j)$ of realizations of random variable the $X(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$. Given this simulated sample of realizations of random variables with a homogeneous distribution, the simulated $n(i,j)$ realizations of the phenotypic random variables $W(i,j)$ would be computed by using the formula $W(i,j) =\mid X(i,j) \mid$ for every genotype $(i,i) \in \mathbb{G}_2$.

This choice of σ^2_{uc}, for example, would result in a simulated samples with a variance that would be close to that in the original data so that unrealistic outliers would occur with small probabilities in the simulated data. The choice of μ_{uc}, however, is less sensitive than that for σ^2_{uc}. For it is interesting to note that if the hypothesis H_0 were true, then all effects in the vector $\boldsymbol{E}$ are 0 for any choice of μ. For example, consider the first order effect

$$\alpha(1) = \mu(1) - \mu \tag{5.4}$$

see (3.14) in section 3. By inspecting (3, 13) in section 3, it can be seen that if H_0 is true, then $\mu(1) = \mu$ so that $\alpha(1) = 0$.

Similar arguments may be used to see that if H_0 is true, then all the effects in the vector $\boldsymbol{E}$ would be 0. It is interesting to note that for any choice

© 2015 Global Journals Inc. (US)

of μ in a simulation procedure, all the effects in the vector $\boldsymbol{E}$ would be 0. From the point of view of simulating $n(i,j)$ realizations of the random variables $W(i,j) =\mid X(i,j) \mid$ for every genotype $(i,i) \in \mathbb{G}_2$, it can be seen that if the expectations of the random variables $X(i,j)$ are some constant μ, then the expectations of the random variables $W(i,j)$ would also be some constant. Hence, the estimated effects based on the means of folded normal distribution would also be constant. If a reader is interested in further details regarding the folded normal distribution, it is suggested that the appendix be consulted.

To carry out a test of statistical significance using the Monte Carlo simulation procedure just outlined, an investigator would start with the array $\boldsymbol{DATA}$ in (2.2) . Then the first step in a Monte Carlo simulation procedure would be that of simulating a quasi-data set, $\boldsymbol{QUASIDATA}$, consisting of n realizations of a random $\boldsymbol{W}$ with a normal distribution with expectation μ_{uc} and variance σ^2_{uc} as suggested as indicated above. Let $\boldsymbol{SIMW}(1,1)$ denote the first $n(1,1)$ elements of the array , $\boldsymbol{QUASIDATA}$ and define the arrays $\boldsymbol{SIMW}(0,0)$ and $\boldsymbol{SIMW}(1,0)$ similarly for the numbers of individuals $n(0,0)$ and $n(1,0)$ of genotypes $(0,0)$ and $(1,0)$, respectively. Then, the simulated quasi array of data for the three genotypes may be represented in the form

$$\boldsymbol{QUASIDATA} = \boldsymbol{SIMW}(1,1), \boldsymbol{SIMW}(0,0), \boldsymbol{SIMW}(1,0) \tag{5.5}$$

just as the real data in (4.2).

Given the simulated data set in (5.5), the next step in the Monte Carlo simulation procedure would be that of computing estimates of the five effects in the 5×1 vector $\boldsymbol{E}$. By repeating this step just outlined $N > 1$ times, a version of the $5 \times N$ matrix $\boldsymbol{R}$ in (4.5) could be computed. Then, by using the procedure outlined in section 4, a test statistical significance for any estimated effect the vector $\boldsymbol{E}$ could could be accomplished. Similarly, a test for the joint statistical significance of the five estimated effects could be carried out, by using the procedure outlined in section 4 by summing the rows of the matrix $\boldsymbol{R}$. In such an experiment, an investigator would be testing the null hypothesis that all the five effects in the vector $\boldsymbol{E}$ are zero.

To simplify the notation, from now on the symbols $\mu(i,j)$ and $\sigma^2(i,j)$ will denote estimates of the corresponding expectation and variance for each genotype $(i,j) \in \mathbb{G}_2$. The rational for considering simulated non-negative random variable is that the majority of measurements for some quantitative traits are usually non-negative numbers as stated above. Given this information, an investigator could estimate the genotypic distribution $\{p(i,j) \mid (i,j) \in \mathbb{G}_2\}$ as well as the 5×1 vector $\boldsymbol{E}$ of effects. In setting up this Monte Carlo experiments , an investigator may wish to entertain the null hypothesis H_0 that there are positive numbers μ and σ^2 such that

$$\mu(i,j) = \mu_{uc} \text{ and } \sigma^2(i,j) = \sigma^2_{uc} \tag{5.6}$$

for all genotypes $(i,j) \in \mathbb{G}_2$ see (5.2) and (5.3).

This choice of σ^2, for example, would result in a simulated samples with a variance that would be close to that in the original data so that unrealistic outliers would occur with small probabilities in the simulated data. The choice of μ, however, is less sensitive than that for σ^2. For it is interesting to note that if the hypothesis H_0 were true, then all effects in the vector $\boldsymbol{E}$ are 0 for any choice of μ. For example, consider the first order effect

$$\alpha(1) = \mu(1) - \mu \tag{5.7}$$

see (14) in section 3. By inspecting (3.16) in section 3, it can be seen that if H_0 is true, then $\mu(1) = \mu$ so that $\alpha(1) = 0$. Similar arguments may be used to see that if H_0 is true, then all the effects in the vector $\boldsymbol{E}$ would be 0. It is

© 2015 Global Journals Inc. (US)

interesting to note that for any choice of μ in a simulation procedure, all the effects in the vector $\boldsymbol{E}$ would be 0.

To carry out a test of statistical significance using the Monte Carlo simulation procedure just outlined, an investigator would start with the array $\boldsymbol{DATA}$ in (4.2) . Then the first step in a Monte Carlo simulation procedure would be that of simulation a quai-data set, $\boldsymbol{QUASIDATA}$, consisting of $n(i.j)$ realizations of random variables $\boldsymbol{X}(i,j)$ for every genotype $(i,j) \in \mathbb{G}_2$ with a normal distribution with expectation μ and variance σ^2 that would be transformed to $n(i,j)$ folded normal random variables $W(i.j)$ for every $(i,j) \in \mathbb{G}_2$. Let $\boldsymbol{SIMW}(1,1)$ denote the first $n(1,1)$ elements of the array , $\boldsymbol{QUASIDATA}$, and define the arrays $\boldsymbol{SIMW}(0,0)$ and $\boldsymbol{SIMW}(1,0)$ similarly for the numbers of individuals $n(0,0)$ and $n(1,0)$ of genotypes $(0,0)$ and $(1,0)$, respectively. Then, the simulated quasi array of data for the three genotypes may be represented in the form

$$\boldsymbol{QUASIDATA} = \boldsymbol{SIMW}(1,1), \boldsymbol{SIMW}(0,0), \boldsymbol{SIMW}(1,0) \tag{5.8}$$

just as the real data in (4.2).

Given the simulated data set in (5.8), the next step in the Monte Carlo simulation procedure would be that of computing estimates of the five effects in the 5×1 vector $\boldsymbol{E}$. By continuing this step just outlined $N > 1$ times, a version of the $5 \times N$ matrix $\boldsymbol{R}$ in (4.5) could be computed. Then, by using the procedure outlined in section 4, a test of statistical significance for any estimated effect the vector $\boldsymbol{E}$ could could be accomplished. Similarly, a test for the joint statistical significance of the five estimated effects could be carried out, by summing the rows of the matrix $\boldsymbol{R}$ and counting the numbers of each of the values $0, 1, 2, 3, 4, 5$. In any computer experiment of the type under consideration, an investigator would be testing the null hypothesis that all the five effects in the vector $\boldsymbol{E}$ are 0.

One of the principal goals of the type of Monte Carlo simulation under consideration is that of computing p-values on which a judgment of statistical significance for each of the five effects that were estimated form the data can be made. The $5 \times N$ matrix $\boldsymbol{R}$ plays a fundamental role in estimating the p-values. Consider, for example, an element by element representation of this matrix of the form

$$\boldsymbol{R} = [r_{ij}] \tag{5.9}$$

and let $\alpha^2(i)$ denote squared effect estimated form the data in the vector $\boldsymbol{E}$ for $i = 1, 2, 3, 4, 5$. Similarly, let $\alpha^2(i,j)$ be an estimate of the squared effect i of replication j of the Monte Carlo simulation experiment for $j = 1, 2, \cdots, N$. Then, for each $i = 1, 2 \cdot \cdot, 5$ and $j = 1, 2, \cdots, N$, r_{ij} may be interpreted as a Bernoulli indicator such that $r_{ij} = 1$ if the event

$$\left[\alpha^2(i,j) \geq \alpha^2(i)\right] \tag{5.10}$$

occurs and $r_{ij} = 0$ if event defined in (5.10) does not occur.

Given a null hypothesis H_0 let

$$E[r_{ij} \mid H_0] = 1P[r_{ij} = 1 \mid H_0] + 0P[r_{ij} = 0 \mid H_0] = P[r_{ij} = 1 \mid H_0] = p(i \mid H_0) \tag{5.11}$$

denote the conditional probability that the event in (5.10) occurs, given H_0 for $j = 1, 2, \cdots, N$. The technical details of the random number generator used in the Monte Carlo simulation experiments reported in following sections will not be discussed here. But, because the randomness in the properties of the sequences of uniform random numbers taking values in the interval $[0, 1)$, the assumption that the events denoted in (5.10) are independent for all $j = 1, 2, \cdots, N$ so that for each $i = 1, 2, \cdots, 5$, it is highly plausible to assume that the sequence

© 2015 Global Journals Inc. (US)

of Bernoulli indicator functions are independently distributed with a common expectation $p\,(i \mid H_0)$ defined in (5.11) . Therefore, by invoking the law of large numbers, it follows that

$$\lim_{N\uparrow\infty} \frac{1}{N} \sum_{j-1}^{N} r_{ij} = p\,(i \mid H_0) \tag{5.12}$$

for every $i = 1, 2, \cdots, 5$. If the strong law of large numbers is invoked, then the limit in (5.12) holds with probability one. In general the larger the choice of the number N,the greater is the reliability of the estimate in (5.14) . In the experiments that will be reported in subsequent sections of this chapter, N was chosen as 10,000. With this choice of N the computing run time of each of the experiments reported in the sections to follow was in the range of 2 to 3 minutes, which would be acceptable if a quantitative trait under consideration involved repeating an experiments for 10 to 15 loci which were thought to be involved in the expression of the trait.

At this point in the development of ideas making up the procedure for tests of significance under consideration, it will be helpful to express the ideas in the last paragraph of section 4 more formally. Let

$$s_j = \sum_{i=1}^{5} r_{ij} \tag{5.13}$$

denote the sum of the indicators in column j or the matrix $\boldsymbol{R}$ for $j = 1, 2, \cdots, N$. Then the array $SUMCOLUMNS$ defined in the last paragraph of section 4 has the form

$$SUMCOLUMNS = (s_1, s_2, \cdots, s_N) \ . \tag{5.14}$$

Let

$$[j \mid s_j = x] \tag{5.15}$$

for $x = 0, 1, 2, 3, 4, 5$. Then

$$m\,(x) = \sum_{j \in [j \mid s_j = x]} s_j \tag{5.16}$$

for all x and

$$\sum_{x=0}^{5} m\,(x) = N \ . \tag{5.17}$$

Then in terms of the formal system developed in this section

$$p_{joint}\,[x \mid H_0] = \frac{m\,(x)}{N} \tag{5.19}$$

for all x is the conditional distribution for judging the joint statistically significance of the 5 effects under consideration. Observe that the conditional distribution defined in (5.19) has the property

$$\sum_{x=0}^{5} p_{joint}\,[x \mid H_0] = 1 \tag{5.21}$$

as it should. In any simulation experiment, it is useful to check that this equation holds as part of tests for the correctness of the software. Equation (5.19) is justified, because sets in the collection of sets in (5.15) are a disjoint partition of the set $SUMCOLUMNS$ in (5.14).

© 2015 Global Journals Inc. (US)

VI. Monte Carlo Simulation Experiments on Simulating Data and Testing the Null Hypotheses

There is a vast literature on Monte Carlo simulation procedures that have been used in many fields of science. For example, the paper by Mode and Gallop (2008) [1], as it turned out, provided the authors into a window on an extensive literature on Monte Carlo simulation procedures as used in many fields of science, see the internet link

$$http \quad : \quad //biomedupdater.com/urlu8c?srk = 2a20a8d6419d877$$
$$ad74198186ff9a1c0ae53b88acaa0ade5587ca0baf302b72$$

Furthermore, the book cited in [2] and edited by the author, also contains an extensive collection of papers on the application of Monte Carlo simulation methods in various fields of biology and related sciences. In this section, the random number generator with a very long period set forth in Mode and Gallop (2008) [1] in section on Monte Carlo simulation methods will be used as well as in all sections to follow, containing accounts of Monte Carlo simulation experiments used to test various version of the null hypothesis. There is a caveat that a reader should be aware of when reading the experimental results reported in this and the following sections is that the random number generator used in all experiments reported in this paper was designed for computers based on 32 bit words. The computers used to conduct Monte Carlo simulation experiments reported in this paper, however, were based on 64 bit words. Algorithms for random number generators for 64 bit words may be found in the papers cited in Mode and Gallop (2008) [1], but to implement these algorithms for use on computers based on 64 bit words would require an extensive period of development, using array manipulating programming languages such as APL. It seems plausible, however, that if the Monte Carlo simulation experiments reported in this paper were based on a random number generator designed for a 64 bit word computers rather than the 32 bit word generator, that the results and conclusions would not be significantly different.

The first step in setting up a Monte Carlo experiment to test some null hypothesis H_0 is to simulate the data that will used to estimate all parameters and effects as functions of parameters. In the best of all worlds, data of the type under consideration would be posted on the internet so it could be downloaded by investigators and used to present concrete examples of the application of new statistical procedures. But, unfortunately, getting permission to use such data is often impossible, unless you are a member of the group that has assembled the data. The parameter values used to simulate the data used in the experiments discussed in the section are shown in Table 6.1 below.

Table 6.1 Parameter Values Used to Simulate Data

$\mu(1,1)$	30
$\sigma(1,1)$	$0.25\mu(1,1)$
$\mu(0,0)$	40
$\sigma(0,0)$	$0.25\mu(0,0)$
$\mu(1,0)$	60
$\sigma(1,0)$	$0.25\mu(1,0)$

By way of interpreting the chosen values in Table 6.1, on some scale of hypothetical units used to measure the expression of some quantitative trait under consideration, the expected values for the three genotypes $(1,1)$, $(0,0)$ and $(1,0)$ were chosen as $\mu(1,1) = 30, \mu(0,0) = 40$ and $\mu(1,0) = 60$. The rational used in choosing these numbers was to assign different values to each of these expected values so that the estimated genetic variance would be positive. The rational for not choosing $\mu(1,0) = 50$ but as $\mu(1,0) = 60$ was to consider the

© 2015 Global Journals Inc. (US)

Ref

1. Mode, C. J. and Gallop, R. J. (2008) A Review on Monte Carlo Simulation Methods as They Apply to Mutation and Selection as Formulated in Wright- Fisher Models of Evolutionary Genetics. Mathematical Biosciences 211: 205-225.

case there was a heterotic effect for heterozygotes of genotype $(1,0)$, i.e., it was assumed that there was some interaction of alleles 1 and 0 in individuals, whose genotype was the heterozygote. The reason for choosing the standard deviations $\sigma(1,1)$, $\sigma(0,0)$ and $\sigma(1,0)$ as a common fraction 0.25 of the expectation for each genotype was that the estimates of the environmental valance for each genotype seemed plausible as observed in preliminary experiments. The sample sizes chosen for the genotypes were $n(1,1) = 100$, $n(0,0) = 200$ and $n(1,0) = 450$. These sample sizes resulted in the allele frequencies $p(1) = 0.433$ and $p(0) = 0.567$. The estimated heritability based on the simulated data for the three genotypes was $H = 0.4280$.

Two null hypotheses were considered in the illustrative examples on testing the statistical significance of the squared effects estimated from the simulated data. Presented in table 6.2 are the assigned parameter values used in testing the two null hypotheses under consideration.

Table 6.2 Parameter Valued Used in Testing Two Null Hypotheses Based on Monte Carlo Simulation Methods

$Parn$	$H_0(1)$	$H_0(2)$
$\mu(1,1)$	μ_{uc}	0
$\sigma(1,1)$	σ_{uc}	σ_{uc}
$\mu(0,0)$	μ_{uc}	0
$\sigma(0,0)$	σ_{uc}	σ_{uc}
$\mu(1,0)$	μ_{uc}	0
$\sigma(1,0)$	σ_{uc}	σ_{uc}

In table 6.2, the subscript uc stands for unconditional expectations and standard deviations as shown in (5.1). Moreover, the estimates of these parameters were computed from the simulated data, using formulas (5.2) and (5.3). The estimate σ_{uc} was computed using the formula

$$\sigma_{uc} = \left(\sigma_{un}^2\right)^{\frac{1}{2}}, \tag{6.1}$$

see (5.2). Table 6.3 contains the symbolic form of the squares of the estimated effects, the estimates of these parameters based on the simulated data and the p-values computed in tests of statistical significance of the null hypotheses $H_0(1)$ and $H_0(2)$ using Monte Carlo simulation methods under consideration.

Table 6.3. Statistical Test of Significance of the Estimates of the Squared Effects Based on Monte Carlo Methods

ESQ	EST	$H_0(1)$	$H_0(2)$
$\alpha^2(1)$	0.03185	0.6124	0.3948
$\alpha^2(0)$	0.00557	0.7796	0.64411
$\alpha^2(1,1)$	376.7118	0	0
$\alpha^2(0,0)$	92.48211	0	0
$\alpha^2(1,0)$	106.8706	0	0

In the simulation experiment designed to test the null hypothesis $H_0(1)$, the squares of the estimated effects were computed with 10,000 Monte Carlo replications, using the parameter assignments listed in the second column of table 6.2. Similarly, to test the null hypothesis $H_0(2)$ 10,000 Monte Carlo replications of the squared effects were again computed. The p-values listed in columns 3 and 4 of table 6.3 were computed using the 10,000 Monte Carlo replication as set forth in equation (5.12) with $N = 10,000$ for each null hypothesis being tested. From rows 1 and 2 of table 6.2, it can be seen that if the null hypothesis

© 2015 Global Journals Inc. (US)

$H_0(1)$ is true, then $\alpha^2(1) = 0$ and $\alpha^2(0) = 0$. The estimates of these two squared effects based on the simulated data are 0.03185 and 0.00557, with the corresponding p-values 0.6124 and 0.7796, respectively, under null hypothesis $H_0(1)$, and are not sufficiently small to reject the null hypothesis being tested. An investigator may therefore conclude that the additive effects of alleles 1 and 0 are not statistically different from 0. As can be seen from the second column of table 6.3, the estimates of the squared effects for interactions effects $\alpha^2(1,1)$, $\alpha^2(0,0)$ and $\alpha^2(1,0)$ are 376.7118, 92.48211 and 106.8706, respectively, with corresponding p-values $0, 0, 0$. The number 0 is the smallest possible p-value; consequently, the squared effects for allelic interaction are highly significantly different form zero under the null hypothesis $H_0(1)$. It is interesting to note if the null hypothesis $H_0(2)$ were tested using the same methods, the statistical conclusions just stated for the five squared effects under consideration would not change even though the p-values for the additive effects differ from those that were computed under the null hypothesis $H_0(1)$.

The last set of p-values that will be presented in this section are those described in (5.13) through (5.21), which were estimated by summing the columns $5 \times 10,000$ matrix $\boldsymbol{R}$ of realized Bernoulli indicator functions as described in section 5. Presented in table 6.4 are the are estimates of these p-values under null hypotheses $H_0(1)$ and $H_0(2)$.

Table 6.4 Estimates of the Joint p-Values Under Each Null Hypothesis

Values	0	1	2	3	4	5
$H_0(1)$	0.2204	0.1672	0.6124	0	0	0
$H_0(2)$	0.3559	0.2403	0.3948	0	0	0

Observe that for each row in this table the listed p-values are a distribution satisfying equation (5.21). For example, for null hypothesis $H_0(1)$

$$0.2204 + 0.1672 + 0.6124 = 1. \tag{6.2}$$

An equation of the same form would be also be valid for the row in table 6.4 corresponding to the null hypothesis $H_0(2)$. By way of interpreting this table, with an estimated probability 0.2204, the sum of a column of the indicator matrix $\boldsymbol{R}$ would be zero under the null hypothesis $H_0(1)$. Similarly, under this null hypotheses, 0.6124 was the estimated probability that the sum of a column of indicators was 2. Observe that, because this probability is largest in the distribution under the null hypothesis $H_0(1)$, the number 2 is the mode of this distribution of joint p-values. It is also interesting to note that the estimated probability that a column sum of indicators in the matrix $\boldsymbol{R}$ had the value $3, 4$ or 5 was 0 under both null hypotheses under consideration. It is interesting to also note that the number 2 was also the mode of the distribution of joint probabilities under null hypothesis $H_0(2)$. Moreover, under both null hypotheses, the events that two columns of the indicator matrix $\boldsymbol{R}$ were both occurred more often and has a greater influence on the p-values for judging the statistical significance of each squared effect that was estimated by summing the rows of the indicator matrix $\boldsymbol{R}$.

Software to test the statistical significance of the estimated heritability $H = 0.4280$ was also developed so that p-values of testing null hypotheses of the form H_0: $H = 0$ could be tested in Monte Carlo simulation experiments. For all the null hypotheses tests described in this section, the p-value was zero. Hence, the estimate of heritability was, in a statistical sense, highly significantly different from zero.

© 2015 Global Journals Inc. (US)

VII. Two Monte Carlo Simulation Experiments to Simulate Data and Test Null Hypotheses

In this section two sets of simulated data, for experiments A and B, will be used in testing null hypotheses. Presented in Table 7.1 are the chosen parameter values for simulation data used in experiments A and B reported in this section.

Table 7.1 Parameter Values Used for Simulation Data in Experiments A and B

Prams	*Exp A*	*Exp B*
$\mu(1,1)$	30	30
$\sigma(1,1)$	$0.25\mu(1,1)$	$\mu(1,1)$
$\mu(0,0)$	40	40
$\sigma(0,0)$	$0.25\mu(0,0)$	$\mu(0,0)$
$\mu(1,0)$	60	60
$\sigma(1,0)$	$0.25\mu(1,0)$	$\mu(1,0)$
$n(1,1)$	10	100
$n(0,0)$	1,000	200
$n(1,0)$	15	450

From this table it can see seen that, with the exception of the sample sizes $n(1,1), n(0,0)$ and $n(1,0)$, the values of the parameters chosen for experiment A in the second column of the table 7.1 are the same as those in Table 6.1. In experiment A, it was assumed allele 1 was a rare mutation that arose in some ancestral population from which the sample evolved. The number of individuals of genotype $(1,1)$ in the sample was chosen as $n(1,1) = 10$, and the sample sizes for genotypes $(0,0)$ and $(1,0)$ were chosen as $n(0,0) = 1,000$ and $n(1,0) = 15$. The objective of experiment A was to provide some insights concerning what impact the low frequency of allele 1 in the sample would have on the estimates and test of statistical significance reported in section 6. With the exceptions of the standard deviations, which were chosen as $\sigma(i,j) = \mu(i,j)$ for all genotypes $(i,j) \in \mathbb{G}_2$, the sample sizes and expectations of a quantitative trait under consideration were the same as those in table 6.1. The objective of experiment B was to provide some insights into the effects of higher environmental variances would have on the estimates of parameters when compared with the experiments and tests of statistical significance reported in section 6.

In experiment A, the estimated of the frequencies of alleles 1 and 0 were about $p(1) = 0.01$ and $p(0) = 0.9$. The estimate of heritability in experiment A was $H_A = 0.0902$. In experiment B, the estimates of the frequencies of alleles 1 and 0 were $p(1) = 0.4333$ and $p(0) = 0.5667$, and the estimate of heritability was $H_B = 0.0932$.

Contained in table 7.2 are the parameter values used to test the null hypotheses in experiments A and B.

Table 7.2 Parameter Values for Testing Null Hypotheses

Parms	*Exp A*	*Exp B*
$\mu(1,1)$	$\mu_{uc}(A)$	$\mu_{uc}(B)$
$\sigma(1,1)$	$\sigma_{uc}(A)$	$\sigma_{uc}(B)$
$\mu(0,0)$	$\mu_{uc}(A)$	$\mu_{uc}(B)$
$\sigma(0,0)$	$\sigma_{uc}(A)$	$\sigma_{uc}(B)$
$\mu(1,0)$	$\mu_{uc}(A)$	$\mu_{uc}(B)$
$\sigma(1,0)$	$\sigma_{uc}(A)$	$\sigma_{uc}(B)$

In table 7.2 the same subscripts are used of the designated parameter values used to test to define the null distributions for testing null hypotheses in experiments A and B. The values of these parameters were chosen by using the formulas in

© 2015 Global Journals Inc. (US)

equations (5.2) and (5.3), but it is clear from the assigned parameter values in table 7.1 that the estimated values of the parameters in table 2 would differ in experiments A and B.

Table 7.3 contains estimates of the squares of effects and p-values estimated by using 10, 000 Monte Carlo replications in both experiments A and B.

Table 7.3 Estimated Squared Effects and P-Values for Experiments A and B

Parms	*Est A*	$p-values\ A$	*Est B*	$p-values\ B$
$\alpha^2(1)$	101.5387	0	54.4873	0
$\alpha^2(0)$	99.4566	0.1525	56.0312	0
$\alpha^2(1,1)$	105.7899	0	1771.4965	0
$\alpha^2(0,0)$	403.1152	0	1043.0865	0
$\alpha^2(1,0)$	409.493	0	7.1615	0.0111

From table 7.3, the second column contains the estimates of the squared effect using the simulated data. With the exception of the estimate of additive effect $\alpha^2(0)$, which was 99.4566, the estimated squares of the remaining effects have zero p-values, are highly statistically different from zero. By way of contrast, in table 6.3 all the squared additive effects, are not statistically different under either null hypotheses $H_0(1)$ or $H_0(2)$. From this example, it can be seen that the small numbers of genotypes $(1,1)$ and $(1,0)$ had a significant effect on the reported p-values in column 3 of the table. In column 5 of the table where the p-values for experiment B are displayed, it can be seen from the estimated p-values that, with the exception that for estimate of the squared effect $\alpha^2(1,0)$, the estimates of the four other squared effect are highly statistically different from zero. It is also interesting to note from an estimated p-value of 0.0111, it may be concluded at about the one percent level, the estimate 7.1615 of $\alpha^2(1,0)$ statistically different from zero. The results of this experiments suggest that, even with a low estimate of heritability resulting from higher assigned values of the environmental variances, there may be interesting cases using real data that would lead to squared effects that were statistically different from zero.

The next set of p-values to be presented in this section are two joint distributions for each of the null hypotheses under consideration as shown in table 7.4

Table 7.4 Estimates of the Joint p-Values Under Each Null Hypothesis

Values	0	1	2	3	4	5
$H_0(A)$	0.8475	0.1525	0	0	0	0
$H_0(B)$	0.9889	0.0111	0	0	0	0

From table 7.4, it can be seen that for both hypotheses the mode of the distribution was 0. It is also interesting to note that for hypothesis $H_0(A)$ the p-value at the number 1 is the same as the p-value in table 7.3 corresponding to the estimate of the parameter $\alpha^2(0)$. Similarly, the p-value for hypothesis $H_0(B)$ at the number 1 is the same as the p-value corresponding to the estimate of the parameter $\alpha^2(1,0)$ in table 7.3. From these observations, it follows that for both null hypotheses there were no columns of the $5 \times 10,000$ matrix $\boldsymbol{R}$ with ones for the values 2,3,4,5. It should also be mentioned that the estimates of heritability for both hypotheses had zero p-values even though both estimates were small. Recall that $H_A = 0.0902$ and $H_B = 0.0932$.

VIII. Further Developments and Potential Applications

By using various statistical and other methods, researchers have identified a number regions in the human genome that are associated with diseases such

© 2015 Global Journals Inc. (US)

as Altzheimer's. A recent paper on such regions has been reported in Raj et al. (2012) [5], in which, among other things, eleven regions of the human genome, associated with susceptibility to Altzheimer's disease, have been identified. Evidence is also reported on the existence of a protein net work involving four of these of these regions that is sustained in the human genome by natural selection. The number of individuals in this sample are about 5,000 that the DNA of each individual in the sample has been sequenced. In a related paper Rossin et al. (2011) [6] report that proteins coded by identified regions of the human genome associated with immune-mediated diseases, physically interact and suggest some underlying basic biology. Altzheimer's disease may also be viewed as a quantitative trait whenever its expression is measured on some numerical scale. Moreover, if for each of the 11 genomic regions may be identified in two alternative forms, then from the point of view of quantitative genetics these 11 regions may be referred to as loci with two alleles at each locus.

When an investigator considers 11 loci and two alleles per locus, the number of effects that may be estimated directly will become very large even if only three genotypes per locus may be identified as discussed in the published in chapters 3 and 4. For example for the case of four loci for which only three genotypes may be identified per locus, the number of identifiable genotypes with respect to 11 loci would be

$$3^{11} = 1.771\,5 \times 10^5 \tag{8.1}$$

However, if only four loci were under consideration, then the number of identifiable genotypes would be

$$3^4 = 81\ . \tag{8.2}$$

As expected this is a much smaller number than that in (8.1), but, nevertheless, when an array with 81 cells is under consideration, a problem that may arise is whether the number of individuals in each cell are large enough to draw statistically reliable statistical inferences. Such problems suggest that an investigator should explore the data to estimate the genotypic frequency of each genotypes as well the frequencies of each allele at the four loci under consideration. Similar questions will arise whenever an investigator wishes to explore the data to determine whether there is a sufficient number of observations in each cell to draw reliable statistical inferences, when N, the number of loci under consideration, is such that $N > 4$. A step that should be included in any exploration of the data would be that of determining if all loci under consideration were autosomal, for if one or more sex linked loci are included in the sample, then such loci would need to treated separately.

When all the loci are autosomal, one approach to determine the number of loci that are such that each cell in a multidimensional would have a sufficiently large number of observations to draw reliable statistical inferences is to investigate each locus under consideration. In this investigation one of the goals would be to determine, among other things, whether the frequency of the two alleles at each locus are sufficiently large enough to be included in the construction of arrays of data with respect to two or more loci that will contain a sufficient number of observations in each cell to draw reliable statistical inferences. For the case of the data on Altzheimer's disease mentioned above, an investigator would need to do an exploratory experiment involving 11 loci with two alleles at each locus. But, even in a sample of 5,000 individuals, the frequency of some alleles at one locus or two or more loci may not be sufficiently large to construct multidimensional arrays that involve low frequency alleles.

These observations suggest that it would be expedient for the above case of a sample of 5,000 individuals to estimate the frequency at each of the 11 loci to obtain information as to whether each allele at each locus has a sufficiently high frequency to be included in multidimensional arrays with respect to two or more loci. But, there are other criteria that could also be used to judge as to what loci would be included in multidimensional arrays. For example, if an estimate of

Ref

5. Raj, T., Shulman, J. M., Keenan, B. T. ,Lori B. Chibnik, L. B., Evans, D. A. Bennett, D. A., Stranger, B. E. and De Jager, P. L. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection DOI 10.1016/j.ajhg.2012.02.022. ˙2012 The American Society of Human Genetics.

© 2015 Global Journals Inc. (US)

heritability at some locus is low, then including this locus in a multidimensional array may not be fruitful. An investigator could also use estimates of the five effects that may be estimated when one autosomal locus is under consideration and carry out statistical test significance on the squares of the effects to judge which ones are statistically significant for each of the 11 loci. If there were loci for which none of squares of effects were not statistically significant, then an investigator may not want to include this locus in a multidimensional array involving two or more loci.

With regard to further developments of the software, it would be possible to create a front end to the APL programs to do the analyses reported in this paper so that the existing APL software could be used to carry out the type of exploratory experiment described above using any computer platform. But, before data consisting on multiple arrays involving two or more autosomal loci can be analyzed, an investigator would need to find either existing software or write software to accommodate multidimensional arrays of data on two or more autosomal loci. If APL were used to write this software, then the existing software for the case of two alleles at one autosomal locus could become part of the extended software when the additive effects and intra-locus effects are estimated at each of the loci under consideration. But, to estimate effects involving two or more loci, new programs would need to be written. It seems very plausible that an array manipulating programming language such as APL would be helpful writing succinct code designed to process multidimensional data on multiple loci. If an investigator wanted to consider one or more quantitative traits, then a modification of the ideas presented in chapter4 to accommodate the case in which only three genotypes per locus can be recognized, then the APL software used in this chapter for the case of one locus would need to be extended to handle two or more traits with respect to each locus. For those cases in which two or more loci and two or more traits are under consideration, an array manipulating programming such as APL would be very helpful in writing code to do the required matrix operation described in chapter 4. Just as the ordering of the three genotypes considered in this chapter which played a basic role in writing the software, some expeditious ordering of the genotypes with respect to two or more loci will also be a crucial step in developing computer code to accommodate cases of multiple loci with three recognizable genotypes at each locus.

Appendix

As an aid to developing a deeper understanding as to the properties of the absolute normal distribution that was used in Monte Carlo simulation experiments described in previous sections, in this appendix formulas for the expectation and variance of this distribution will be derived. In the literature on probability and statistics, the absolute normal distribution is called the folded normal and for the case $\mu = 0$ and $\sigma = 1$, this distribution is known as the half normal. The formulas the will be derived below may be found on the internet, and if a reader is interested in more details, it is suggested the web site: en.wikipedia.org/wiki/Folded_normal_distribution be consulted, where some references are also listed. Some proofs of the formulas derived below may also be found on the internet, but many of these proof lack transparency. In what follows, attempts will be made to included enough details with the hope that the derivation of the formulas will be transparent.

The first distribution to be described is the half normal. Let Z denote a normal random variable with expectation 0 and variance 1. In symbols $Z \sim N(0,1)$. The *pdf* of Z is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right] \tag{A.1}$$

© 2015 Global Journals Inc. (US)

for $z \in (-\infty, \infty) = \mathbb{R}$, the set of real numbers. The distribution function of Z is, therefore,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left[-\frac{1}{2}y^2 dy\right] \tag{A.2}$$

for $z \in \mathbb{R}$. Let $\mu \in \mathbb{R}$ and $\sigma \in [0, \infty)$. Then, by definition, a random variable $X = \mu + \sigma Z$ has a normal distribution with expectation

$$E[X] = \mu + \sigma E[Z] = \mu \tag{A.3}$$

and variance

$$var[X] = E\left[(X-\mu)^2\right] = \sigma^2. \tag{A.4}$$

The random variable $Y = \mid Z \mid$, the absolute value of Z, maps $\mathbb{R}$ into $[0, \infty)$,and has the distribution function

$$\begin{aligned} F(y) &= P[Y \leq y] = P[-y \leq Z \leq y] = \frac{1}{\sqrt{2\pi}} \int_{-y}^{y} \exp[-\frac{1}{2}y^2] dy \\ &= \frac{2}{\sqrt{2\pi}} \int_{0}^{y} \exp[-\frac{1}{2}y^2] dy = \sqrt{\frac{2}{\pi}} \int_{0}^{y} \exp[-\frac{1}{2}y^2] dy \end{aligned} \tag{A.5}$$

Therefore, the *pdf* of Y

$$f(y) = \frac{dF(y)}{dy} = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}y^2} \tag{A.6}$$

for $y \in [0, \infty)$. By definition, the expectation of Y is

$$E[Y] = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} y e^{-\frac{1}{2}y^2} dy. \tag{A.7}$$

But,

$$\int_{0}^{\infty} y e^{-\frac{1}{2}y^2} dy = -\lim_{y \uparrow \infty} e^{-\frac{1}{2}y^2} - \left(-e^{-\frac{1}{2}0}\right) = 1 \tag{A.8}$$

so that

$$E[Y] = \sqrt{\frac{2}{\pi}}\ . \tag{A.9}$$

From the definition of Y, it follows that $Y^2 = Z^2$ so that

$$E\left[Y^2\right] = E\left[Z^2\right] = 1 \tag{A.10}$$

because $Z \sim N(0,1)$. As is well known, the variance of Y may be expressed in the form

$$var[Y] = E\left[Y^2\right] - (E[Z])^2 \tag{A.11}$$

so that

$$var[Y] = 1 - \frac{2}{\pi} < 1. \tag{A.12}$$

The numerical value of this expression is

$$1 - \frac{2}{\pi} = 0.363\,38\ , \tag{A.13}$$

© 2015 Global Journals Inc. (US)

which will be helpful in the numerical evaluation of some formulas to follow.

When formulating a distribution so that it yields non-negative realizations of random variables in Monte Carlo simulation experiments, one approach would be to consider the random variable defined by

$$W = \mu + \sigma Y = \mu + \sigma \mid Z \mid . \tag{A.14}$$

From the above results, it can be seen that

$$E[W] = \mu + \sigma E[Y] = \mu + \sigma\sqrt{\frac{2}{\pi}} . \tag{A.15}$$

Furthermore, the variance of W has the formula

$$\begin{aligned} var[W] &= E\left[(W - E[W])^2\right] \\ &= E\left[\left(\mu + \sigma Y - \mu - \sigma\sqrt{\frac{2}{\pi}}\right)^2\right] \\ &= \sigma^2 E\left[\left(Y - \sqrt{\frac{2}{\pi}}\right)^2\right] \\ &= \sigma^2 var[Y] = \sigma^2\left(1 - \frac{2}{\pi}\right) < \sigma^2 . \end{aligned} \tag{A.16}$$

An advantage of this formulation is that the theoretical expectation and variance are easy to evaluate numerically. Thus, if the random variable defined in $A.14$ were used in testing a null hypothesis, as described in previous sections, the formulas for the expectation and variance in $A.15$ and $A.16$ could be used to estimate the expectation and variance of the random variable W.

From now on attention will be devoted to the folded normal distribution. In the Monte Carlo simulation experiments described in the forgoing sections of this paper is a primary task was to compute realizations of a phenotypic random variable W defined by

$$W = \mid X \mid \ , \tag{A.17}$$

where $X = \mu + \sigma Z \sim N\left(\mu, \sigma^2\right)$ .Some authors call the distribution of the random variable W the folded normal distribution. As a first step in finding the expectation of the random variable W, is to recall the definition of the function $\mid x \mid$ and observe that $\mid x \mid = x$ if $x \geq 0$ and $\mid x \mid = -x$ if $x \leq 0$. Let A denote the set

$$A = [z \in \mathbb{R} \mid \mu + \sigma z \geq 0] . \tag{A.18}$$

Equivalently,

$$A = [z \in \mathbb{R} \mid \mu + \sigma z \geq 0] = \left[z \in \mathbb{R} \mid z \geq \frac{-\mu}{\sigma}\right] . \tag{A.19}$$

Similarly, let B denote the set

$$B = [z \in \mathbb{R} \mid \mu + \sigma z \leq 0] = \left[z \in \mathbb{R} \mid z \leq \frac{-\mu}{\sigma}\right] . \tag{A.20}$$

Then, then the expectation of the random variable W may be represented in the form

$$E[W] = E\left[W \mid Z \geq \frac{-\mu}{2}\right] - E\left[W \mid Z \leq \frac{-\mu}{2}\right] . \tag{A.21}$$

© 2015 Global Journals Inc. (US)

Notes

Observe that

$$E\left[W \mid Z \geq \frac{-\mu}{\sigma}\right] = z\int_{\frac{-\mu}{\sigma}}^{\infty}(\mu+\sigma z)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}d$$

$$= \mu\left(\int_{\frac{-\mu}{\sigma}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz\right)+\sigma\int_{\frac{-\mu}{\sigma}}^{\infty}z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz \quad (A.22)$$

It can be seen that the coefficient of μ in $A.22$

$$\int_{\frac{-\mu}{\sigma}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = 1-\Phi\left(\frac{-\mu}{\sigma}\right). \quad (A.23)$$

Next, observe that

$$\int_{\frac{-\mu}{\sigma}}^{\infty}z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = -\left(-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\mu^2}{\sigma^2}}\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\mu^2}{\sigma^2}}. \quad (A.24)$$

Therefore

$$E\left[W \mid Z \geq \frac{-\mu}{\sigma}\right] = \mu\left(1-\Phi\left(-\frac{\mu}{\sigma}\right)\right)+\sigma\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\mu^2}{\sigma^2}}. \quad (A.25)$$

Next consider

$$E\left[W \mid Z \leq \frac{-\mu}{\sigma}\right] = \mu\left(\int_{-\infty}^{\frac{-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz\right)+\sigma\int_{-\infty}^{\frac{-\mu}{\sigma}}z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz. \quad (A.26)$$

By definition

$$\int_{-\infty}^{\frac{-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = \Phi\left(-\frac{\mu}{\sigma}\right), \quad (A.27)$$

and

$$\int_{-\infty}^{\frac{-\mu}{\sigma}}z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = -\frac{1}{\sqrt{2\pi}}e^{-\frac{\mu^2}{2\sigma^2}}. \quad (A.28)$$

By applying $A.21$ and doing the algebra, it follows that

$$\mu_W = E[W] = \mu(\left(1-2\sigma^2\Phi\left(-\frac{\mu}{\sigma}\right)\right)+\sigma\sqrt{\frac{2}{\pi}}e^{-\frac{\mu^2}{2\sigma^2}}. \quad (A.29)$$

Because $W^2 = X^2$, it follows that

$$E\left[W^2\right] = E\left[X^2\right] = \sigma^2+\mu^2. \quad (A.30)$$

The validity of this equation follows form the equation

$$E[W^2] = var[W]+(E[W])^2. \quad (A.31)$$

Therefore,

$$var[W] = \sigma^2+\mu^2-(\mu_W)^2. \quad (A.32)$$

© 2015 Global Journals Inc. (US)

From this equation, it can be seen that if $\mu > 0$ and large, then for every fixed $\sigma > 0$

$$\Phi\left(-\frac{\mu}{\sigma}\right) \approx 0 \tag{A.33}$$

and

$$e^{-\frac{\mu^2}{2\sigma^2}} \approx 0\ . \tag{A.34}$$

Therefore,

$$\mu_W = E[W] \approx \mu\ , \tag{A.35}$$

and

$$var[W] \approx \sigma^2 \tag{A.36}$$

is valid. These results are of interest, but under the condition listed above, the mean and variance of the folded normal distribution would be near that of the original normal distribution.

In all Monte Carlo simulation experiments reported in the previous sections on testing null hypotheses, σ was chosen as $\sigma = p\mu$, where $0 < p < 1$. In such cases

$$\Phi\left(-\frac{\mu}{\sigma}\right) = \Phi\left(-\frac{1}{p}\right) \tag{A.37}$$

If p is small, then $\Phi\left(-\frac{1}{p}\right) \approx 0$. For example, suppose $p = 1/4$. Then

$$\Phi(-4) \tag{A.38}$$

would be small, and if an algorithm were available to evaluated $\Phi(z)$ for any $z \in \mathbb{R}$, then the number in $A.38$ could be computed. Next observe that if $\sigma = p\mu$ then

$$\exp\left[-\frac{\mu^2}{2p^2\mu^2}\right] = \exp\left[-\frac{1}{2p^2}\right]. \tag{A.39}$$

In particular, if $p = 1/4$, then

$$\exp[-8] = 3.3546 \times 10^{-4} \tag{A.40}$$

It seems plausible, therefore, that for values of p such that $p \leq 1/2$ the approximations in $A.35$ and $A.36$ would be near the actual values of μ_W and $var[W]$.

There is another approach to approximating μ_W and $var[W]$ by using Monte Carlo methods and the law of large numbers. For example, let $W_1, W_2, \cdots, W_N$ be independent realizations of the random variable W. Then

$$\widehat{\mu}_W = \lim_{N\to\infty} \frac{1}{N}\sum_{\nu=1}^{N} W_\nu = \mu_W \tag{A.41}$$

with probability one. Similarly,

$$\widehat{E}\left[W^2\right] = \lim_{N\to\infty} \frac{1}{N}\sum_{\nu=1}^{N} W_\nu^2 = E\left[W^2\right] \tag{A.42}$$

with probability one. Therefore, if N is large, $N \geq 10,000$, then

$$\widehat{\mu}_W \approx \mu_W \tag{A.43}$$

and

$$\widehat{E}\left[W^2\right] \approx E\left[W^2\right]\ . \tag{A.44}$$

Hence,

$$\widehat{var}[W] = \widehat{E}\left[W^2\right] - \left(\widehat{\mu}_W\right)^2 \approx var[W]\ . \tag{A.45}$$

Notes

© 2015 Global Journals Inc. (US)

When these approximations are compared with the numerical value of μ_W in $A.29$ and that of $var[W]$ in $A.31$, then an investigator may judge how well the approximations in $A.43$ and $A.45$ are acceptable.

Lastly observe that there is another check on the correctness of formulas $A.29$ and $A.52$.For if $\mu = 0$ and $\sigma = 1$, then from $A.29$ and $A.32$, it follows that

$$E[W] = \sqrt{\frac{2}{\pi}}\,. \tag{A.46}$$

and

$$var[W] = 1 - \frac{2}{\pi}\,. \tag{A.47}$$

By definition if $\mu = 0$ and $\sigma = 1$, the random variable W has a half normal distribution with an expectation given by $A.46$ and variance $A.47$ which match the formulas in $A.9$ and $A.12$. This demonstration shows that the half normal distribution is a special case of the folded normal distribution as was expected.

Notes

References Références Referencias

1. Mode, C. J. and Gallop, R. J. (2008) A Review on Monte Carlo Simulation Methods as They Apply to Mutation and Selection as Formulated in Wright- Fisher Models of Evolutionary Genetics. Mathematical Biosciences 211: 205-225.
2. Mode, C. J., Raj, T. and Sleeman, C. K. (2011) Monte Carlo Implementations of Two Sex Density Dependent Branching Processes and their Applications in Evolutionary Genetics. Pages 273 - 296 in Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science. INTECH, an internet company, edited by C. J. Mode.
3. Mode, C. J. (2014-a) Estimating Effects and Variance Components in Models of Quantitative Genetics in an Era of Sequenced Genomes. Global Journal of Science Frontier Research-Mathematics and Decisions Sciences. Issue 5 Version 1.0 Online ISSN 2249-4626.
4. Mode C., J. (2014-b) Estimating Statistical Measures of Pleiotropic and Epistatic Effects in the Genomic Era. International Journal of Statistics and Probability. Viol. 3, No 2. ISSN 1927 -7032.
5. Raj, T., Shulman, J. M., Keenan, B. T. ,Lori B. Chibnik, L. B., Evans, D. A. Bennett, D. A., Stranger, B. E. and De Jager, P. L. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection DOI 10.1016/j.ajhg.2012.02.022. ˙2012 The American Society of Human Genetics.
6. Rossin E. J., Lage K., Raychaudhuri S., Xavier R. J., Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7(1): e1001273. doi:10.1371/journal.pgen.1001273 InTech-ISBN 978-953-307-427-6.

© 2015 Global Journals Inc. (US)

This page is intentionally left blank

© 2015 Global Journals Inc. (US)