Letters to the Editor

Am. J. Hum. Genet. 77:513–514, 2005

On Rapid Simulation of *P* Values in Association Studies

To the Editor:

In the March issue of the *Journal*, Seaman and Müller-Myhsok (2005) proposed a method for rapid simulation of *P* values in association studies. The authors kindly discussed my article (Lin 2005), which was electronically published in September 2004. Unfortunately, their discussion is inaccurate. In particular, their assertion that the variance formula given in my article ignores the variation due to the estimation of nuisance parameters is untrue.

Both my article (Lin 2005) and that of Seaman and Müller-Myhsok (2005) are based on the simulation of the (same) joint distribution for a set of test statistics, although the actual simulation procedures are somewhat different. If a statistic is asymptotically normal, then it can be approximated by a sum of independent terms. Specifically, the statistic for testing the jth null hypothesis H_j:$\beta_j = 0$ can be written as

$$U_j = \sum_{i=1}^{n} U_{ji} , \qquad (1)$$

up to an asymptotically negligible term, where U_{ji} involves only the data from the ith subject, and n is the sample size. Under H_j:$\beta_j = 0$ ($j = 1, \ldots, J$), the set of statistics $(U_1, \ldots, U_J)$ is asymptotically multivariate zero-mean normal with covariance

$$V_{jk} = \sum_{i=1}^{n} U_{ji} U_{ki}^T \qquad (2)$$

between U_j and U_k ($j,k = 1, \ldots, J$). In my article, I proposed to simulate this joint distribution by

$$\tilde{U}_j = \sum_{i=1}^{n} U_{ji} G_i ,$$

where $G_1, \ldots, G_n$ are independent standard normal random variables, because $(\tilde{U}_1, \ldots, \tilde{U}_J)$ has the same joint distribution as $(U_1, \ldots, U_J)$. Seaman and Müller-Myhsok proposed to fit a joint model that includes all β_j terms and to simulate the joint distribution of $(U_1, \ldots, U_J)$ by a multivariate normal random vector with mean 0 and with covariance matrix $\{V_{jk}; j,k = 1, \ldots, J\}$. The two proposals simulate from (essentially) the same joint distribution and are both very fast. In particular, my proposal involves the evaluation of the $\tilde{U}_j$ terms, which is of the order nJ and which can be done in seconds or minutes, even for large values of n and J. One advantage of my proposal is that missing genotype data for one statistic do not affect any other statistics. (If the ith subject has no genotype data for calculating the jth statistic, then we simply set U_{ji} to 0. There is no need to impute missing data.) By contrast, the proposal of Seaman and Müller-Myhsok can include in the analysis only those subjects with complete genotype data on all the SNPs, unless all the missing genotypes are imputed. Imputation can adversely affect the type I error and power.

Seaman and Müller-Myhsok (2005) focused on the parametric statistics under generalized linear models, whereas my article (Lin 2005) covered all possible statistics, parametric or nonparametric. As described in the appendix of my article, all the commonly used association statistics can be written in the form of equation (1), in which U_{ji} is the ith subject's *efficient* score function for β_j. In the special case of parametric statistics,

$$U_{ji} = S_{\beta_j,i} - A_{\beta_j\alpha_j} A_{\alpha_j\alpha_j}^{-1} S_{\alpha_j,i} , \qquad (3)$$

where $S_{\beta_j,i}$ and $S_{\alpha_j,i}$ are the ith subject's score functions for β_j and α_j, α_j is the set of nuisance parameters, and $A_{\beta_j\alpha_j}$ and $A_{\alpha_j\alpha_j}$ are the appropriate submatrices of the limiting Fisher information matrix for β_j and α_j. As mentioned in the appendix of my article, this expression can be found in mathematical statistics texts, such as that by Bickel et al. (1993, p. 28). It was also given as equation (A1) of Lin and Zou (2004). In this case, $n^{-1}\sum_{i=1}^{n} U_{ji} U_{ji}^T$ converges to $A_{\beta_j\beta_j} - A_{\beta_j\alpha_j} A_{\alpha_j\alpha_j}^{-1} A_{\alpha_j\beta_j}$, the limiting covariance matrix of $n^{-1/2} U_j$, and the joint distribution of $(\tilde{U}_1, \ldots, \tilde{U}_J)$ indeed provides a valid approximation to that of $(U_1, \ldots, U_J)$. Thus, Seaman and Müller-Myhsok's statement that my variance formula ignores the term $A_{\beta_j\alpha_j} A_{\alpha_j\alpha_j}^{-1} A_{\alpha_j\beta_j}$ is simply untrue. Had I used the wrong variance formula, the numerical results presented in my article would not have been sensible.

Seaman and Müller-Myhsok (2005) might have been

confusing score functions with *efficient* score functions. The score function for β_j involves the nuisance parameters α_j, which are replaced by $\hat{\alpha}_j$, the maximum-likelihood estimators of α_j under H_j:$\beta_j = 0$. To account for the extra variation caused by this estimation, we use the Taylor series expansion to express the score function for β_j (with α_j replaced by $\hat{\alpha}_j$) as a sum of independent terms, which is in the form of equation (1) with U_{ji} as given in equation (3), so that equation (2) provides the correct variance-covariance expression (Lin and Zou 2004). The efficient score functions U_{ji} involve the unknown parameters α_j. When α_j in U_{ji} is replaced by $\hat{\alpha}_j$, the resulting U_j, V_{jj}, and T_j are (essentially) the same as the $U_{\beta(l)}$, $V_{\beta(l)}$, and T_l given by Seaman and Müller-Myhsok (2005). Again, the framework of my article (Lin 2005) extends far beyond the parametric setting.

In fact, the parametric setting considered by Seaman and Müller-Myhsok (2005) does not demonstrate the full power of the simulation approach. In their setting, the calculation of each statistic is of the order n, so that the permutation test is very feasible, even for large values of n. There is a stronger case for the simulation approach when the calculation of each statistic is time consuming or when the null distribution cannot be properly generated by permutation, as discussed in my article (Lin 2005).

Incidentally, equation (2) in Seaman and Müller-Myhsok (2005) is confusing. The term in the middle is the score function for β, which is a function of α, whereas the term on the far right involves $\hat{\alpha}$ instead.

D. Y. Lin

Department of Biostatistics
University of North Carolina
Chapel Hill

References

Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation in semiparametric models. The Johns Hopkins University Press, Baltimore

Lin DY (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21: 781–787

Lin DY, Zou F (2004) Assessing genomewide statistical significance in linkage studies. Genet Epidemiol 27:202–214

Seaman SR, Müller-Myhsok B (2005) Rapid simulation of *P* values for product methods and multiple-testing adjustments in association studies. Am J Hum Genet 76:399–408

Address for correspondence and reprints: Dr. Danyu Lin, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu

0002-9297/2005/7703-0018$15.00

Am. J. Hum. Genet. 77:514–515, 2005

Reply to Lin

To the Editor:

We are very grateful to Lin (2005*b* [in this issue]) for pointing out that, contrary to what was written in our article (Seaman and Müller-Myhsok 2005), his variance formula (Lin 2005*a*) does take into account the estimation of nuisance parameters. We apologize to Lin and readers of the *Journal* for this error. As Lin supposes, we had failed to appreciate the difference between score functions and efficient score functions.

Both Lin's method (Lin 2005*a*) and our method (Seaman and Müller-Myhsok 2005) involve estimation of the same covariance matrix of the vector of score statistics, $U^T = (U_1^T, \ldots, U_J^T)$. However, our estimators of this matrix are not the same. Under the joint null hypothesis, H_0:$\beta_1 = \ldots = \beta_J = 0$, vector U is asymptotically multivariate normal distributed with mean zero and covariance $E(\sum_{i=1}^n U_{ji}U_{ki}^T)$ between U_j and U_k, where $U_{ji} = S_{\beta_j i} - A_{\beta_j\alpha_j}A_{\alpha_j\alpha_j}^{-1}S_{\alpha_j i}$. Lin (2005*a*) estimates this covariance by $V_{jk}^{\text{lin}} = \sum_{i=1}^n U_{ji}U_{ki}^T$. Let V^{lin} denote the matrix whose (j,k)th block element is V_{jk}^{lin}.

In our article (Seaman and Müller-Myhsok 2005), we considered tests derived from a single generalized linear model (GLM). Here, the covariance matrix for U can be estimated by $V^{\text{dsa}} = V_{\beta\beta} - V_{\beta\alpha}V_{\alpha\alpha}^{-1}V_{\alpha\beta}$, where $V_{\beta\beta}$ and so forth are submatrices of the Fisher information matrix of the GLM. Whereas Lin (2005*a*) uses the estimator V^{lin} and simulates from $N(0,V^{\text{lin}})$, we use V^{dsa} and simulate from $N(0,V^{\text{dsa}})$.

Let us examine V^{lin} and V^{dsa} for the GLM based on the binomial or Gaussian distribution. Assume that there are no environmental covariates; hence, the nuisance parameter vector, α, consists of just an intercept term. Let X_{ji} be individual i's locus score at locus j, $\bar{X}_j = \sum_{i=1}^n X_{ji}/n$, and $B_{jki} = (X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k)$. Let Y_i be individual i's trait value and $\bar{Y} = \sum_{i=1}^n Y_i/n$. The ($j$,$k$)th element of V^{dsa} is

$$V_{jk}^{\text{dsa}} = \frac{nR}{W}\sum_{i=1}^{n}\frac{B_{jki}}{n} \quad . \tag{1}$$

For the binomial GLM, $R = \sum_{i=1}^n (Y_i - \bar{Y})^2/n$ and $W = 1$. For the Gaussian GLM, $R = 1$ and $W = \sum_{i=1}^n (Y_i - \bar{Y})^2/n$. Now, $U_{ji} = (Y_i - \bar{Y})(X_{ji} - \bar{X}_j)/W$ and, hence,

$$V_{jk}^{\text{lin}} = \frac{1}{W^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2(X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k) \quad .$$

Note that, under H_0, $E(V^{\text{lin}}) = E(V^{\text{dsa}})$. We can rewrite V^{lin}_{jk} as

$$V^{\text{lin}}_{jk} = \frac{nR}{W}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 B_{jki} \Big/ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \ . \qquad (2)$$

From equations (1) and (2), it can be seen that V^{lin}_{jk} is proportional to a weighted mean of B_{jki} terms, with weights $(Y_i - \bar{Y})^2$, and V^{dsa}_{jk} is the corresponding unweighted mean. (This is also approximately true for the Poisson GLM, provided that $\text{Var}(Y) = E(Y)$; i.e., that there is no over- or underdispersion.) The weighted mean will have greater variance than the unweighted mean. Thus, $\text{Var}(V^{\text{dsa}}) \leqslant \text{Var}(V^{\text{lin}})$; that is, V^{dsa} is a more efficient estimator than is V^{lin}. The use of V^{dsa} should therefore produce more-reliable estimates of adjusted P values and product P values.

For a case-control study with equal numbers of cases and controls, weights $(Y_i - \bar{Y})^2 = 0.25$ are all equal, and $V^{\text{lin}} = V^{\text{dsa}}$. Thus, $\text{Var}(V^{\text{dsa}}) = \text{Var}(V^{\text{lin}})$. However, when the case:control ratio is $1{:}M$, $(Y_i - \bar{Y})^2$ equals $1/(M+1)^2$ for controls and $M^2/(M+1)^2$ for cases. That is, each case receives M^2 times as much weight as each control does. For a continuous trait, weights $(Y_i - \bar{Y})^2$ obviously vary. Computer simulations suggest that V^{dsa} is more efficient than V^{lin} also when there are environmental covariates.

It is true that a limitation of our method is that it cannot handle missing data, so missing genotypes must be imputed. This will lead to a conservative test if many genotypes are missing (assuming imputation does not use trait values). In this situation, the method of Lin (2005*a*) should be preferred. However, when there are few missing data, our method might be preferred because of the greater efficiency of V^{dsa}.

When there are no environmental covariates, V^{lin} can be adapted to yield an estimator that allows missing data while sharing the efficiency of V^{dsa}. We now derive this for the binomial and Gaussian GLMs. Analogous estimators exist for other GLMs. Let $C_j \subseteq \{1, \ldots, n\}$ be the set of individuals with complete data for test j. The covariance between U_j and U_k is $E(\Sigma_{i \in C_j \cap C_k} U_{ji} U_{ki})$, since $U_{ji} = 0 \forall i \notin C_j$. Lin (2005*a*) estimates this using

$$V^{\text{lin}}_{jk} = \sum_{i \in C_j \cap C_k} U_{ji} U_{ki}$$

$$= \frac{1}{W_j W_k}\sum_{i \in C_j \cap C_k}(Y_i - \bar{Y}_j)(Y_i - \bar{Y}_k) B_{jki} \ ,$$

where $\bar{Y}_j = \Sigma_{i \in C_j} Y_i / |C_j|$, $\bar{X}_j = \Sigma_{i \in C_j} X_{ji} / |C_j|$, and $B_{jki} = (X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k)$. For the Gaussian GLM, $W_j = \Sigma_{i \in C_j}(Y_i - \bar{Y}_j)^2 / |C_j|$. For the binomial GLM, $W_j = 1$. Under H_0, X_{ji} and Y_i are independent, and thus

$$E\left(\sum_{i \in C_j \cap C_k} U_{ji} U_{ki}\right) = \frac{1}{|C_j \cap C_k|}$$

$$\times E\left[\frac{1}{W_j W_k}\sum_{i \in C_j \cap C_k}(Y_i - \bar{Y}_j)(Y_i - \bar{Y}_k)\right]$$

$$\times E\left(\sum_{i \in C_j \cap C_k} B_{jki}\right) .$$

Hence, the new estimator is

$$V^{\text{new}}_{jk} = \frac{1}{|C_j \cap C_k|}\frac{1}{W_j W_k}$$

$$\times \sum_{i \in C_j \cap C_k}(Y_i - \bar{Y}_j)(Y_i - \bar{Y}_k)\sum_{i \in C_j \cap C_k} B_{jki} \ .$$

Note that, if $C_j \cap C_k = \{1, \ldots, n\}$ (i.e., there are no missing data), then $V^{\text{new}}_{jk} = V^{\text{dsa}}_{jk}$. The approach above also works for tests based on different traits. Let Y'_{ji} denote individual i's value for the trait variable of test j and let $\bar{Y}'_j = \Sigma_{i \in C_j} Y'_{ji} / |C_j|$. The term $(Y_i - \bar{Y}_j)(Y_i - \bar{Y}_k)$ in V^{new}_{jk} is replaced by $(Y'_{ji} - \bar{Y}'_j)(Y'_{ki} - \bar{Y}'_k)$, and $(Y_i - \bar{Y}_j)^2$ in W_j is replaced by $(Y'_{ji} - \bar{Y}'_j)^2$ (and similarly in W_k).

S. R. Seaman and B. Müller-Myhsok
Max-Planck Institute of Psychiatry
Munich

References

Lin DY (2005*a*) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21:781–787

——— (2005*b*) On rapid simulation of *P* values in association studies. Am J Hum Genet 77:3–513 (in this issue)

Seaman SR, Müller-Myhsok B (2005) Rapid simulation of *P* values for product methods and multiple-testing adjustments in association studies. Am J Hum Genet 76:399–408

Address for correspondence and reprints: Dr. Shaun Seaman, Max-Planck Institute for Psychiatry, Kraepelinstraβe 2-10, Munich 80804, Germany. E-mail: shaun@mpipsykl.mpg.de

0002-9297/2005/7703-0019$15.00