A Likelihood-Ratio Test of Twin Zygosity Using Molecular Genetic Markers

Stephen Erickson

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America

The importance of using multiple polymorphic genetic markers to determine unambiguously whether a twin pair is monozygotic (MZ) or dizygotic (DZ) has long been recognized. Concordance among a set of markers is used as evidence of monozygosity, as it would be improbable for DZ twins to be concordant at a large number of polymorphic loci. Several sources give a formula for the probability of two DZ twins sharing the same genotype at a locus, assuming knowledge of allele frequencies but not of either twin's genotype; this probability can be used to determine whether a set of markers will reliably distinguish between MZ and DZ status in a randomly selected twin pair. If the shared genotype is known, however, the likelihood-ratio test (LRT) of the null hypothesis of dizygosity against the alternative hypothesis of monozygosity takes into account the observed genotype and, by the Neyman-Pearson lemma, is the most powerful test of its size. The LRT is equivalent to conditioning on the genotype of one of the twins, and computing the probability, assuming DZ status, of the other twin sharing that genotype. The resulting *p* values are frequently lower than those produced by the unconditional probability, especially if rare alleles are observed. The unconditional probability can be recapitulated from conditional probabilities by averaging across all of the conditioned sibling's possible genotypes. To illustrate properties of the LRT applied to multiple markers, the probability distribution of the LRT *p* value is computed from allele frequencies of twelve unlinked markers published in Elbaz et al. (2006) and compared with the *p* value computed from unconditional probabilities.

The importance of using multiple polymorphic genetic markers to determine unambiguously whether a twin pair is monozygotic (MZ) or dizygotic (DZ) has long been recognized. Typically, a set of unlinked markers is genotyped for each twin, and concordance at all loci is taken as evidence of monozygosity. The significance of the observed concordance can be expressed quite naturally in terms of classical statistical hypothesis testing, with a null hypothesis of dizygosity tested against an alternative of monozygosity.

Several sources give a formula for the probability of a genotype match at a given locus between two DZ twins (e.g., Selvin, 1977; Becker et al., 1997; Nyholt, 2006). Nyholt states the probability as

$$M(\mathrm{DZ}) = \frac{1}{4}\left[\sum_{i=1}^{n} p_i^4 + \sum_{i=1}^{n}\sum_{j=i+1}^{n} (2p_i p_j)^2\right] + \frac{1}{2}\sum_{i=1}^{n} p_i^2 + \frac{1}{4}, \quad (1.1)$$

where $p_1,\ldots,p_n$ are the allele frequencies for that locus. This formula assumes knowledge of population allele frequencies, but not of the genotype of either sibling, and typically would be used to determine whether a set of markers will reliably distinguish between MZ and DZ status in a randomly selected twin pair.

Likelihood-ratio test of twin zygosity

Suppose, on the other hand, a pair of twins is observed to share the A_iA_j genotype at a locus. The likelihood-ratio test (LRT) statistic of the null hypothesis of dizygosity against the alternative hypothesis of monozygosity is given by

$$\begin{aligned}
&\Lambda(\mathrm{Sib1} = A_iA_j \,\&\, \mathrm{Sib2} = A_iA_j) \\
&= \frac{L(\mathrm{DZ})}{L(\mathrm{MZ})} \\
&= \frac{\Pr(\mathrm{Sib1} = A_iA_j \,\&\, \mathrm{Sib2} = A_iA_j \mid \mathrm{DZ})}{\Pr(\mathrm{Sib1} = A_iA_j \,\&\, \mathrm{Sib2} = A_iA_j \mid \mathrm{MZ})} \\
&= \frac{\Pr(\mathrm{Sib1} = A_iA_j \,\&\, \mathrm{Sib2} = A_iA_j \mid \mathrm{DZ})}{\Pr(\mathrm{Sib1} = A_iA_j)}.
\end{aligned} \quad (1.2)$$

Gaines and Elston (1969) refer to this ratio as the relative probability of monozygosity for concordant twins. This statistic can also be derived as the probability of a DZ genotype match, conditioned on one of the siblings' genotype:

$$\begin{aligned}
&\Pr(\mathrm{Sib2} = A_iA_j \mid \mathrm{DZ} \,\&\, \mathrm{Sib1} = A_iA_j) \\
&= \frac{\Pr(\mathrm{Sib1} = A_iA_j \,\&\, \mathrm{Sib2} = A_iA_j \mid \mathrm{DZ})}{\Pr(\mathrm{Sib1} = A_iA_j)}.
\end{aligned} \quad (1.3)$$

Received 16 July, 2007; accepted 24 October, 2007.

Address for correspondence: Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35233 USA. E-mail: erickson@uab.edu

https://doi.org/10.1375/twin.11.1.41 Published online by Cambridge University Press

Table 1

Genotype Transmission Probabilities

Parental genotype	Pr(PG)	Pr(A_1A_1 | PG)	Pr(A_1A_2 | PG)	Pr(A_2A_2 | PG)
(A_1A_1, A_1A_1)	p_1^4	1	0	0
(A_1A_1, A_1A_2)	$2p_1^3p_2$	½	½	0
(A_1A_1, A_2A_2)	$p_1^2p_2^2$	0	1	0
(A_1A_2, A_1A_1)	$2p_1^3p_2$	½	½	0
(A_1A_2, A_1A_2)	$4p_1^2p_2^2$	¼	½	¼
(A_1A_2, A_2A_2)	$2p_1p_2^3$	0	½	½
(A_2A_2, A_1A_1)	$p_1^2p_2^2$	0	1	0
(A_2A_2, A_1A_2)	$2p_1p_2^3$	0	½	½
(A_2A_2, A_2A_2)	p_2^4	0	0	1

Note: There is a row for each possible unphased parental genotype. Pr(PG) is the probability of observing a given parental genotype in a randomly selected parent-pair assuming Hardy-Weinberg equilibrium, while Pr(A_iA_j | PG) is the conditional probability of an offspring receiving the A_iA_j genotype under simple Mendelian transmission, $i, j = 1,2$.

If the twins are discordant at the locus, the null hypothesis of dizogosity is clearly not rejected, because MZ twins a priori must share all genotypes. If the twins are concordant, on the other hand, the size α test rejects the null when Λ greater than α, because it happens that Λ is exactly the probability of falsely rejecting the null, per (1.3). That is, Λ is not just the test statistic, but is also the p value itself. The Neyman-Pearson lemma, furthermore, ensures that this test is the most powerful of its size.

To compute (1.3), the denominator follows quite simply from allele frequencies. The numerator, however, can not be computed simply as the product of individual probabilities, because the genotypes for siblings are not independent; they depend on parental genotypes. For simplicity in computing the numerator, assume the locus is diallelic with alleles A_1 and A_2. Table 1 has a row for each possible unphased parental genotype. Pr(PG) is the probability of observing a given parental genotype in a randomly selected parent-pair assuming Hardy-Weinberg equilibrium, while Pr(A_iA_j | PG) is the conditional probability of an offspring receiving the A_iA_j genotype under simple Mendelian transmission. Conditional on the parental genotype, we assume Pr(A_iA_j | PG) is independent between non-MZ siblings, and therefore

$$\begin{aligned}&\Pr(\text{Sib1} = A_iA_j \,\&\, \text{Sib2} = A_iA_j)\\ &= \sum_{\text{PG}} \Pr(\text{PG})[\Pr(A_iA_j \mid \text{PG})]^2.\end{aligned} \tag{1.4}$$

Comparison of Unconditional and Conditional Probabilities

Figure 1 compares the conditional probabilities (1.3) computed in this manner to the unconditional probabilities (1.1) computed assuming knowledge of neither sibling's genotype. Probabilities are plotted as a function of the minor allele (A_2) frequency ranging from 0 to 1/2. If A_1A_1 is observed, the unconditional and conditional probabilities are comparable, and both probabilities approach one as the minor allele frequency approaches zero. If one or two minor alleles are observed, on the other hand, the conditional probability becomes much lower than the unconditional probability as the minor allele frequency approaches zero. If a rare allele is observed in a concordant twin pair, therefore, the LRT p value will be markedly lower than the p value computed from unconditional probabilities.

All of the probabilities are bounded from below by 1/4, which is the probability of alleles being identical by descent (IBD). At all allele frequencies, furthermore, the unconditional probability is within the range of the three conditional probabilities. Indeed, as one would expect, the unconditional probability is the weighted average of conditional probabilities

$$\begin{aligned}M(\text{DZ}) = \sum_{i,j} [&\Pr(\text{Sib2} = A_iA_j \mid \text{DZ} \,\&\, \text{Sib1} = A_iA_j)\\ &\times \Pr(\text{Sib1} = A_iA_j)],\end{aligned}$$

Multiple Loci

The previous results concern a single locus. When a set of multiple unlinked loci are concordant, however, locus-specific p values can simply be multiplied because of the probabilistic independence of unlinked markers. As an illustration, consider the set of twelve genotype frequencies published in Elbaz et al. (2006) shown in Table 2; the frequencies are those of the White non-Hispanic controls collected by the Nelson research team. Alleles are generically labeled and ordered so that A_1 is always the major allele, and SNP 9 is removed for simplicity's sake because it is located

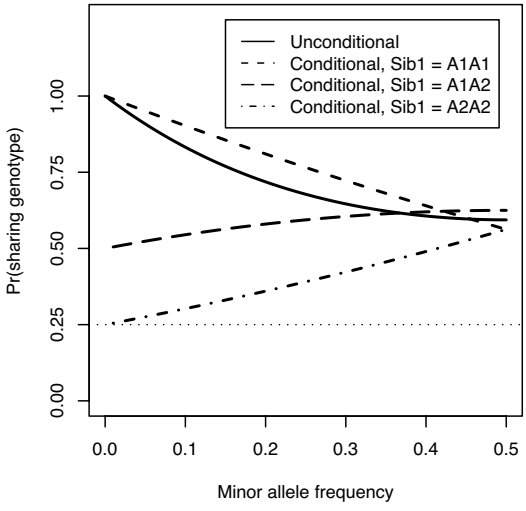

Figure 1

Probability of DZ twins sharing the same genotype as a function of minor allele frequency.

Note: Probabilities shown are conditioned on each of a sibling's three possible observed genotypes, as well as unconditional.

https://doi.org/10.1375/twin.11.1.41 Published online by Cambridge University Press

Table 2

Genotype Frequencies (published by Elbaz et al., 2006)

SNP	A_1A_1	A_1A_2	A_2A_2	$\hat{p}_{A_2}$
1	352	158	9	.170
2	424	94	2	.094
3	444	72	3	.075
4	391	118	9	.131
5	382	116	19	.149
6	416	96	7	.106
7	428	86	4	.091
8	366	132	22	.169
10	414	98	4	.103
11	329	173	16	.198
12	285	200	35	.260
13	284	197	35	.259

Note: The frequencies are those of the White non-Hispanic controls collected by the Nelson research team. Alleles are relabeled and reordered so that A_1 is always the major allele, and SNP 9 is removed for simplicity's sake because it is located on the X chromosome.

on the X chromosome. These data were chosen as an example of the variety of genotype frequencies within a certain population over a set of markers.

There are $3^{12}=531{,}441$ possible genotypes among this set of markers, so it is quite feasible to compute the probability distribution of the LRT *p* value for a randomly selected MZ twin-pair, weighting each possible *p* value by the corresponding genotype probability, assuming HWE, no mutations, and no genotyping errors. In figure 2, the distribution is shown in log (base 10) scale because of the skewed distribution of *p* values, and a vertical reference line corresponding to the unconditional probability of .047 is included. This reference line falls near the mode of the distribution, but the majority of the LRT distribution falls to the left (i.e., more significant) side. The median LRT *p* value is .039, while the unconditional probability of .047 is at the 59th percentile of LRT *p* values. In most cases, therefore, the LRT *p* value will be lower than one computed from unconditional probabilities. These results, which reflect actual SNP frequencies across several markers in a sizeable sample of a population, agree with observations made in the single-locus case.

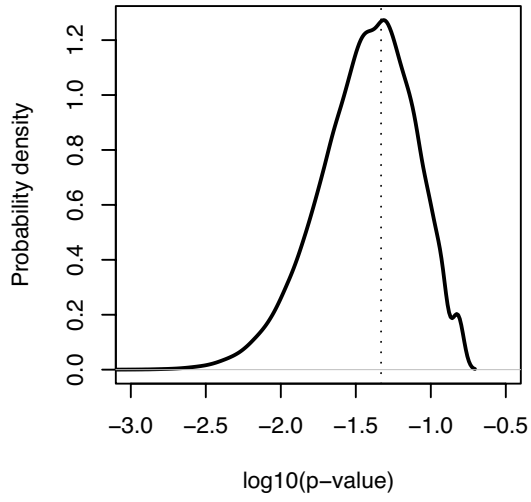

Figure 2

Distribution of the LRT *p* value for a randomly selected MZ twin-pair.

Note: LRT *p* values are computed per equation (1.2) and are shown in log (base 10) scale. The distribution shown is a kernel density estimate weighted by the genotype frequencies in Table 2, assuming HWE, no mutations, and no genotyping errors. The vertical reference line corresponds to the unconditional probability of .047.

Discussion

The likelihood ratio test described in this paper is a relatively straightforward application of the Neyman-Pearson lemma. I have shown that the LRT statistic Λ can also be derived as the probability of one DZ twin having the same genotype as the other twin, conditioned on the other twin's observed genotype. This interpretation makes intuitive sense and shows that Λ is the probability of type I error when rejecting the null hypothesis of dizygosity.

As shown in figures 1 and 2, the LRT is not guaranteed to yield a lower *p* value than one derived from unconditional probabilities. The unconditional *p* value will be lower when, for example, only major or common alleles are observed in a twin pair. The example based on genotype frequencies published in Elbaz et al. (2006), however, suggests that the LRT *p* value for most twin pairs will be lower than a *p* value computed from unconditional probabilities.

Acknowledgments

The author was supported by NIH grant 5T32HL 072757-04 and acknowledges the helpful comments of David B. Allison, Hemant Tiwari, and Michael C. Neale.

References

Becker, A., Busjahn, A., Faulhaber, H.-D., Bähring, S., Robertson, J., Schuster, H., & Luft, F. C. (1997). Twin zygosity: Automated determination with microsatellites. *Journal of Reproductive Medicine, 42*, 260–266.

Elbaz, A., et al. (2006). Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: A large-scale international study. *Lancet Neurology, 5*, 917–923.

Gaines, R. E., & Elston, R. C. (1969). On the probability that a twin pair is monozygotic. *American Journal of Human Genetics, 21*, 457–465.

Nyholt, D. R. (2006). On the probability of dizygotic twins being concordant for two alleles at multiple polymorphic loci. *Twin Research and Human Genetics, 9*, 194–197.

Selvin, S. (1977). Efficiency of genetic systems for diagnosis of twin zygosity. *Acta Geneticae Medicae Gemellolologiae, 26*, 81–82.

https://doi.org/10.1375/twin.11.1.41 Published online by Cambridge University Press