The Power of the Classical Twin Study to Resolve Variation in Threshold Traits

Michael C. Neale,1,3 Lindon J. Eaves,2 and Kenneth S. Kendler1,2

Received 3 Dec. 1993—Final 13 Feb. 1994

We explore the power of the twin study to resolve sources of familial resemblance when the data are measured at the binary or ordinal level. Four components of variance were examined: additive genetic, nonadditive genetic, and common and specific environment. Curves are presented to compare the power of the continuous case with those of threshold models corresponding to different prevalences in the population: 1, 5, 10, 25, and 30%. Approximately three times the sample size is needed for equivalent power in the continuous case when the threshold is at the optimal 50%, and this ratio increases to about 10 times when 10% are above threshold. Some power may be recovered by subdividing those above threshold to form three or more ordered classes, but power is determined largely by the lowest threshold. Non-random ascertainment of twins (i) through affected twins and examining their cotwins or (ii) through ascertainment of all pairs in which at least one twin is affected increases power. In most cases, strategy i is more efficient than strategy ii. Though powerful for the rarer disorders, these methods suffer the disadvantage that they rely on prior knowledge of the population prevalence. Furthermore, sampling from hospital cases may introduce biases, reducing their value. A useful approach may be to assess the population with a screening instrument; the power calculations indicate that sampling all concordant and half of the discordant pairs would be efficient, as along as the cost of screening is not too high.

KEY WORDS: Twin study; threshold traits; variance power; research design; ascertainment; sampling; selection.

INTRODUCTION

The classical twin study, which uses monozygotic (MZ) and dizygotic (DZ) twins reared together from birth to adulthood, is perhaps the most widely used behavior genetic research design. Although twins share prenatal and postnatal environments to a much greater degree than other relatives, the physical and psychological consequences of this sharing per se appear to be relatively minor, with the exception of birth weight and some verbal IQ measures (Record et al., 1970). Thus a strong advantage of twin studies is that twins are representative of the general population for a wide variety of phenotypes. Other assumptions of the method, e.g., that the degree of sharing of environmental factors is equal for MZ and DZ pairs, where tested, have almost always been found valid (Kendler et al., 1993b; Loehlin and Nichols, 1976).

Previous exploration of the statistical power of twin study has usually been limited to the continuous, normally distributed variables. Eaves (1969, 1972) and Martin et al. (1978) studied the power of the twin study to resolve additive (A) and non-additive (D) genetic and common (C) and specific (E) environment effects. Likewise, Heath and Eaves (1985) studied the power of twin-family and adoption studies to resolve cultural transmission and assortative mating. We know of only two brief reports of twin study power for ordinal data. First, Neale

1 Department of Psychiatry, Medical College of Virginia, Box 718, Richmond, Virginia 23298-0710.
2 Department of Human Genetics, Medical College of Virginia, Box 5, Richmond, Virginia 23298-0003.
3 To whom correspondence should be addressed.
et al. (1989) showed that sample sizes required for ordinal data were generally much larger than for the continuous case. Second, Neale and Cardon (1992) illustrated the methods used for power calculations for both the continuous and the binary variable case. This treatment was for didactic purposes and used only one example of a true model with 30% A, 20% C, and 50% E. Thus there is no published account of a detailed study of power for binary or ordinal twin data. Such data are common in the study of medical or psychiatric illnesses, where the presence or absence of disease may be the only possible assessment.

Our aim here is to calculate the power for variables that are measured at the binary or ordinal level. A comprehensive study of these factors would prove prohibitively long, so we limit our treatment to the exploration of several key issues. First, we vary the variance components of the population under study (the ‘true world model’). Second, the parameter (a, c, or d) that will be fixed to zero under the false model is changed. Third, we alter the threshold of affection, corresponding to various prevalence rates in the population. Fourth, the ratio of MZ-todZ pairs is varied from 2:1 to 1:2 for some simple cases. Fifth, we explore the advantages of ordinal data (with more than two ordered classes) over binary data. Sixth, the value of two schemes for nonrandom ascertainment is examined: probandwise (single selection) and pairwise (or truncate). The relation between these schemes and common research strategies is discussed. Finally, because the individual researcher may wish to explore other research designs or to find the power to test other specific hypotheses, we present sample Mx scripts (Neale, 1991) for power calculation.

**METHODS**

**Statistical Theory**

The basic theory behind power calculations has been described in detail in several earlier papers (e.g., Martin et al., 1978; Heath et al., 1985; Neale and Cardon, 1992). Essentially we want to know what chance we have of rejecting a particular false model, when some other true world model generated the data. As mentioned by Neale and Cardon, this probability, or power, will depend on at least six factors: (i) the effect being considered, e.g., \(a^2\) or \(c^2\); (ii) the size of effect in the true world; (iii) the probability level selected for rejection of the model; (iv) the sample size of the study; (v) the proportion of MZ-to-DZ twins in the sample; and (vi) the type of measurement—ordinal or continuous.

To compute the power to reject a particular hypothesis, e.g., that \(a^2 = 0\), we first generate data from the true world model. The second step is to fit the false model to these generated data and obtain a \(\chi^2\) goodness-of-fit of this model, which is known as the centrality parameter (\(\lambda\)) of the non-central \(\chi^2\) distribution. Tables of \(\lambda\) were published by Pearson and Hartley (1972) and Haynam et al. (1970); they indicate critical values for probability \(\beta\) of rejecting the model at significance level \(\alpha\) for a variety of degrees of freedom. \(\beta\) is known as the power of the test. Several computer programs will compute these values and will interpolate for values not found in the table (NAG, 1990; IMSL, 1987). Once the power of a given test is known, it is relatively straightforward to calculate the sample size required for a given level of power (\(\lambda_g\)), using the formula

\[ N^* = \frac{\lambda N_0}{\lambda_g} \]

where \(N^*\) is the sample size required, and \(N_0\) is the sample size of the generated data.

For complete data, we have two choices for calculating the centrality parameters. We could follow the procedure described by Neale and Cardon and use PRELIS (Jøreskog and Sörbom, 1993) to calculate tetrachoric correlations and their asymptotic weights and fit the (false) model by weighted least squares. Alternatively, we could generate contingency tables and fit the model directly to these data by maximum likelihood. For non-randomly ascertained samples we do not have this choice because PRELIS will not compute correlations for such data. However, we can still use the maximum-likelihood method as long as we correct for ascertainment, so for consistency we use this method throughout.

**Threshold Model**

For this study, we assume a threshold model based on an underlying normal distribution of liability (Falconer, 1960; Reich et al., 1979; Neale et al., 1986; Neale and Cardon, 1992, pp. 41-49). With ordinal data that have \(p\) categories, the model uses \(p - 1\) thresholds to subdivide the liability distribution. The expected proportion of individuals
Power of the Twin Study

in category $i$ is

$$
\int_{u-1}^{u} \phi(x) \, dx
$$

where $t_0 = -\infty$, $t_\phi = \infty$, and $\phi(x)$ is the normal probability density function, given by

$$
\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}
$$

We extend this model to the bivariate case to calculate the proportion of pairs of twins in each cell of the contingency table of twin responses. For example, if the response variable is binary, we need to compute expected proportions for the four cells of the $2 \times 2$ contingency table corresponding to the pairwise twin 1/twin 2 outcomes: no/no, yes/no, no/yes, and yes/yes. The expected proportion of pairs in cell $ij$ is,

$$
p_{ij} = \int_{u-1}^{u} \int_{v-1}^{v} \phi(x_1, x_2) \, dx_2 \, dx_1
$$

where $\phi(x_1, x_2)$ is the bivariate normal density function,

$$
[2\pi\Sigma]^{-\frac{n^2}{2}} \exp \left\{ -\frac{1}{2} (x_i - \mu_i)' \Sigma^{-1} (x_i - \mu_i) \right\}
$$

where $\Sigma$ is the population covariance matrix and $\mu_i$ is the (column) vector of means of the variables, and $|\Sigma|$ and $\Sigma^{-1}$, respectively, denote the determinant and inverse of the matrix $\Sigma$. In our application, we assume the means, $\mu_i$, to be zero, and the variances to be unity. Thus the expected proportions in the cells depend entirely on the thresholds ($t_i$ and $t_j$) and the correlation between the twins' liabilities. Note that the thresholds are abrupt, or "hard"; the ordinal categories are precisely—not probabilistically—related to the continuous liability distribution.

Under this bivariate normal liability model, twice the log-likelihood of the data is

$$
2\ln L_{(\text{model})} = \sum_{i=1}^{r} \sum_{j=1}^{c} 2n_{ij} \ln \left\{ \frac{n_{ij}}{n_{..}} \phi(x_1, x_2) \right\}
$$

(1)

where $n_{ij}$ is the observed frequency in cell $ij$, $p_{ij}$ is the expected proportion in cell $ij$, and $n_{..}$ is the total number of observations in the contingency table. Since $n_{..}$ is not estimated, the number of degrees of freedom associated with an $r \times c$ table is $rc - 1$. To compute an approximate $\chi^2$ statistic, twice the likelihood of the data under the model is subtracted from twice the likelihood of the observed data themselves, which is

$$
2\ln L_{(\text{data})} = \sum_{i=1}^{r} \sum_{j=1}^{c} 2n_{ij} \ln \left\{ \frac{n_{ij}}{n_{..}} \right\}
$$

(2)

The likelihood-ratio obtained by subtracting Eq. (1) from Eq. (2) yields a $\chi^2$ which is asymptotically unbiased, but it may be unsuitable for small samples or sparse contingency tables (Agresti, 1990). For the purposes of power calculation, we can make the sample size arbitrarily large to give relatively unbiased estimates.

Nonrandom Ascertainment

To deal with selected samples of twins, we use an ascertainment correction for the likelihood. Effectively, as we omit certain classes of person from observation, the likelihood of observing the remaining individuals increases. Mathematically, we wish to “normalize” the proportions in the remaining classes of individuals to sum to unity. This ascertainment proportion can be calculated by subtracting the proportions in all omitted classes from the total population proportion (i.e., 1.0). First, consider the case where all pairs in which at least one twin is above the threshold are sampled, which we call pairwise (Morton, 1982; p. 48; $\pi = 1$). The ascertainment correction is

$$
A_M = \int_{-\infty}^{t} \int_{-\infty}^{t} \phi(x_1, x_2) \, dx_2 \, dx_1
$$

(3)

where $t$ is the ascertainment threshold, $x_1$ and $x_2$ are the liability values of twins 1 and 2, and $\phi$ is the multinormal probability density function. The likelihood corrected for ascertainment would simply be the likelihood as obtained before [Eq. (1)] but divided by $1 - A$. Second, consider the case where we have a sample of probands and examine all their cotwins, and where the chance that a cotwin is also a proband is essentially zero. This we call probandwise ascertainment ($\pi \to 0$). Thus, where the “proband” is not affected, neither affected nor unaffected relatives will be observed. The ascertainment correction is therefore

$$
A_p = \int_{-\infty}^{t} \int_{-\infty}^{t} \phi(v_1, v_2) \, dv_2 \, dv_1 + \int_{-\infty}^{t} \phi(v_1, v_2) \, dv_2 \, dv_1
$$

(4)

$$
= \int_{-\infty}^{t} \phi(v_1) \, dv_1
$$

(5)
In practice, we may observe intermediate selection schemes (some proportion of the concordant pairs are both probands or $0 < \pi < 1$). We do not calculate the power for these cases, but note that they might be obtained through a four-group analysis of MZ and DZ pairs ascertained under each of the two methods.

**Continuous Variable Case**

For reference, we calculated the power available for the continuous variable case, using a documented script from the Mx manual (Neale, 1991). This script goes through three steps: (i) it generates MZ and DZ covariance matrices under the true model, (ii) it fits the false model to these data, and (iii) it computes the required sample sizes for certain specific levels of power. To a good approximation, we were able to recover the results presented by Martin et al. (1978) for these models. Some differences are expected because we summarized the data as covariance matrices, whereas Martin et al. (1978) used mean squares.

**Mx Scripts for Power Calculation**

All analyses reported in this paper were performed using Mx (Neale, 1991), which has special features for generating data, fitting models to (possible non-randomly ascertained) contingency tables, and power calculations. Appendix 1 contains a documented example script (Fig. A1) which generates the contingency tables for MZ and DZ twins. Appendix 2 shows the Mx script (Fig. A2) which reads the contingency tables generated by the script in Appendix 1 (Fig. A1) and then fits the false model to these data. The power=.05,1 option automatically calculates the power to reject the hypothesis at the .05 level of significance for a 1-df test. The software then calculates the total sample size required to reject the hypothesis at various levels of power.

**Simulated and Generated Data**

We draw a distinction between generated data and simulated data. When we use particular parameter values to calculate predicted covariances under the model, the variances of the MZ and DZ twins (both twin 1 and twin 2) are necessarily equal; we call these statistics generated data. The true model will always fit these data. In contrast, if we create a simulated pseudo sample by the use of a random number generator, there will be stochastic error in the variances and covariances, so there is a possibility of failure of the true model. This chance of failure also exists when we fit the model to contingency tables. For example, with generated data the contingency tables are necessarily symmetric so the fit of the true model would be perfect. However, with simulated data the model might fail because of asymmetry in the table.

To calculate the power curves in this paper, the $\chi^2$ found for the fit of the false model—which contains one fewer parameter than the true model—is appropriate for power calculation with 1 df. This is because we use generated data. To confirm the validity of our results, we also used simulated data for several cases. With a true world model of 50% additive genetic, 0% common environment, and 50% random environment, we used SAS (1988) to simulate four types of data. According to the results for continuous data, 90% power to reject (at $p=.05$) the false CE model would be obtained with approximately 430 pairs of twins. We first, therefore, simulated 215 pairs of MZ and 215 pairs of DZ twins and computed the covariance matrices. Second, we fitted the true (ACE) model and obtained the $\chi^2$ test of fit. Third, we fitted the false (CE) model and obtained its $\chi^2$. Finally, we computed the difference between these $\chi^2$ statistics. This experiment was repeated 1000 times. We found that in 893 cases, the difference $\chi^2$ exceeded the 3.841 critical value for .05 significance of the $\chi^2$ distribution with 1 df. Similar experiments (with total sample sizes of 3658, 710, and 468 pairs) were performed with the random, pairwise, and probandwise samples of binary outcome data, and we found the false model to be rejected in 896, 887, and 881 cases of 1000 trials, respectively. Although all four samples were slightly below the 900 expected, this pattern was not found when the experiment was repeated. These results show good agreement between the simulation and the generation methods of calculating power.

**RESULTS**

**Introduction**

We decided that graphs (created by SAS graph) would be the most informative and concise way to present the results. Required sample size is plotted on a logarithmic scale of sample size, because of the wide variation in the numbers. We recognize
that precise sample sizes are not available with this method, especially for larger numbers, so we have established a database of tables of the results which are available on request or through internet via anonymous ftp to sapphire.gen.vcu.edu (Directory "ftp/pub/powercat"). In all cases, we use the .05 significance level to define model rejection. The graphs we present plot only the curves for 80% power, which is commonly used in research proposals. To compute other power levels from these figures is, however, very easy, because almost all the tests have a single degree of freedom. To obtain sample sizes for 25, 50, 60, 70, 80, 90, 95, and 99% power the reported numbers given in the tables and figures should be multiplied by the constants .210, .489, .624, .786, 1, 1.339, 1.656, and 2.341, respectively. For 2 df, the corresponding constants are .234, .514, .645, .799, 1, 1.313, 1.603, and 2.221, respectively. For example, when the threshold gives a 10% prevalence, a random sample size of 3659 pairs would be required to reject the false CE model when the true world model is 50% A and 50% E. For 50% power the required sample size would be \(0.489 \times 3659 = 1789\).

Single-Threshold Model for Binary Data

Many studies of twins, especially those with clinical foci, collect data that constitute binary outcomes—the presence or absence of disease, for example. We confine our attention here to univariate analyses of these data, i.e., a single outcome variable measured on a sample of twin pairs. Initially, pairs are assumed to be sampled at random from the population under study, with equal numbers of MZ and DZ twins. We report results of altering the MZ:DZ sampling ratio in this section and consider nonrandom ascertainment of pairs below.

Rejecting the AE, CE, and E Models When the True World Is CE or AE

Figure 1 shows samples sizes required to reject the hypothesis that familial aggregation is entirely environmental in origin (CE) when the true model is AE. For the purposes of calculating power, we consider the true model to be an ACE model with C zero. Figure 1 shows curves indicating sample sizes for a variety of prevalences, obtained by adjusting the threshold value so that a certain proportion of the population is above threshold (e.g., prevalence .10 would mean that 10% of the population is above the threshold). Also shown is a reference curve drawn for the continuous variable case. The power with binary outcomes is greatest when there is a 50:50 ratio of those above and below threshold. Even in this most powerful case, about three times as many pairs of twins are required to reject the model as would be needed if continuous measures were used. This ratio increases with the distance of the threshold from the population mean. With a threshold of 1.282, giving a 90:10 ratio, some 10 times the sample size required for the continuous case is needed to obtain equivalent power.

Conversely, we may wish to find the power to reject the AE model when the true model is CE, which may be done by consulting Fig. 2. For equivalent MZ twin correlations in the true world there is more power to reject the CE model than the AE model. This makes sense because the DZ correlation is larger and therefore more information under the CE model than under the AE model. The result is in agreement with that for the continuous case reported by Martin et al. (1978).

Finally, we calculated the power to reject the
E model of no familial resemblance when the true world is AE or CE. Figures 3 and 4 show that there is much greater power to reject the E model than to reject the CE or AE models. For example, with a 5% prevalence and $a^2 = .4$, we need about 10,000 pairs to reject the CE model but only 1000 pairs to reject the E model.

**Fig. 2.** Curves illustrating the effect of prevalence and true common environment variance on the power to reject the AE model at the .05 level of significance. The lowest curve is for continuous data.

**Fig. 3.** Curves illustrating the effect of prevalence and true heritability on the power to reject the E model at the .05 level of significance. The lowest curve is for continuous data.

Rejecting the AE, CE, and E Models When the True World Is ACE or ADE

We now consider slightly more complex true world models, where three sources of variance are operating. Figure 5 shows the effect of varying the amount of true world $a^2$ on the power to reject the AE model. The left-most points on this graph correspond to the 10% prevalence line in Fig. 2. Thus, the number of pairs required to reject the AE model when $a^2 = .6$ and $c^2 = .3$ ($r_{MZ} = .9$, $r_{DZ} = .6$) is slightly less than 3000. When there is no additive genetic variation ($r_{MZ} = r_{DZ} = .3$), this number increases to about 5500. Although this contrast is itself quite dramatic, it should be noted that the lines in Fig. 5 have a relatively shallow gradient, so that the gain in power caused by the background $a^2$ variation is not great unless the change in $a^2$ is great.

Similar effects are noted when we try to reject the CE model in the presence of some $c^2$ in the true model (Fig. 6). As the background amount of common environmental variation increases, the number of pairs of twins required decreases. This is because, all other things being equal, larger correlations have smaller errors than small ones. Common-environment variation increases the MZ and DZ correlations by the same amount, consequently reducing their error and making it easier to discriminate between them. The gain in power is greater with increasing background $c^2$ variation than with background $a^2$ variation, because the latter does not increase the DZ correlation so much.

Even more pronounced gains in power are observed for the 2-df tests when the E model of no familial resemblance is fitted to data generated by an ACE model (Fig. 7). The detection of familial
resemblance with twin studies of binary outcome variables presents little difficulty for the moderate- to large-sized twin study.

It is well-known that the power to detect genetic dominance is low in the classical twin study. The power is extremely low when binary outcomes are under investigation, as shown in Fig. 8. For example, if the true world variation is 30% additive, 30% dominance, and 40% random environment, we would need about 20,000 pairs of twins to reject the AE model.

The MZ:DZ Sampling Ratio

In common with Martin and Eaves (1977), we note that the power to reject alternative models depends on the ratio of MZ-to-DZ twins. Figure 9 shows the effects of sample constitution and the type of hypothesis being tested on statistical power. The x axis in Fig. 9 represents the true world proportion of either additive genetic or common-environment variance, according to whether the CE or AE model is being rejected, respectively. For rejecting the CE model, the sample sizes are uniformly higher than for the AE model. There is relatively little effect of changing the sample constitution on the power to reject the CE model; departures from the 1:1 ratio in either direction decrease the power to approximately the same degree. To reject the AE model, there is more power if the DZ sample size is twice that of the MZ and less power when the ratio is reversed.

Multiple Thresholds

We now show how some of the loss of power incurred by dichotomizing the liability continuum may be regained by further subdivision of the sample. To illustrate this effect, we considered up to four thresholds simultaneously. Continuous data may be thought of as ordinal data with an infinite number of infinitesimally spaced thresholds. Thus the continuous case represents the limit to the advantages of subdividing the data into small classes, so we plot this in the figures for reference. The power
to reject the CE and the AE models for various pairwise combinations of thresholds is shown in Figs. 10 and 11. In both cases, the largest sample sizes are needed if the thresholds are at the 5 and 1% levels. Note how the lines tend to cluster into four groups: those involving lower thresholds of 25%, those involving lower thresholds of 10%, and the one with a threshold of 5%. Thus the power is driven by the placement of the lower threshold (nearest the 50% mark), so that there is relatively modest gain by adding thresholds more extreme than the one of primary interest.

Some further gain in power may be obtained by adding several thresholds. Comparison of Figs. 2, 11, and 13 shows that the sample size required to reject the AE model when the true model is 60% C and 40% E goes from 450 to 360 to 330 when we have thresholds of 25%, 25 and 10%, and 25, 10, and 5%, respectively. Similar results hold for the power to reject the false CE model, as may be seen by comparing Figs. 1, 10, and 12. Again, such increases in power from multiple thresholds are relatively minor compared to the placement of the lowest threshold.

Nonrandom Ascertainment

We considered two types of nonrandom ascertainment of twin pairs. First, a sample of all pairs in which at least one twin is above threshold. Such a sample might be obtained through complete national health records cross-matched against twin registry information. We call such ascertainment pairwise or multiple complete; it corresponds to an ascertainment probability (often called \( \pi \)) of unity. Another way to obtain this kind of sample might be with a screening instrument. Selection of pairs in which at least one is above a certain critical value on the screening instrument could yield a sample mathematically equivalent to one obtained through complete hospital records. Especially for traits with high thresholds (low prevalence), there is much more power than when sampling at random. Figure 14 shows curves for both the random and the pairwise ascertainment cases, for direct comparison. Note that the order of the curves is, for the most part, opposite to that of the random sampling case; more power is available for less prevalent traits. How-
Fig. 8. Curves illustrating the weakness of the classical twin study to resolve genetic dominance by rejecting the AE model at the .05 level of significance. Prevalence is 10%.

Fig. 9. Curves illustrating the effects of true heritability, true common environment, and ratio of MZ to DZ pairs in the study on the power to reject the CE or AE model at the .05 level of significance. Prevalence is 10%.

However, the power is still considerably less than for the continuous variable case.

The second scheme we examined involved only obtaining two cells of the contingency table, which we call *probandwise* sampling. Such a sample might be obtained if, for a given set of subjects found to be above threshold, we examine their cotwins. Importantly, the cotwins of these individuals are expected never to be probands themselves. Thus the ascertainment probability $\pi$ is close to zero. Figure 15 plots the sample sizes required to reject the CE model under this type of nonrandom ascertainment. We note that the decrease in required sample size is considerably greater than for the *pairwise* sampling considered above. *The increase in power is much greater for rarer traits, which approaches that of the continuous variable case.*

**DISCUSSION**

We have explored several issues in the use of binary and ordinal variables with the classical twin study, which we discuss in turn. Following this discussion, we consider the limitations of the methods used and the implications for research design.

**Twin Study Power**

There were few surprises in the effects of varying the true world model, or varying the false model to be tested, largely because these results map onto those found by Martin et al. (1978) for the continuous variable case. In brief, we found that there was more power to reject false models when the true world involved common environmental effects than when familial aggregation was genetic. False models involving no familial aggregation are relatively easy to reject, requiring fewer than 1000 pairs for most cases and often fewer than 100 pairs. Models involving an incorrectly specified source of resemblance (e.g., AE instead of CE) are typically difficult to reject when the outcome is binary. The problem is that the $a$ and $c$ parameters are highly correlated. The range of required sample sizes is quite dramatic: For the 5% prevalence case we may...
Fig. 10. Curves showing the power to reject, at the .05 level, the CE model when the data consist of three ordered classes. Different curves correspond to different combinations of thresholds. The lowest curves plots continuous data for reference. The line for .25 and .10 is superimposed on that for .25 and .05; likewise, the lines for .10 and .05 and .10 and .01 are indistinguishable.

Fig. 11. Curves showing the power to reject, at the .05 level, the AE model when the data consist of three ordered classes. Different curves correspond to different combinations of thresholds. The lowest curve plots continuous data for reference. Some superimposing of curves occurs, which can be discriminated at the right.

need as many as 20,000 pairs to reject the AE model when the true world is 20% $c^2$ and 80% $e^2$; when these percentages are reversed, some 400 pairs are sufficient. The power to reject an AE model when the true world model is ADE is especially low. Even when dominance variance is 80% and additive genetic variance is 10%, about 2500 pairs of twins would be needed to have 80% power to reject the false AE model at the .05 level.

Clearly, binary data are much less informative than continuous data. This is true even when the threshold is at the most informative 50% split; some two to five times as many pairs are needed to achieve a power equal to that of a study of continuous data. The picture gets markedly worse the farther the threshold from the mean; a study of a binary variable with 10% prevalence (such as major depression) has about one-tenth the power of its continuous counterpart. To take a more extreme clinical example, schizophrenia has a prevalence of about 1%. If familial aggregation were additive genetic, and MZ twins correlated .7, a random sample of about 9000 pairs of twins would be needed to reject the false CE model.

On the whole, there is not a great deal of power to be regained by subdividing those above threshold into subgroups. For example, an analysis of depressives classified as mild, moderate, and severe cases would not be much more powerful than one where mild, moderate, and severe were treated as a single affected group. Nevertheless, the 5–10% reduction in sample size may translate into quite large savings in a field study involving personal interviews or other costly procedures. These savings are minor compared to those that may be obtained through the assessment of milder cases. In reality, such efforts to increase power are tempered by the possibility that the various subgroups are not strictly ordered on the hypothesized underlying
Fig. 12. Curves showing the power to reject, at the .05 level, the CE model when the data consist of two to five ordered classes. Different curves correspond to different number of threshold and the lowest curve is for continuous data.

Fig. 13. Curves showing the power to reject, at the .05 level, the AE model when the data consist of two to five ordered classes. Different curves correspond to different numbers of thresholds and the lowest curve is for continuous data.

continuum of liability. While the methods provide, in principle, for some test of this assumption, power to detect departures from bivariate normality in ordinal data is low.

Review of the results using the nonrandom ascertainment seems at first to offer a reprieve for the researcher wishing to investigate the genetic architecture of a binary variable. When pairs are selected because at least one member of the pair is above threshold (pairwise ascertainment), the required sample sizes are only about double those needed for the continuous case. This result does not depend greatly on the prevalence of the variable in question. In fact, smaller samples are needed for more extreme thresholds, though there is not much difference between, say, 5 and 25% prevalence. With a more extreme threshold, we omit more of the large, uninformative cell containing the concordant below threshold pairs. This effect is magnified when selection is probandwise—based on one pair member rather than on both. Under this type of ascertainment, the power is close to that of the continuous variable case for quite large ranges of thresholds and genetic parameters values. We note, however, that these methods depend quite heavily on knowledge of the population threshold. The more extreme the trait, the larger the sample required to obtain a given confidence interval on the threshold. Yet the method assumes that the threshold is known without error! Ideally, the population threshold should be estimated from the same population using the same diagnostic criteria at the same time of study. These data could then be analyzed jointly, but the cost to collect this additional sample may defeat the object of nonrandom ascertainment.

Similar circular problems affect the cost effectiveness of the use of a screening instrument. The additional efficiency of the ascertained sample will be tempered by the cost of the initial screening. If admission to hospital is used as the screening criterion, this cost might be considered already born, but unfortunately the sample may not be representative of cases in the general population. Such unrepresentativeness is seen in Berkson’s bias, whereby
those with two disorders are more likely to present for treatment than would be expected given the presentation rates for the disorders alone. The study of comorbidity is considerably hampered by this problem. Clearly, both the screening instrument and the hospitalization criteria methods will be flawed if they are not 100% specific for the liability dimension.

Limitations

Certain aspects of our choice of models and methods limit the generality of the results. First and foremost is the use of the classical twin study. This design has some distinct limitations, particularly the confounding of the effects of age, assortative mating, common environment, \( g-e \) covariance, genetical dominance, and other forms of nonadditivity. The twin study should be regarded as a first step in the study of genetic architecture, rather than as a definitive way to apportion variation. Second, we have not considered an exhaustive set of univariate models. The effects of social interaction, which could lead to different prevalences in MZ, DZ, and non-twin samples, have not been modeled. Nor have we considered the effects of sex limitation or other forms of interaction. If these factors are not excluded by design, they will lead to a reduction of statistical power. For example, if we were to study both male and female pairs and discovered that these groups were heterogeneous with respect to the causes of variation ("nonscalar sex limitation"), then a larger total sample size would be needed to obtain equivalent statistical power within each gender.

Our analyses here are strictly univariate. Some increase in power might be obtained through the use of multiple indices, if these measures decrease measurement error on the factor of interest. Indeed, continuous measures are commonly derived from sets of items measured at the binary level. Yet different symptoms are often unreliable or invalid in-
dicates the trait of interest, and the number of valid indicators is very limited—frequently to reassessment of the same trait at a different time. A sense of the possible gain in power can be drawn by considering the twin correlations for the latent trait of interest, rather than the indices themselves. An example where this would apply is our study of test-retest data on depression (Kendler et al., 1993a). On the first wave, subjects were asked to report about depression during the 12-month period prior to interview. A year later, they were asked about not only the preceding 12-month interval, but also the 12 months prior to the first interview. Results suggested that depression was a highly heritable, moderately reliable disorder. Based on this higher heritability, the power to reject false models of no heritability would be much greater than that available at either time point alone.

Some statistical issues deserve discussion. First, we used likelihood ratio tests to compute our centrality parameters. For small samples, these statistics may not provide the most powerful or the least biased estimates of true power; it has been shown that chi-square tests asymptote faster than maximum-likelihood ones (Agresti, 1990). The same may be true for other methods of analyzing ordinal data, including logistic regression (Sham et al., 1994), asymptotically weighted least squares (Browne, 1984; Jöreskog and Sörbom, 1993; Neale, 1991; Neale and Cardon, 1992), and the DeFries–Fulker (1983) method. However, for most practical purposes our results will provide useful guidelines.

A second problem with using maximum likelihood is that extensions to the multivariate case rapidly become computationally expensive or impractical with current computer hardware and software. Extension to the bivariate case with binary data (requiring calculation of four-dimensional integrals) seems feasible, but quadrature becomes very CPU intensive beyond this point. Nevertheless, ML has the advantage of being able to handle non-randomly ascertained samples; we know of no software that will compute asymptotic weights for such data. Lee et al. (1992) report methods for dealing with data missing completely at random, but these are not appropriate for the systematic pairwise and probandwise strategies described here. Gibbs sampling (McCulloch, 1993) might be used to compute weight matrices under these circumstances.

Finally, the threshold model may well be wrong. Of course, no model is “correct” when applied to real data, but the departure from reality may be quite dramatic for the threshold model. It seems quite unlikely that there is not an abrupt threshold of disease, such that individuals infinitesimally below it are normal, while those infinitesimally above it have the full-blown disorder. More reasonable would be some soft threshold, although these are comparatively rarely used. Nevertheless, the abrupt threshold model may give fairly accurate results if the soft thresholds have a relatively steep form. It is also possible that the idea of an underlying continuum of liability is incorrect; it may be that individuals actually belong to different classes, perhaps because of some major genetic or major environmental effect. Computation of power to test these and other alternative models is beyond the scope of this article.

Implications for Research Design

Clearly, it is to the investigator’s advantage to use continuous, normally distributed variables wherever there is a choice. Unfortunately, there is often no choice—there is no good quantitative index of schizophrenia, for example. For an epidemiologic study, very large samples will often be required, making studies with practical constraints or expensive protocols infeasible. However, the inclusion of continuous correlates of a qualitative phenotype is advantageous, and there may be no better place to develop improved quantitative indices than within behavior genetic designs. When continuous measures cannot be made, there is much more advantage to assessing a broader diagnostic category than there is to subdividing those above threshold. Of course, either approach critically depends on the assumption that the various subdivision are on the same continuum of liability. Unfortunately, power to test whether this is the case also seems low (Neale et al., 1994).

The problems with binary or ordinal outcomes are not limited to univariate analyses. Multivariate analysis of ordinal data presents special problems of its own, often requiring large sample sizes (Jöreskog and Sörbom, 1993). Because behavior genetic methods rely on multiple groups, these large sample size criteria have to be applied to each group.

Nonrandom ascertainment of twins can, in principle, lead to enormous gains in power. We have shown that probandwise ascertainment, which uses only half of the discordant pairs in the population, is more powerful than pairwise. Unfortu-
nately, while such strategies are initially attractive, they have several inherent weaknesses. The first is the reliance on an estimate of the population threshold. Inaccuracies here can lead to serious bias in the estimation of genetic parameters. For example, under probandwise ascertainment, with a true world model of 50% A and 50% E, if the population prevalence is 10% but we mistakenly fix it at 5%, the model estimates variance to be 42% A, 20% C, and 38% E. Conversely, if the true world prevalence is 5%, but we fix it at 10%, then the estimates become 32% A and 68% E. The corresponding figures for mistaking 1 for 2%, and vice versa, are 45% A, 12% C, 43% E, and 40% A and 60% E.

The second weakness with proband sampling is that if, for example, we were to ascertain a sample through hospital records, the probands might well be unrepresentative of individuals with the disorder in the population. In alcoholism, for example, there is some evidence to suggest that the genetic architecture of hospitalized cases may differ from that of cases in the general population (Kendler et al., 1992; Pickens et al., 1991; McGue et al., 1992). Such unrepresentativeness is a natural consequence of Berkson’s bias, whose presence will disrupt the study of comorbidity and symptoms common to several disorders. The method is also problematic for mild disorders, such as phobia, that typically do not require hospitalization. Obviously, both random and nonrandom sampling will have their limitations if liability is not distributed as a single continuous distribution but is a mixture of two or more dimensions. The discernment of heterogeneity in the extremes of a distribution would seem to need a suitable reference population, so at best a combination of random and nonrandom sampling might be used.

One useful scenario for epidemiological study may be to screen the population with a relatively inexpensive protocol and then select pairs for further study. However, the sensitivity and specificity of the screening instrument become important in this case. For a screen that is either 100% sensitive or 100% specific, it may be reasonable to assume that the screen has a threshold that is lower or higher than that of the clinical diagnosis but on the same scale. When neither of these special cases applies, some imperfect relationship with the liability dimension is apparent. Data collected in this way would need to be analyzed jointly with the screening instrument, so that the relationship between the screen and the clinical diagnosis can be modeled. As noted above, such multivariate analysis can be difficult or computationally intensive with ordinal data.

Finally, let us consider how sampling in practice maps onto the results presented here. Various research strategies led to the types samples explored in this article. First and foremost is random sampling, which has the enormous advantage of a built-in estimate of the population prevalence. Although limited by subject compliance and ease of tracing, the method is largely free of sampling bias and may be capable of detecting volunteer bias through the comparison of pairs concordant and discordant for participation (Neale and Eaves, 1993). Other strategies may give rise to effectively random samples. Suppose that (i) we had access to national database of hospital diagnoses for a disorder for which hospitalization is universal (or fully representative) and (ii) we could match this with a random or complete sample of twin births. We could then select all pairs in which at least one member was affected. If hospital diagnoses were considered sufficient criteria for epidemiological research, then such a study would require no further data collection and would be inexpensive. Despite the use of hospital records, for statistical purposes we can consider this as a random sampling strategy because the proportion of concordant unaffected pairs is measured.

Alternatively, consider the approach where all cases detected in a particular hospital or district are asked if they are a twin, and if they are, we assess the pair. Here several subcategories emerge, depending on the probability that the cotwin, if affected, will also be a patient at the hospital or catchment area. If that probability is unity, we have our “pairwise” model. If the probability is very small, then we have the “probandwise” model. In the latter case we avoid assessing half of the discordant pairs—those in which the affected member does not attend the hospital. Most epidemiological studies require assessment of the ascertained probands to ensure that they meet the researchers’ diagnostic criteria. This procedure transforms the initial hospital diagnosis into a form of screening instrument. The efficiency of sampling in this way will depend on the sensitivity and specificity of the hospital diagnosis to detect the researchers’ diagnostic criteria. The power calculations presented here are appropriate only if the hospital diagnosis is a perfectly sensitive screen of a random sample of cases from the general population.
APPENDIX 1

Figure 1A shows the Mx input script to generate MZ and DZ contingency tables for later use in power calculations performed with the script in Appendix 2 (Fig. A2). The first step uses two groups to compute the predicted cell proportions for MZ and DZ twins. Then two further groups are used to compute tables and write these out to the files power1.mz and power1.dz. See the text and comments in the script (following ! symbols) for details. At the time of writing, this file may be downloaded via anonymous ftp to sapphire.gen.vcu.edu, where it is scored as powercat1.mx in directory ftp/pub/mx/doc. The Mx software is also available at this site.

Categorical data power analysis: generate MZ Twin data

Data Ninput=2 Ngroup=4 ! Data line - input variables and groups
Ctable 2 2 ! Contingency table to get percentages
 0 0
 0 10000
Matrices
T Full 2 1 ! Population threshold matrix
G Full 1 3 ! Gamma matrix
P Symm 6 6 ! Phi matrix
I Iden 2 2
Thresholds T /
Covariance (IQ(\sqrt(G)))'P*(IQ(\sqrt(G)))' / ! Usual specification for G*P*G'
! sets up G matrix of the form a c e 0 0 0
! 0 0 0 a c e
MA G
.100 .000 .900 ! TRUE world values for a-2, c-2, e-2
MA P ! PHI matrix for MZ Twins
1
0 1
0 0 1
1 0 0 1
0 1 0 0 1
0 0 0 0 0 1

Fig. A1A.
start 0.6 all
start 1.645 T 1 1 T 2 1 ! Z score for hypoth. pop. prevalence
OU NONE

Categorical data power analysis: DZ Twins
Data NInput=2
CTable 2 2
n n
0 10000
Matrices
T Full 2 1 =T1 ! Same matrix as for MZ twins
G Full 1 3 =G1 ! Same matrix as for MZ twins
P Symm 6 6
I Iden 2 2
Thresholds T /
Covariance (I*(\sqrt(G)))*P*(I*(\sqrt(G)))'/
Matrix P ! Phi matrix for DZ twins
1
0 1
0 0 1
.5 0 0 1
0 1 0 0 1
0 0 0 0 1
Option NO_output
End

Calculate expected frequencies for MZ Twins
Data Calculation
Matrices:
F Full 2 2 =%p1 ! Expected proportions from group #1
A Full 2 2 ! Matrix to describe ascertainment
! of cells: 1=yes, -1=no
N FU 1 1 ! Matrix with MZ twin sample size
COMPUTE NO(A,F)/ ! Calculates expected frequencies
Fig. A1B.
Matrix N
10000 ! Hypothesised MZ twin sample size
Matrix A
1 1 ! random sampling here
1 1
Option MX%E=POWER1.MZ ! Save results to file
End

Calculate expected frequencies for DZ Twins
Data Calculation
Matrices
F Full 2 2 =%p2 ! Expected proportions from group #2
A Full 2 2 ! Ascertainment scheme matrix
N Full 1 1 ! Matrix with DZ twin sample size
COMPUTE N0(A,F)/ ! Calculates expected frequencies
Matrix N
10000 ! Hypothesised DZ twin sample size
Matrix A
1 1
1 1
Option MX%E=POWER1.DZ ! Save results to file
End

Fig. A1C.
APPENDIX 2

Figure A2 shows the Mx input script to perform power calculations using the contingency tables generated by the script in Appendix 1 (Fig. A1). The first two groups fit the false model to the contingency table data read in from the power1.mz and power2.dz files. The third group constrains predicted variance to equal unity, because variance and threshold values are confounded in these data. See the text and comments in the script (following ! symbols) for details. At the time of writing, this file may be downloaded via anonymous ftp to sapphire.gen.vcu.edu, where it is stored as powercat2.mx in directory ftp/pub/mx/doc.

```
Power Calculation - MZ Twins - (false) CE Model
Data NInput=2 NGroups=3
CTable 2 2
File=POWER1.MZ
Matrices
  G Full 2 6 ! Matrix of loadings on latent variables
  P Symm 6 6 ! Matrix of correlations among latent variables
  T Full 2 1 ! Thresholds
Thresholds T /
Covariances G*P*G' /
Specify G
  0 2 3 0 0 0
  0 0 0 0 2 3
Matrix P
  1
  0 1
  0 0 1
  1 0 0 1
  0 1 0 0 1
  0 0 0 0 1
Start 0.6 all
Start 1.645 T 1 1 T 2 1 ! Z score of 1 - hyp pop prevalence
Bound .00001 .99999 G 1 1 - G 2 6
```

Fig. A2A.
Power of the Twin Study

Option RS
End

Power Calculation - DZ Twins
Data NInput=2
CTable 2 2
File=POWER1.DZ
Matrices
G Full 2 6 =G1 !same matrix of loadings as for MZ's
P Symm 6 6 !different matrix of correlations among latent variables
T FU 2 1 =T1 !thresholds
Thresholds T /
Covariances G*P*G' /
MA P
1
0 1
0 0 1
.5 0 0 1
0 1 0 0 1
0 0 0 0 1
Option RS power=.05, 1
End

Constraint Group to Ensure that a*a + c*c + e*e + l*l = 1
Data Constraint NInput=1
Matrices
S Full 1 3
I Iden 1 1
Covariance I - S*S' /
Specify 'S
0 2 3
Option none
End

Fig. A2B.
ACKNOWLEDGMENTS

We are grateful for support from PHS Grants MH-40828, HL-48148, MH-45268, MH-49492, MH-4953, AA-09095, and RR 08123. Special thanks go to John Myers for computing, tabulating, and graphing the results.

REFERENCES


Edited by N. G. Martin