Estimating the statistical power to detect population subdivision using mitochondrial DNA

Barbara L. Taylor, Susan J. Chivers and Andrew E. Dizon
Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, CA 92038 U.S.A.

Abstract

We develop a technique to estimate statistical power for detecting population subdivision using hypothesis testing. Case-specific simulations are used to capture the spatial relation and abundances of the putative populations. The actual level of genetic differentiation between neighboring populations varies through time because of genetic drift. We capture this uncertainty about the level of population differentiation, which is the effect size, by sampling the populations through time. Each time period a p-value is calculated for a series of statistics used to detect population subdivision. Statistical power is the proportion of time that the null hypothesis of panmixia was correctly rejected for a given α-level. Results are presented as Type 1 versus Type 2 error tradeoff curves, which does not necessitate the researcher choosing an α-level.

Introduction

Conservation biologists are often interested in defining population structure when mixing between populations is low from a demographic point of view (for example around 1% per year) but is high from an evolutionary point of view. At this level, differences between populations are not fixed genetic differences but rather small differences in gene frequencies. Because analytical tree building methods provide no resolution at this level of population mixing, researchers usually test hypotheses about population structure: null hypothesis (H_0)--populations are panmictic (each individual has an equal chance of mating with any other individual in the population), alternate hypothesis (H_A)--there is population structure. Often the result is that H_0 cannot be rejected. In this case it is important to know the answer to questions like: "If dispersal between these areas was one percent per year, what is the probability I would have detected it?" This is a question of statistical power. This paper develops a method to estimate statistical power to answer such applied questions. For further discussion of the importance of calculating statistical power see Taylor & Dizon (1997).

Statistical power depends on three things: the levels of error chosen to define "significance", the effect size and the precision of the estimate. Terms used in hypothesis testing are presented in Table 1. Typically researchers choose a critical level to define significance of a 5% probability of falsely rejecting H_0 ($\alpha = 0.05$).

	Result of statistical test	
	Do not reject H_0	Reject H_0
H_0 is true	Correct decision made with probability 1 - α	Type 1 error made with probability α
H_0 is false	Type 2 error made with probability β	Correct decision made with probability 1 - β (power)

Table 1. Possible logical outcomes and types of statistical error when testing a null hypothesis H_0

In conservation biology the Type 1 error is the probability of under-protecting the resource and the Type 2 error is the probability of over-protecting the resource. Statistical power, the probability of correctly rejecting H_0 (1 - β), is not currently not calculated in studies of population structure.

The greater the difference between hypotheses, the easier it becomes to distinguish between the hypotheses (high power). This difference in magnitude between H_0 and H_A is called effect size. The expected effect size for genetic differentiation (F_{ST}) depends on effective population size (N) and dispersal (m) (Wright's formula (Wright 1931) modified for mitochondrial DNA [mtDNA] (Takahata & Palumbi 1985)):

$$F_{ST} = \frac{1}{2Nm + 1} \qquad (1$$

For an ideally panmictic population that is falsely subdivided into two strata $F_{ST} = 0$. For a population of 1,000 with one migrant per generation (Nm = 1), which is a dispersal rate that would allow populations to evolve, $F_{ST} = 1/3$. This level of genetic differentiation would be easy to detect statistically. However, for the same population dispersing at 1%/year with a generation time of 10 years, $F_{ST} = 0.005$, which would be much more difficult to differentiate from the panmictic case where $F_{ST} = 0$. Because we expect small differences in genetic differentiation with relatively high dispersal, we consequently expect low power. Note that differentiation also depends on abundance. To obtain the same level of differentiation ($F_{ST} = 0.005$) for a population of 10,000 would mean that dispersal would need to be ten times lower (an annual dispersal of 0.1%/year). Power also depends on the precision of our estimate, which in turn depends on sample size. Because we expect low power we can also expect that we will need high precision (high sample size) to have a good chance of detecting the small difference.

So, why isn't power routinely calculated? First, calculating power requires a specific alternate hypothesis, such as "dispersal is 1%/year". Many biologists do not know the level of dispersal relevant to their question about population structure and therefore fall back on a non-specific H_A: there is population structure. A non-specific H_A does not allow calculation of statistical power (Taylor and Gerrodette 1993; Taylor 1997; Taylor & Dizon, in press). A second problem is that equation 1 assumes that all populations are of equal size. Taylor et al. (in press) showed that unequal population size is important to the level of genetic differentiation, which is the expected effect size. Further, the simulations used to examine the effect of unequal population size revealed that the level of genetic differentiation varied, often dramatically, through time due to genetic drift (Fig. 1a). Thus, the effect size was not a single fixed value of F_{ST} (or the analogous ϕ_{ST} shown here) but rather a distribution of values (Fig. 1b). Because calculations of statistical power usually assume a fixed effect size it is not immediately obvious how to calculate power with a distribution of effect sizes. Clearly, statistical power would be different if $\phi_{ST} = 0.005$ than if it was 0.100. The problem is that the stochastic process of genetic drift makes either of those values possible for the same dispersal rate. Somehow we need to incorporate our uncertainty about the effect size into the calculation of statistical power. Note that neither distribution in Fig. 1b are Normally distributed. Fits of statistical distributions to such distributions are poor. This makes an analytical approach to incorporating uncertainty in effect size sub-optimal.

This paper solves that problem by using a simulation method to calculate power. The model allows populations to be of unequal size and accounts for temporal fluctuations in effect size. The technique simulates the hypothesized population structure with the dispersal rate of interest (called hereafter the critical dispersal rate) and calculates power for any statistic commonly used to investigate population structure (Hudson et al. 1992, Excoffier 1992, Roff & Bentzen 1989).

Methods

Our approach to estimating statistical power requires a case-specific simulation model designed to capture the most important dynamics that would influence gene frequencies. The primary reason for a case-specific approach is that the abundances of populations and their spatial relation to one another have large affects on gene frequencies that cannot be captured with analytical equations (Taylor et al., in press). For simplicity, we illustrate the technique with a simple model that considers only mtDNA. In principle, the same simulation technique to estimate power could be used for nuclear DNA, but the model would need to be much more complex to account for different mating structures and for the number, type and evolution of different loci. For example, even among microsatellites some loci are highly polymorphic while others have only a few alleles, which may indicate different tempos and perhaps different modes of evolution. The standard cautions apply concerning complex models that require the estimation of many parameters. The more uncertainty that is built into the model, the more uncertain the interpretation of the data will be. However, in principle any model that scientists feel captures the essential dynamics of their case could be used to estimate power.

Our illustration model is a birth and death Monte Carlo model (Appendix). The model is a stepping-stone model that allows annual dispersal to nearest neighbors. We chose a dispersal rate of 1%/year because it is a difficult case to detect population structure. Initially all individuals in the five populations had a single haplotype. We ran the model until the distributions for a number of parameters remained essentially constant: haplotypic diversity (H_T), the number of haplotypes and the statistical measures of genetic differentiation. This stochastic equilibrium had occurred after 100,000 years. The length of time required will depend on generation time (four years in our case) and the mutation rate ($\mu = 0.0001$ for each of 40 variable sites).

Once populations were in stochastic equilibrium we gathered genetic data every 25 generations (100 years). In principle, one could sample every year and, as long as one sampled over a very long time period, the distribution of effect sizes should not be biased. However, because simulations are very computer intensive and the calculation of statistics using randomization techniques is slow, we found periodic sampling to be a more practical approach. This temporal sampling is the real innovation in our method. Sampling through time allows us to incorporate the variability in actual genetic population differentiation caused by genetic drift. Thus, we have managed to incorporate a distribution of effect sizes into our calculation of power.

At each discrete time interval we gathered the following data: haplotype frequencies, haplotypic diversity, the actual measures of population differentiation (χ^2, H_{ST}, F_{ST}, K^*_{ST}, ϕ_{ST}), and the p-values for those measures for different sample sizes (n = 20, 40). Because this is a methods paper, we will only present results for F_{ST} and evaluate the comparative performance of the statistics in a separate paper (Taylor et al. SC/F2K/J5). For each statistic of differentiation, the p-values were estimated by performing 5,000 randomizations (Hudson et al. 1992). Thus, the null distribution for panmixia was formed by randomly assigning each individual to either population A or population B and calculating the statistics of differentiation. The p-value was the proportion of this null distribution that was equal to or greater than the observed value calculated for the sampled individuals.

The simulation was run for 50,000 years yielding 500 sets of statistics. Statistical power is calculated as the proportion of time that H_0 is correctly rejected. For example, Table 1 shows statistics for 10 time intervals for a single pair of populations. For $\alpha = 0.01$, statistical power = 0.1 because we would correctly reject panmixia in one out of ten cases (the last case when p = 0.008). Similarly, when $\alpha = 0.05$ power = 0.5 and when $\alpha = 0.10$ power = 0.6.

Time (years)	p-value for F_{ST}
100,000	0.260
100,100	0.135
100,200	0.190
100,300	0.097
100,400	0.034
100,500	0.026
100,600	0.421
100,700	0.038
100,800	0.034
100,900	0.008

Table 1. P-values for F_{ST} for 10 time periods for two populations with $n_1 = n_2 = 40$.

Because power is inversely related to the chosen α-level and the appropriate ratio of over- to under-protection errors is a policy decision, we present results not only at the standard $\alpha = 0.05$ but also in an α versus β (or Type 1 versus Type 2) trade-off curve. For this curve, the proportion of time H_0 is incorrectly not rejected (β) is recorded for each α-level.

We performed a double-sample check procedure to ensure that both our sampling procedure and our randomization process were producing unbiased estimates of p-values for the statistics. At each time step when we sampled the populations, we also sampled the same population twice. In this case, the null hypothesis is true, i.e. the strata are artificial and only a single population exists. Thus, we expect that 5% of the time, we will get p-values less than 0.05. That is, when we set $\alpha = 0.05$ and expect to falsely reject H_0 5% of the time, that we will actually do so in our simulation procedure. P-values when H_0 is actually true should be uniformly distributed between zero and one.

Results and Discussion

Tradeoff curves show the expected increased power (decrease in Type 2 error) for a given α-level (Type 1 error) with increased sample size (Fig. 2). Presentation of results in a tradeoff curve allows the scientist to completely summarize the statistical results of their study without the need to make a value judgement on the appropriate tradeoff between Type 1 and Type 2 errors.

Consider, for example, if only results for $\alpha = 0.05$ were given. For n = 20 and 40, β would be 0.34 and 0.19 respectively (corresponding to statistical power estimates of 0.66, and 0.81). Thus, by choosing $\alpha = 0.05$ the researcher has chosen to be more willing to commit an under rather than over-protection error by 6.8 times (0.34/0.05), 3.8 times for n = 20 and 40 respectively. A tradeoff curve allows the manager to later choose a ratio of Type 1 to Type 2 errors. For example, if a 1:1 ratio was chosen and samples sizes of 20 were available for each population, the critical α-level would be 0.19 and decisions would be made knowing that power = 0.81. In such a case, H_0 would be rejected if a p-value was less than 0.19 (say 0.15).

The results of the double-check method revealed that in most cases the distribution of p-values when H_0 was true was uniform between zero and one. We did find some interesting exceptions. In cases with both low sample size ($n \leq 20$) and where rare haplotypes were present (at levels where it would be likely in a random draw to sample none of that haplotype for at least one population) there were more p-values than expected between 0.95 and 1.00. Thus, although the randomization procedure does partially correct for low cell size problems for statistics like χ^2, and definitely preforms better than a standard χ^2, small biases are still present. Fortunately, these biases are very small within the range of concern (p-values < 0.5) so from a practical biological standpoint they are not important.

Usually a paper presenting a new method would present verification of the method. This is not possible in this case, other than our check for bias in estimating p-values, because the quality of the estimate of power is primarily related to model choice. In our case, as long as we sampled over a sufficiently long period, the estimate of power faithfully represents the conditions of our model. In real life, however, one must choose a model to represent an actual biological case. Taylor et al. (in press) showed that using analytical formulas is ill-advised for applied problems of dispersal estimation because assumptions that are always violated, such as all populations are of equal size, have important affects on dispersal estimates. However, there are many assumptions made for every model. For example, for simplicity we assumed that abundances remained fairly constant, that dispersal remained constant through time and did not depend on factors such as the density of neighboring populations, that there was a constant number of populations arranged spatially in a stepping-stone fashion, and that mutation was a constant probability and was equal for all sites.

The interpretation of genetic data will depend on how well these model assumptions capture the essential dynamics of the case of interest. As such, it is nonsensical to discuss whether our technique of estimating power is unbiased. The statistical technique of sampling through time, compiling a list of p-values and estimating power as the proportion of time H_0 was correctly rejected for a specific rate of dispersal is general and not affected by the model assumptions. However, the researcher will have to assess potential biases that the case-specific model might introduce for any particular application.

Using genetic markers to estimate dispersal will always be imperfect because our knowledge of the population history, the species' natural history, our initial hypotheses about plausible spatial structure and even our understanding of molecular evolution are all imperfect. Modelers should investigate the robustness of their conclusions to violations in model assumptions. Every model should be accompanied by a list of assumptions and state how altering those assumptions will change conclusions. It could be the case that although there is great uncertainty in model choice, any of the plausible models would result in the same conclusion and thus there is no uncertainty in the decision. In many cases it will be possible to design models to minimize how assumptions will affect conclusions. For example, one could choose to have all assumptions such that they would minimize effect size. One could then say that statistical power is **at least** the estimated amount. Thus, the more the scientist understands how the results of the study could be applied, the more the analysis can be tailored to produce results that are easily incorporated into management.

References

Dizon, A. E., C. A. LeDuc, and R. G. LeDuc. 1993. Intraspecific structure of the northern right whale dolphin (*Lissodelphis borealis*): the power of an analysis of molecular variation for differentiating genetic stocks. CalCOFI Rep. 35:61-67.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Frankham, R. 1995. Effective population size/adult population size ratios in wildlife: a review. Genetical Research 66:95-107.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9:138-151.

Roff, D. A. and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. 6:539-545.

Takahata, N. and S.R. Palumbi. 1985. Extranuclear differentiation and gene flow in the finite island model. Genetics 109:441-457.

Taylor, B. L. 1997. Defining "population" to meet management objectives for marine mammals. In Molecular Genetics of Marine Mammals. eds. A. E. Dizon, S. J. Chivers, and W. F. Perrin. Special Publication 3:347-364 Allen Press, Inc., Lawrence, Kansas, U.S.A.

Taylor, B. L. and A. E. Dizon. 1996. The need to estimate power to link genetics and demography for conservation. Conservation Biology 10:661-664.

Taylor, B. L. and A. E. Dizon. In Press. First policy then science: why a management unit based solely on genetic criteria cannot work. Molecular Ecology.

Taylor, B. L., S. J. Chivers, and A. E. Dizon. 1997. Using statistical power to interpret genetic data to define management units for marine mammals. In Molecular Genetics of Marine Mammals. eds. A. E. Dizon, S. J. Chivers, and W. F. Perrin. Special Publication 3:347-364 Allen Press, Inc., Lawrence, Kansas, U.S.A.

Taylor, B. L. and S. J. Chivers. SC/F2K/J5 Evaluating the performance of different statistics to detect population subdivision.

Taylor, B. L., S. J. Chivers, S. Sexton and A. E. Dizon. In press. Using simulation models that incorporate uncertainty to estimate dispersal rates from mitochondrial DNA data. Conservation Biology.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97-159.

Appendix: The simulation model

We model five sub-populations, linearly distributed in space in a stepping-stone fashion. The model differs in several ways from an analytical stepping-stone model. First, the model is stochastic: events, such as birth and death, are random, though the probability of these events is fixed. The annual probability of dispersal (d) is related to the migration rate/generation (m) as d = m/T. At carrying capacity (K), the probability of birth = probability of death = 0.2. This yields a generation time (T) of 4 years. Populations were not allowed to exceed K and excess individuals were randomly removed. Each individual consisted of a string of forty variable base pairs emulating the variable sites found in mtDNA of an abundant species of dolphin (*Lissodephis borealis*) (Dizon et al. 1993). Genetic distance was the number of base-pair differences between two sequences/number of variable sites (40). The site-specific mutation rate was 0.0001. Lower rates resulted in uncharacteristically low genetic diversity, which gave rather uninteresting results because most individuals in all sub-populations ended up with the same haplotype. Higher rates resulted in higher genetic diversity than observed in the real population. The model was initialized with all individuals having the same haplotype (the most common for *L. borealis*). The rate of transitions versus transversions was also estimated from *L. borealis* and incorporated as probabilities for sites that were already randomly chosen to have a mutation.

Each year the simulation stepped through the following sequence of events for each individual: 1) randomly determine whether the individual gives birth, 2) if the individual gives birth, determine the new haplotype allowing each site the possibility of mutating, 3) randomly determine if the new individual disperses to a neighboring population, 4) randomly determine whether the original individual survives to the next time step, and 5) if the individual survives, randomly determine whether the individual disperses. Note that this birth-and-death model treats all individuals equally, i.e. all are females with no effects of age. As such, the population is close to N_e except that abundance fluctuates slightly due to the random birth and death processes. Because populations were truncated at K, the true N_e will be slightly less than K. We do not attempt to translate from a census population size to N_e here, but in a real application this would be an additional source of uncertainty (Frankham 1995).

Populations at the ends of the distribution send approximately half their dispersers to their single neighbor and retain the other half. Simulations were run until the temporal distribution of population subdivision statistics stabilized. With the mutation rate used, stabilization occurred after approximately 100,000 years.

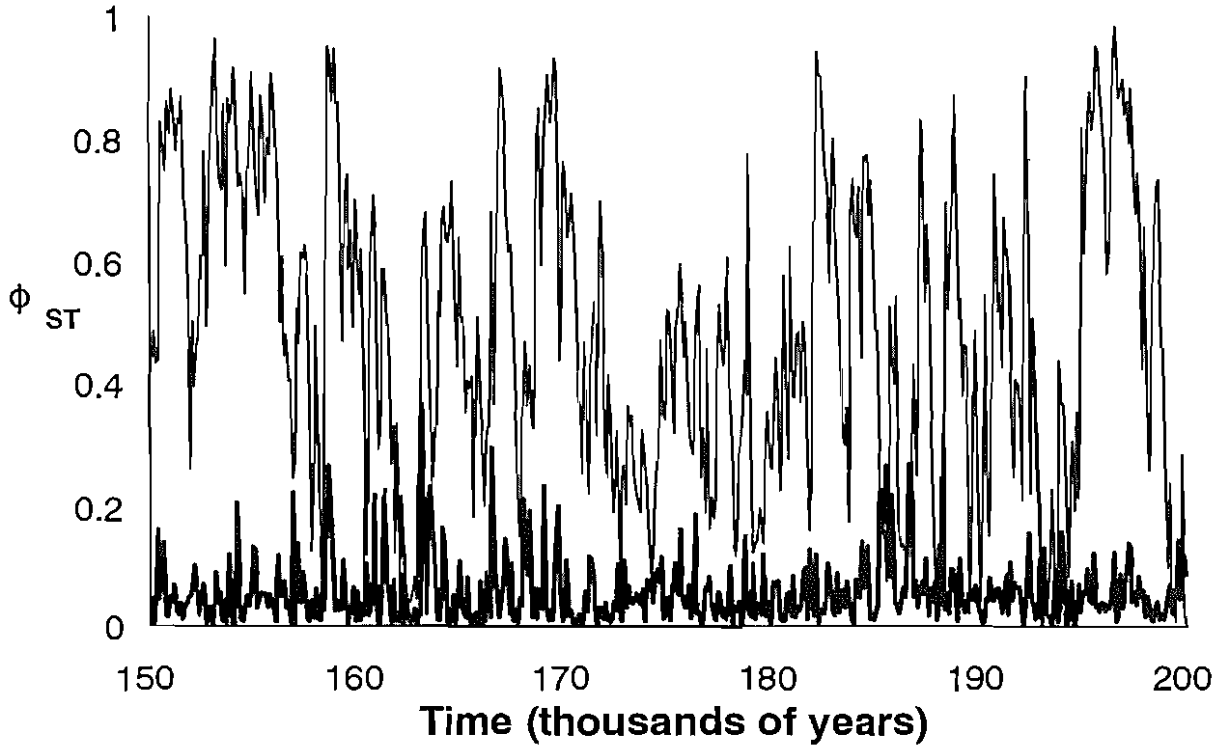

Figure 1a. The statistic ϕ_{ST} through time for a pair of populations both of effective population size of 100 and for dispersal rates of 0.0005 (thin line) and 0.01 (thick line).

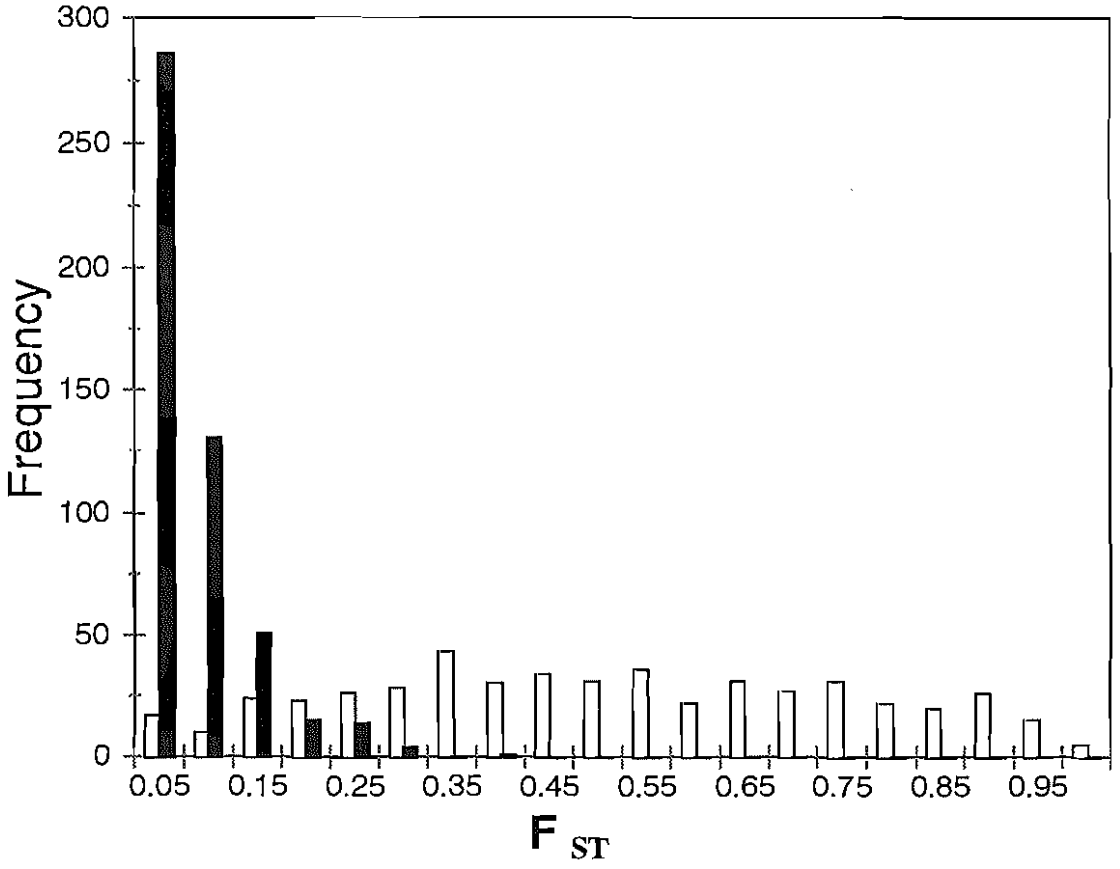

Figure 1b. Frequencies of the statistic ϕ_{ST} through time (seen in 1a) for dispersal rates of 0.0005 (clear bars) and 0.01 (black bars) for N = 100. Note that the calculation of ϕ_{ST} is based on all individuals in both populations and is therefore not an estimate but the observed level of genetic differentiation that is changing through time because of genetic drift.

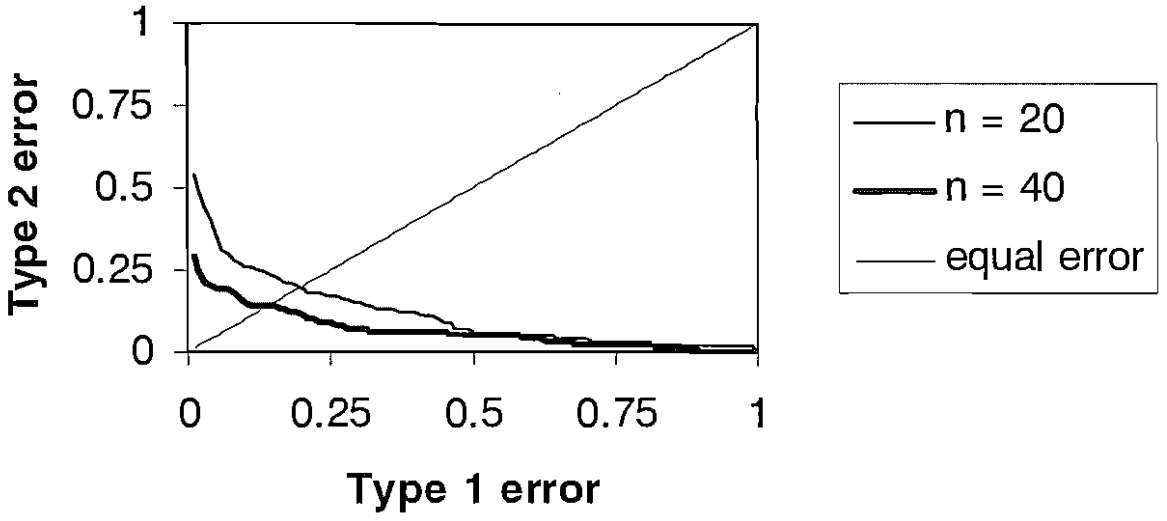

Figure 2. Tradeoff curve between Type 1 and Type 2 errors for a pair of populations both with an effective population size of 100 and a dispersal rate (d) of 1%/year, which for a generation time of 4 years is a migration rate/generation (m) of 4%/year (d = 0.01, m = 0.04). The "equal error" line indicates where Type 1 and 2 errors would be equal. Values on the curve below that line would have lower under than over-protection errors (i.e. be more conservative of the resource) than values above that line.