Forensic Science International: Genetics xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic Population Genetics - Original Research

A general approach to power calculation for relationship testing

Thore Egeland [a,b,*], Nadia Pinto [c], Magnus Dehli Vigeland [d]

[a] IKBM, Norwegian University of Life Sciences, Ås, Norway
[b] Norwegian Institute of Public Health, Oslo, Norway
[c] IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal
[d] Department of Medical Genetics, Oslo University Hospital, Oslo, Norway

ARTICLE INFO

Article history:
Received 11 December 2012
Received in revised form 2 April 2013
Accepted 3 May 2013

Keywords:
Forensics
Power of exclusion
Relationship testing
R
paramlink

ABSTRACT

This paper is motivated by power considerations in connection with relationship testing. Given the true relationship between a set of individuals, a claimed relationship between the same individuals, and a set of genetic markers, we compute the power of exclusion, i.e., the probability that the genotypes will be incompatible with the claimed relationship. If exclusion is impossible, as will be the case if it is required for instance to distinguish between sibs and half sibs, we rather obtain the distribution of the likelihood ratio. The problem we are addressing can also be seen as a standard way of measuring the ability of a battery of tests to resolve claimed family relationships. In particular, simple exclusion probabilities are regularly calculated worldwide as a part of designing forensic marker sets. Our approach to these problems is guided by a natural way of calculating exclusion probabilities on a computer. We present a user friendly implementation for this as part of the R package paramlink, originally designed by one of the authors (MDV) for pedigree manipulations and likelihood computations. By doing so we are able to handle problems more challenging than we have seen in the literature. Specifically, we deal with complex pedigrees with arbitrary inbreeding and conditioning. We present examples for autosomal as well as X-linked markers and some formulae to validate the results. The examples indicate a wide range of applications. Details are presented for an immigration case where previously reported calculations are extended to account for possible inbreeding and known genotypes. The supplementary material includes a tutorial on how to perform these calculations in paramlink.

1. Introduction

Correctly specified relationships are of great importance in many research areas. In forensic science, particularly, relationship testing lies at the very heart of several applications, including paternity cases and identification problems following disasters. In other fields questions of kinship are often encountered as assumptions that need verification. For example, linkage analysis (for locating chromosomal regions involved in disease) relies on correctly specified pedigrees. Similarly, individuals recruited in association analysis are typically required to be unrelated, and this assumption can be tested against alternatives of distant family relationships.

The problem we are addressing resembles statistical power calculations performed before any data is present. Broadly speaking, the objective is to assess if the intended data collection will suffice for reliable conclusions. Such calculations may also give guidance on what further data should be collected.

For kinship problems, genotypes may be inconsistent with a hypothesised family relationship in which case the hypothesis is rejected. More briefly we refer to this as an *exclusion*. We disregard mutations, silent alleles and genotyping errors throughout this paper since then exclusion is impossible. In our applications we find it useful to distinguish between cases where exclusion of the claimed relationship is theoretically possible, and those where it is not. For example, paternity cases fall in the first category: If the alleged father have no alleles in common with the child, the paternity is excluded. In such cases we focus on the *Power of Exclusion* (*PE*), i.e., the prior probability of exclusion, given the true relationship. If *PE* is close to 1, this indicates that sufficient data will be available. If, on the other hand, the power is much less than 1, it may be necessary to genotype more family members than planned or the number of genetic markers may have to be increased.

In many cases, exclusion of the claimed relationship is theoretically impossible. For instance, no autosomal marker can with certainty reject sisterhood between two women. In these cases power analysis involves investigating the *Likelihood Ratio* (*LR*). If the alternative to sisterhood is that the women are unrelated, a *LR* = 100,000 is interpreted as follows: The genotype data is 100,000 times more likely assuming sisterhood as

* Corresponding author at: Norwegian University of Life Sciences, Norway. Tel.: +47 64966391.
E-mail address: Thore.Egeland@umb.no (T. Egeland).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.05.001

compared to the women being unrelated. Our calculations are typically done before genotype data is available and then we have to calculate the probability distribution of the *LR*. This distribution is often hard to obtain explicitly, leaving simulation as the preferred strategy. Based on the simulated *LR* values, we can check how often a specified threshold, say 100,000 is exceeded. If a sufficiently large fraction of simulated values, say 80%, exceeds the threshold, the power is deemed adequate.

The paper [1] is the earliest example dealing with power of exclusion we are aware of. The topic continues to be discussed and general forensic books including [2,3] provide extensive presentations. Regarding the distribution of the *LR*, [4] presents an alternative to simulation based on asymptotic approximation for unlinked markers whereas simulation is used for linked markers.

There appears to be a need for software that can compute *PE* and simulate *LR* in *general* cases, that is allowing for complex pedigrees with loops (inbreeding) for autosomal and X-linked markers. Moreover, some individuals may have been typed at the time of power analysis, requiring the calculation to be performed conditionally on these individuals. Our ambition for this paper has therefore been to formulate a general framework for power calculation in relationship testing and provide freely available software. The program we are presenting, `paramlink`, is an `R` package originally designed for parametric linkage analysis. Several functions have been added to the package specifically for the applications in this paper. The implementation does not require the user to be neither an R expert nor familiar with linkage analysis. Furthermore, the tutorial included in the supplementary material offers examples with detailed explanations.

2. Methods

Initially we consider one marker. There are N individuals with a claimed relationship described by a pedigree Ped_1. Typically there will be more than N individuals in Ped_1 as additional persons may be needed to define the family relationships. We assume that all N will be available for genotyping, but we allow for situations where some genotypes are known from previous genotyping (all analyses are then conditional on these). If the true relationship is described by Ped_2, the marker's power of exclusion is defined as:

$$PE = \Pr(\text{Ped}_1 \text{ incompatible with genotypes} | \text{Ped}_2).$$

For any genotype combination $g = (g_1, g_2, \ldots, g_N)$, and $i = 1, 2$, define

$$\begin{aligned} A_i(g) &= \Pr(g|\text{Ped}_i), \\ I_i(g) &= \begin{cases} 1 & \text{if } A_i(g) = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

A_i is the joint genotype distribution for pedigree Ped_i and I_i is an inconsistency indicator function. Both of these functions are naturally represented as N-dimensional arrays. Summing over all possible genotypes we find:

$$\begin{aligned} PE &= \sum_g I(\text{Ped}_1 \text{ incompatible with g}) \Pr(g|\text{Ped}_2) \\ &= I_1 \circ A_2, \end{aligned} \quad (1)$$

where the final expression means the sum of the entries in the entrywise product of the arrays I_1 and A_2. A simple paternity case serves to illustrate the notation: Individual X claims to be the father of individual Y (corresponding to Ped_1), and we consider testing this by genotyping the two using an autosomal SNP marker with equifrequent alleles A and B. If X in reality is unrelated to Y (corresponding to Ped_2), what is the power of the marker to reject the paternity? Although trivial to compute the answer by hand in this case, we illustrate Eq. (1) by writing out the matrices:

$$I_1 = \begin{array}{c|ccc} & AA & AB & BB \\ \hline AA & 0 & 0 & 1 \\ AB & 0 & 0 & 0 \\ BB & 1 & 0 & 0 \end{array} \qquad A_2 = \begin{array}{c|ccc} & AA & AB & BB \\ \hline AA & 1/16 & 2/16 & 1/16 \\ AB & 2/16 & 4/16 & 2/16 \\ BB & 1/16 & 2/16 & 1/16 \end{array}$$

It follows that the exclusion probability is $I_1 \circ A_2 = 1/8$.

The extension to K independent markers is straightforward. If PE_k denotes the exclusion power for marker k, the total power is

$$PE = 1 - \prod_{k=1}^{K} (1 - PE_k). \quad (2)$$

We proceed now to the category of cases where exclusion is impossible, i.e., $PE = 0$. For a given combination of genotypes $g = (g_1, g_2, \ldots, g_N)$ of a single marker, the likelihood ratio is defined by

$$LR = \frac{\Pr(g|\text{Ped}_1)}{\Pr(g|\text{Ped}_2)} = \frac{A_1(g)}{A_2(g)}. \quad (3)$$

This extends to K independent markers in the obvious way:

$$LR = \prod_{k=1}^{K} LR_k. \quad (4)$$

Based on the distribution of *LR*, the power can be assessed.

There appears to be no generally accepted convention linking *LR* values to verbal statements. There are, however, several suggestions. According to [3] (p. 40), the evidence in favor of Ped_1 is "very strong" if $LR > 100{,}000$ and "strong" if $LR > 1000$. The *LR* can be related to other numerical summaries of the evidence, including a *Bayesian* alternative. If we assign prior probabilities for Ped_1 and Ped_2, the posterior probabilities for the pedigrees can be expressed in terms of these priors and the *LR*. Assigning a prior probability of 0.5 to both pedigrees leads to $\Pr(\text{Ped}_1|data) = LR/(LR + 1)$. The mentioned thresholds of 100,000 and 1000 then correspond to posterior probabilities of 0.99999 and 0.999, respectively. There is also a link between the power of exclusion and likelihood ratio as $PE = \Pr(LR = 0)$.

Most of the genotype data is not available when power calculations are performed. The missing data is then simulated and the *LR* distribution is obtained. Based on this distribution, the power can be assessed. Specifically, the fraction of simulated values exceeding specified thresholds can be found and a conclusion can be drawn as to whether one should proceed with the planned data collection. The conventional requirement of 80% power in other areas of applied statistics then corresponds to requiring 80% of the simulated *LR* values to exceed the specified threshold.

2.1. Implementation

The `R` package `paramlink` (`http://cran.r-project.org/web/packages/paramlink`) provides various functions for likelihood based pedigree analysis, including parametric LOD scores, power analysis for linkage, genotype probability distributions, and many utility functions for plotting and manipulating pedigrees and markers. Likelihoods are calculated using the Elston–Stewart algorithm which works well with pedigrees of any size if the number of linked markers is small. The calculations have been validated by comparing to other software (whenever possible), checking against exact formulae and also by simulation.

Please cite this article in press as: T. Egeland, et al., A general approach to power calculation for relationship testing, Forensic Sci. Int. Genet. (2013), http://dx.doi.org/10.1016/j.fsigen.2013.05.001