Fuzzy P-values in Latent Variable Problems

Charles Geyer
University of Minnesota

joint work with

Elizabeth Thompson
University of Washington

`http://www.stat.umn.edu/geyer/fuzz`

Preprints

Thompson, E. A. and Geyer, C. J. (2005).
Fuzzy P-values in Latent Variable Problems
Technical Report No. 481.
Dept. of Statistics, Univ. of Washington.
Submitted to *Biometrika*.

Geyer, C. J. and Meeden, G. D. (2005).
Fuzzy Confidence Intervals and P-values.
To appear in *Statistical Science* (with discussion).

`http://www.stat.umn.edu/geyer/fuzz`

Genetic Linkage Analysis

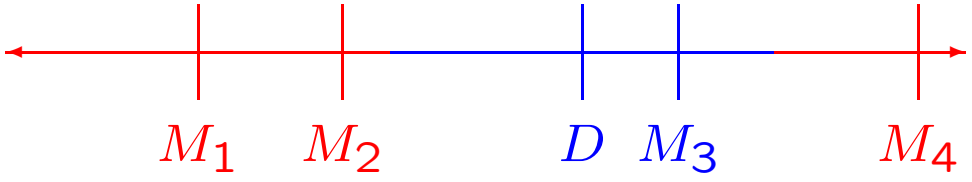
If a disease (or other trait) runs in families, then it may be partly genetic.

If a disease (or other trait) runs in families along with a marker trait associated with a known location in the genome, then some part of the trait may be associated with a nearby location in the genome (may be linked to the marker).

Genetic Linkage Analysis (Cont.)

Chromosomes occur in homologous pairs, one inherited from one parent. Each may be a combination of the homologous pair in the parent.

At each location the DNA may come from the grandfather (blue) or the grandmother (red). The points where the origin changes are called *crossovers*.



In the simplest model crossovers form a Poisson process and marginal segregation probabilities at each location are 50–50.

Completely specifies probability model for inheritance patterns.

Nonparametric Linkage Analysis

Given possibly incomplete data Y on marker status on individuals in a pedigree, we can simulate the inheritance pattern X at any genome location or any set of genome locations (`http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml`).

If willing to hypothesize a disease model, a probability model describing the trait given the underlying genetics, then we could calculate a likelihood (traditional lod score analysis).

Recent work (Whittemore and Halpern, 1994; Kruglyak, Daly, Reeve-Daly and Lander, 1996; Kong and Cox, 1997; McPeek, 1999; Nicolae and Kong, 2004; Thompson and Basu, 2003) avoids disease models.

Nonparametric Linkage Analysis (Cont.)

Let $t_\lambda(X)$ be a function of the inheritance pattern X at a genome location λ that should be larger when that location is associated with the disease than otherwise.

In our example $t_\lambda(X)$ is the size of the largest subset of affected individuals who carry DNA identical by descent at location λ in the realization X.

Problem: $t_\lambda(X)$ is not observable.

Simple Solution: use $w_\lambda(Y) = E\{t_\lambda(X) \mid Y\}$ as test statistic.

Criticism of Simple Approach

Test statistic $w_\lambda(Y) = E\{t_\lambda(X) \mid Y\}$ must be calculated by Monte Carlo (using simulation of X given Y) and is extremely computationally intensive.

Thompson and Basu (2003) point out that mere computation of $w_\lambda(Y)$ loses information in the distribution of $t_\lambda(X)$ given Y and confounds

- the evidence Y provides about X and
- the evidence X provides for linkage.

They proposed "pseudo-p-values" which were not true p-values (not Uniform$(0, 1)$ under the null hypothesis).

The Fuzzy Approach

X is a latent variable. $t_\lambda(X)$ is a latent test statistic.

$$s_\lambda(x) = \Pr\{t_\lambda(X) \geq t_\lambda(x)\}$$

is a latent p-value.

If we could observe $X = x$, then $s_\lambda(x)$ would be the p-value.

Thompson and Geyer (2005) call the random variable $s_\lambda(X) \mid Y$ the *fuzzy p-value* for the test of linkage in this situation.

The connection with Geyer and Meeden (2005) is they both have the same equation

$$E[\Pr\{s_\lambda(X) \leq \alpha | Y\}] = \alpha, \qquad \text{for all } \alpha$$

so the fuzzy p-value is a true p-value in the sense that (marginally, not conditionally on Y) it is Uniform$(0, 1)$.

Calculating Fuzzy P-Values

Need two sets of simulations

- $X_0^{(h)}$, $h = 1, \ldots, m$, from marginal of X under H_0
- $X^{(i)}$, $i = 1, \ldots, n$, from conditional of X given Y under H_0.

For each $X^{(i)}$ estimate $s_\lambda(X^{(i)})$ by

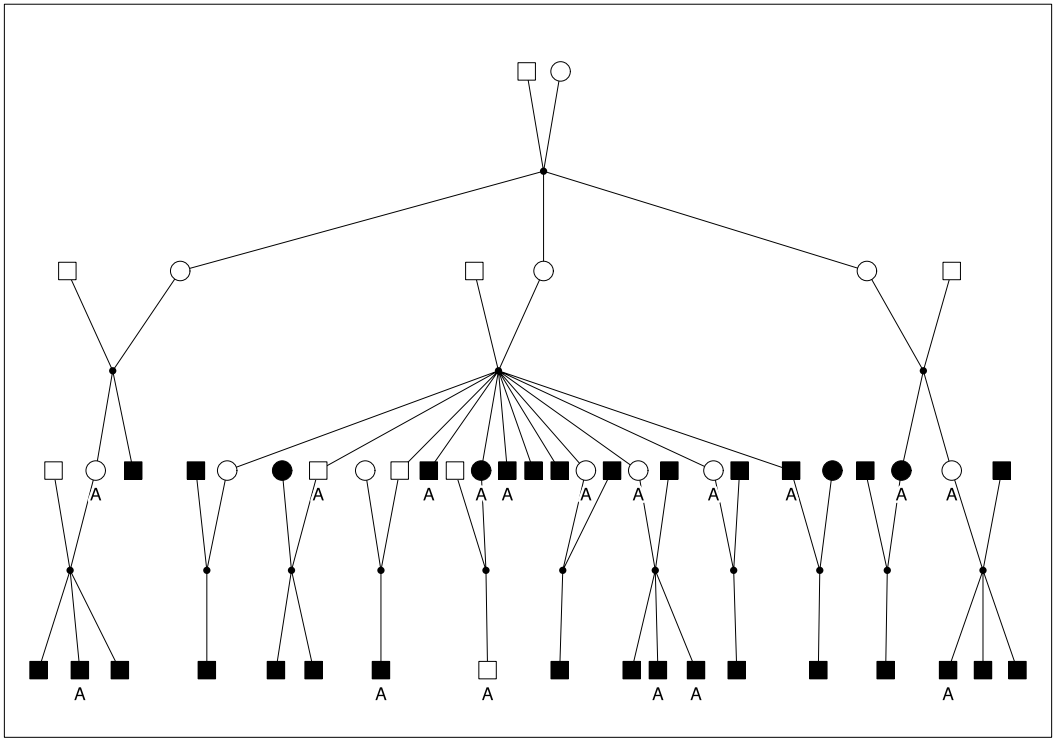
$$\widehat{s}_\lambda(X^{(i)}) = \frac{1}{m} \sum_{h=1}^{m} I\left\{t_\lambda(X_0^{(h)}) \geq t_\lambda(X^{(i)})\right\}$$

The distribution of the $\widehat{s}_\lambda(X^{(i)})$ as indicated by their histogram or empirical c. d. f. approximates the fuzzy p-value.

Virtues of Fuzzy P-Values

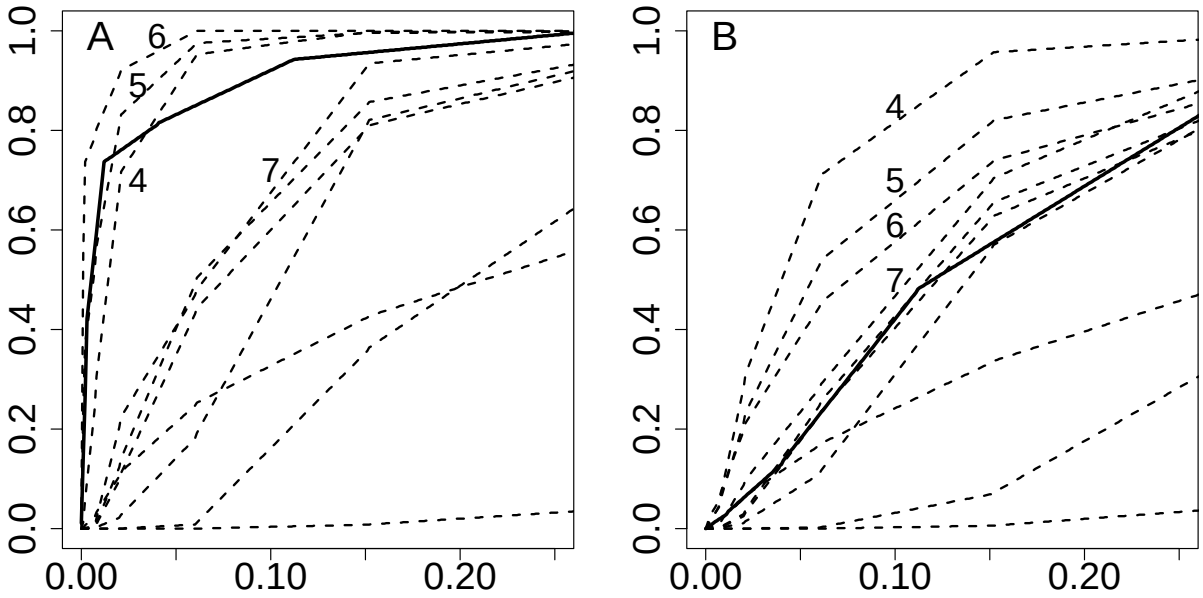
- Exact randomized tests. Simple interpretation.
- No two-stage Monte Carlo required.
- Too much fuzziness in fuzzy p-value indicates more markers needed.
- Only use conditional of Y given X under H_0. Not marginal of Y (as competing methods do).

Example Pedigree



'A' denotes affected. Dark shading denotes typed for at least 8 of the 10 DNA marker loci. No shading denotes no marker data except for two individuals typed at 2 marker loci.

Example Results



c. d. f. of fuzzy p-values. Dashed lines are for hypothesized disease locus at at one marker locus. Solid lines are for omnibus test (explanation follows) corrected for multiple testing. A: using all marker data. B: marker 6 data ignored.

Correction for Multiple Testing

Now consider multiple simultaneous tests of linkage multiple genome locations λ.

The right way to do multiple testing is to conceptually consider you are doing only one "omnibus" test. The procedure is constructed so the omnibus test rejects at level α with probability α so its p-value is Uniform$(0, 1)$.

The natural omnibus latent test statistic is

$$t_{\max}(X) = \max_{\lambda \in \Lambda} t_\lambda(X)$$

Since $t_{\max}(X)$ is just another latent test statistic, we already know what to do.