Fuzzy P-values in Latent Variable Problems

Charles Geyer
University of Minnesota

joint work with

Elizabeth Thompson
University of Washington

Glen Meeden
University of Minnesota

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

Preprints

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

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.

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

Ordinary Confidence Intervals

OK for continuous data, but a really bad idea for discrete data.

Why?

Coverage Probability

$$\begin{aligned} \gamma(\theta) &= \text{pr}_\theta\{l(X) < \theta < u(X)\} \\ &= \sum_{x \in S} I_{(l(x), u(x))}(\theta) \cdot f_\theta(x) \end{aligned}$$

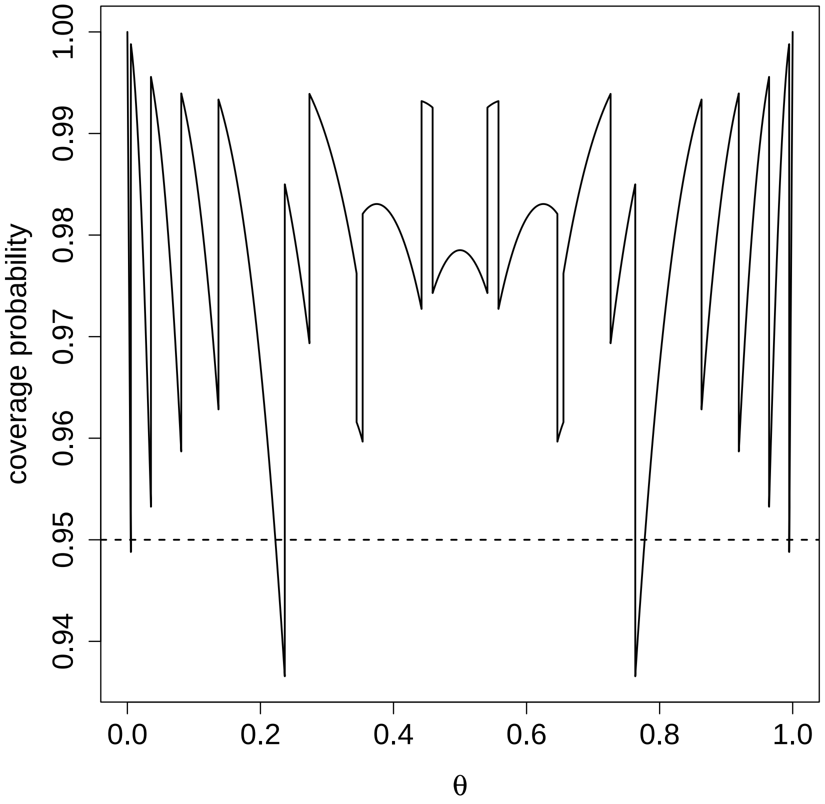
As θ moves across the boundary of a possible confidence interval $(l(x), u(x))$, the coverage probability jumps by $f_\theta(x)$.

Ideally, γ is a constant function equal to the nominal confidence coefficient.

But that's not possible.

Binomial Example

Binomial data, sample size $n = 10$, confidence interval calculated by R function `prop.test`



Fuzzy Tests and Confidence Intervals

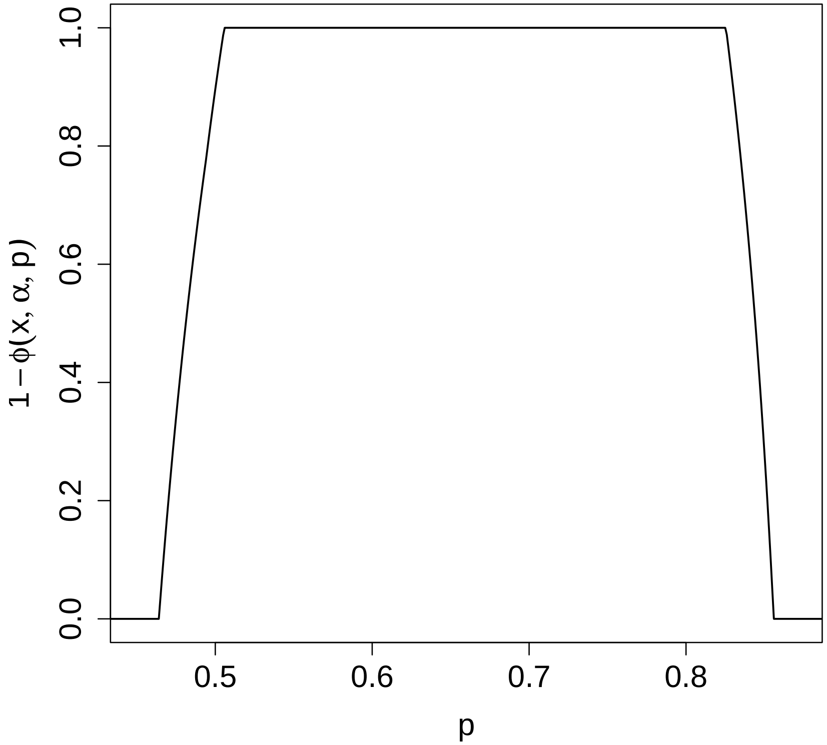
- For fixed α and θ_0,
  $$x \mapsto \phi(x, \alpha, \theta_0)$$
  is the *fuzzy decision* function for the size α test of $H_0 : \theta = \theta_0$.

- For fixed x and α,
  $$\theta \mapsto 1 - \phi(x, \alpha, \theta)$$
  is (the membership function of) the *fuzzy confidence interval* with coverage $1 - \alpha$.

- For fixed x and θ_0,
  $$\alpha \mapsto \phi(x, \alpha, \theta_0)$$
  is (the cumulative distribution function of) the *fuzzy P-value* for test of $H_0 : \theta = \theta_0$.

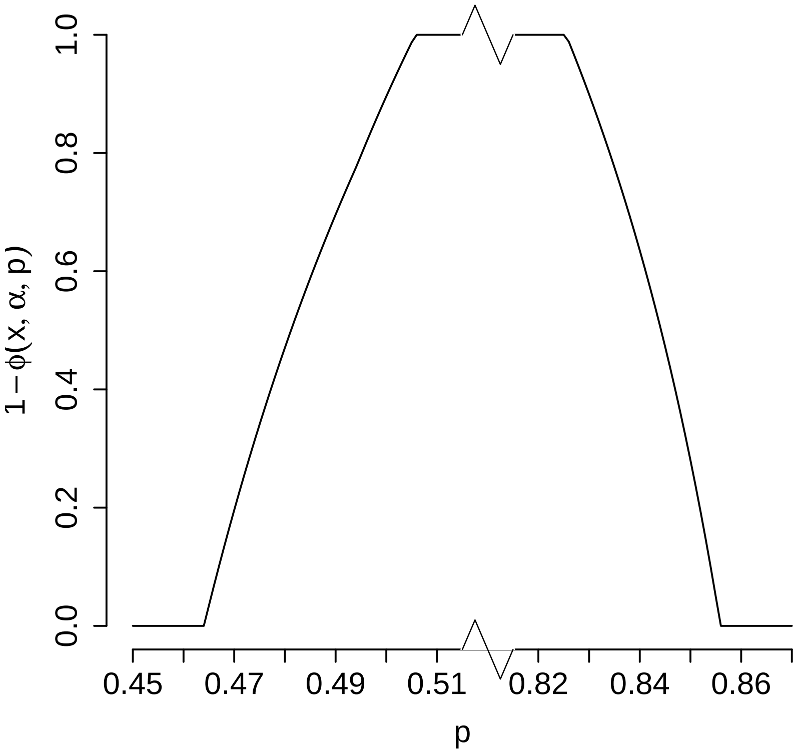
Fuzzy Confidence Interval
Binomial Example

Fuzzy confidence interval associated with the UMPU test, confidence level $1 - \alpha = 0.95$, sample size $n = 25$, data $x = 17$.



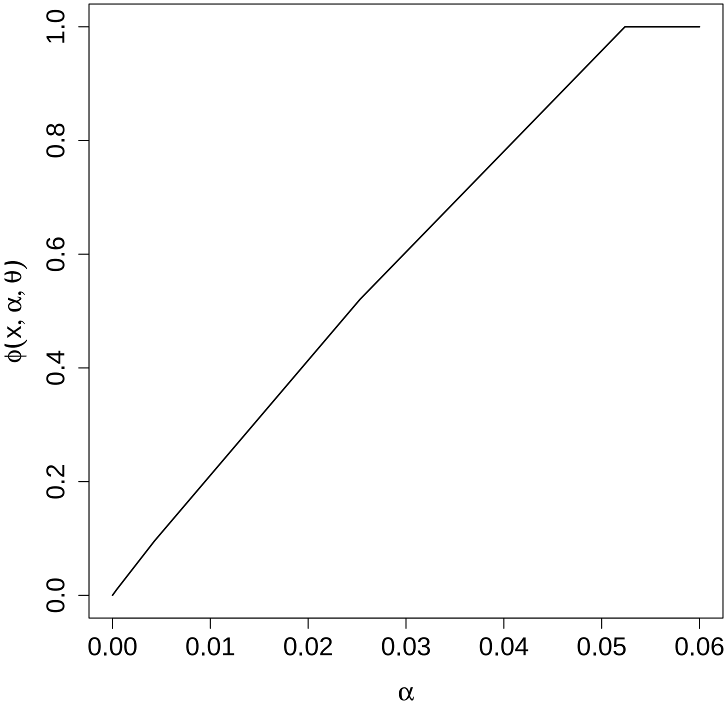
Fuzzy Confidence Interval
Binomial Example (Cont.)

Fuzzy confidence interval associated with the UMPU test, confidence level $1 - \alpha = 0.95$, sample size $n = 25$, data $x = 17$.



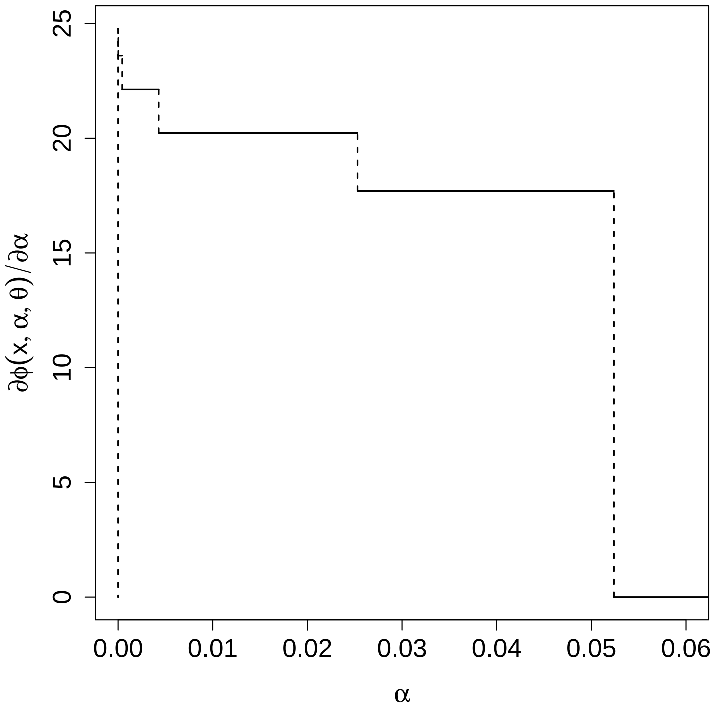
Fuzzy P-Value
Binomial Example

Cumulative distribution function of the fuzzy p-value associated with the UMPU test, sample size $n = 10$, data $x = 10$, null hypothesis $\theta_0 = 0.7$.



Fuzzy P-Value
Binomial Example (Cont.)

Probability density function of the fuzzy p-value associated with the UMPU test, sample size $n = 10$, data $x = 10$, null hypothesis $\theta_0 = 0.7$.



Genetic Linkage Analysis

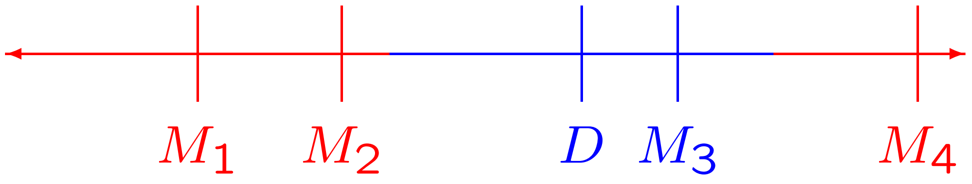
If a disease (or other trait) runs in families, then it may be partly genetic.

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*.



A *recombination* occurs between two locations if an odd number of crossovers occurs between them.

Genetic Linkage Analysis (Cont.)

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.

Only parameters to specify are Poisson intensity (genetic map, recombination probabilities) and population allele frequencies for markers.

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 (`http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml`) or joint inheritance pattern $\mathbf{X}$ at many locations.

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

In our example $t(X)$ is the size of the largest subset of affected individuals who are identical by descent in the realization X.

Problem: $t(X)$ is not observable.

Simple Solution: use $w(Y) = E\{t(X) \mid Y\}$ as test statistic.

Criticism of Simple Approach

Test statistic $w(Y) = E\{t(X) \mid Y\}$ must be calculated by Monte Carlo (using simulation of X given Y).

Null distribution of $w(Y)$ must be calculated by two-stage Monte Carlo

1. Simulate Y under null hypothesis.

2. For each simulated Y, simulate many X given Y and average $t(X)$ to calculate $w(Y)$.

Criticism of Simple Approach (Cont.)

Null distribution of $w(Y)$ is extremely computationally intensive.

Thompson and Basu (2003) point out that mere computation of $w(Y)$ loses information the distribution of $t(X)$ given Y provides about

- evidence Y provides about X
- evidence 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. $t(X)$ is a latent test statistic.

$$s(x) = \Pr\{t(X) \geq t(x)\}$$

is a latent p-value.

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

Thompson and Geyer (2005) call the random variable $s(X) \mid Y$ 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

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

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

For each $X^{(i)}$ estimate $s(X^{(i)})$ by

$$\widehat{s}(X^{(i)}) = \frac{1}{m} \sum_{h=1}^{m} I\Big\{t(X_0^{(h)}) \geq t(X^{(i)})\Big\}$$

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

Virtues of Fuzzy P-Values

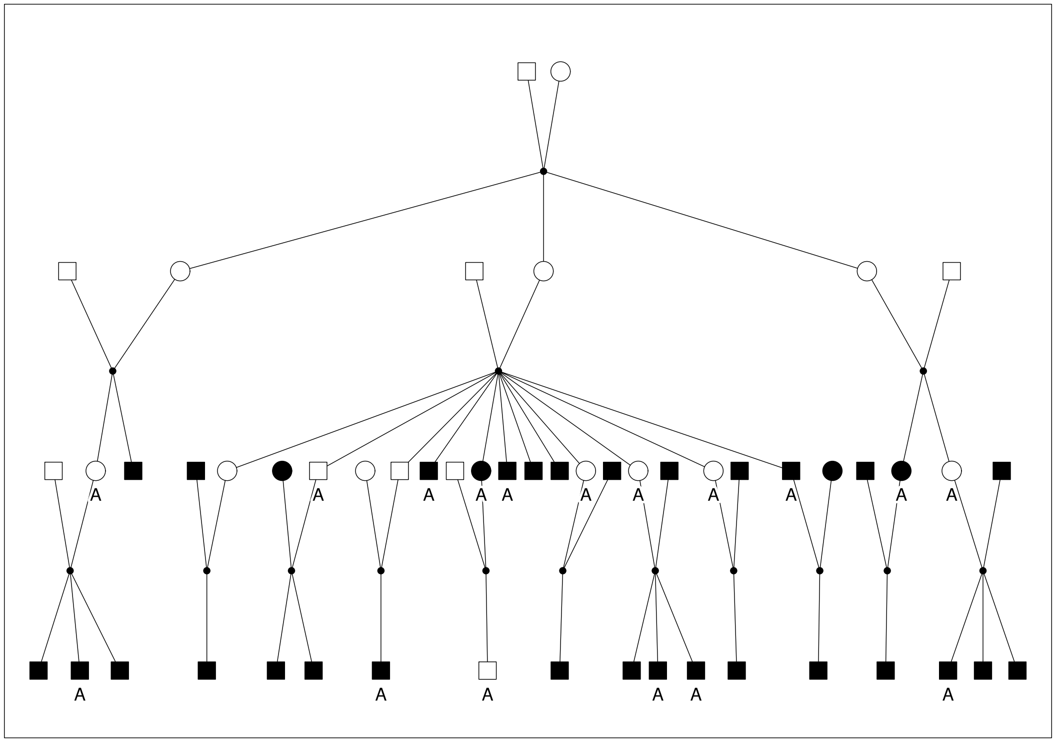
- Exact randomized tests. Simple interpretation.

- No two-stage Monte Carlo required.

- Too much fuzziness in fuzzy p-value indicates more markers needed.

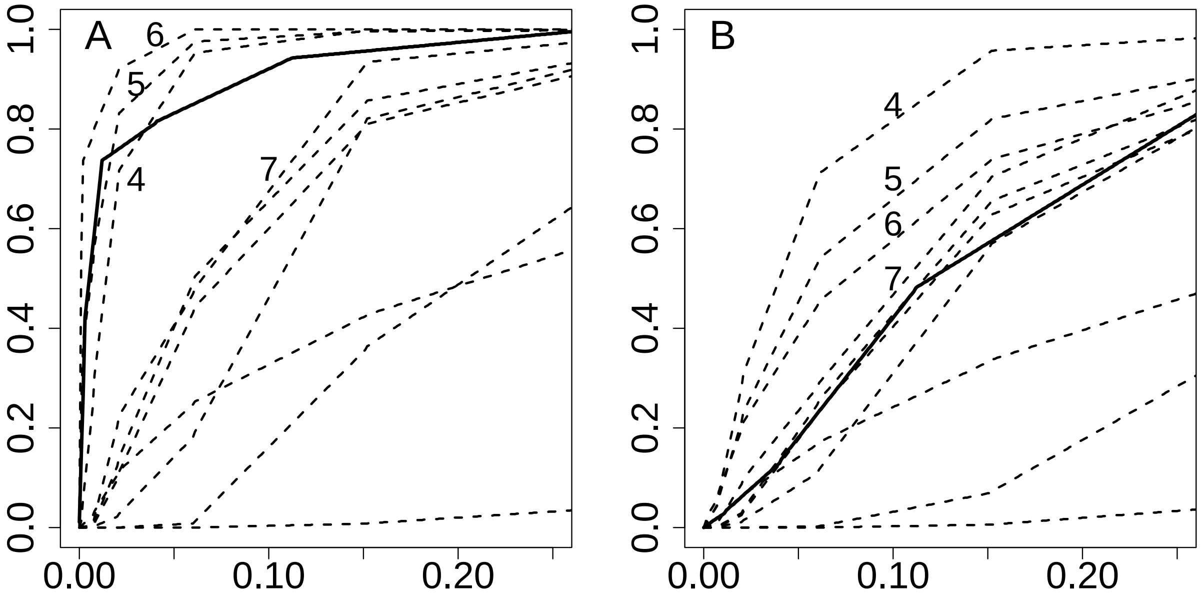
- Need only model conditional distribution of Y given X under H_0. Do not use marginal distribution of Y (competing methods do).

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 based on latent data (inheritance patterns) X_λ at multiple locations λ which we collect as the vector latent variable $\mathbf{X}$.

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

$$t_{\text{max}}(\mathbf{X}) = \max_{\lambda \in \Lambda} t(X_\lambda)$$

Except for notation $t_{\text{max}}(\mathbf{X})$ is just like $t(X)$ so we already know what to do.