Latent variables, Uncertainty and Evidence

Elizabeth Thompson
University of Washington

For Fields Institute, Toronto
Lecture 1, April 3, 2006

The general ideas: work with Charles Geyer, U.Mn.
Examples of this talk: work with student Yanming Di.

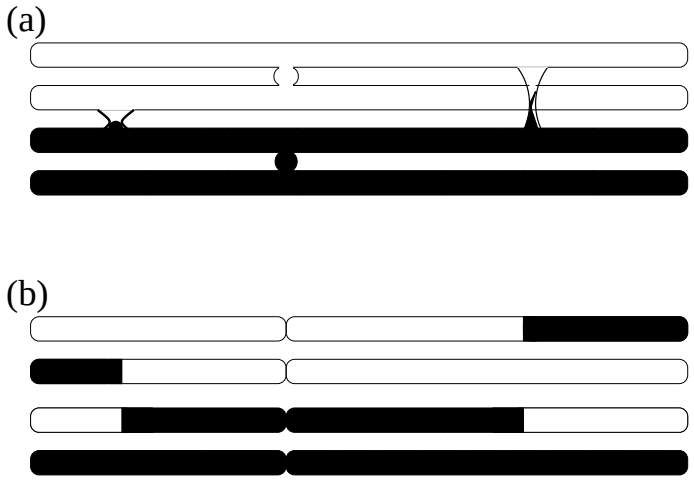
Latent variables

In many areas of genetics/genomics (and other sciences), we do not observe the variables that would make it easy to test hypotheses of interest:

Genetic Data observation	Model or Hypothesis	Latent variables
Offspring gametes	genetic interference	4 meiotic products
Data on pedigrees	genetic linkage	recombinant/non-rec
Variation in popns	coancestry/structure	gene inheritance
Variation btw species	mutation/selection	phylogeny

• How should our uncertainty about latent variables $\mathbf{X}$ be expressed in our inference?

Chromosomes and meiosis



Chromosomes duplicate align and exchange material.

Offspring chromosome consists of segments of two parental chromosomes (length $\approx 10^8 bp$).

There is dependence in DNA inherited at nearby locations: dependence is stronger for closer locations.

Difficulties: statistical and computational

• Often these models have complicated patterns of dependence among observed data components $\mathbf{V}$, resulting from the latent structure $\mathbf{X}$.

• Even computing a likelihood $\mathrm{P}(\mathbf{V})$ or relevant test statistics can be hard, requiring summation over the hidden variables:

$\mathrm{P}(\mathbf{V}) = \sum_{\mathbf{X}} \mathrm{P}(\mathbf{V} \mid \mathbf{X})\mathrm{P}(\mathbf{X})$.

Often, Monte Carlo is needed to compute the statistic or likelihood.

• Assessing significance can therefore be even harder. Even assuming can simulate $\mathbf{V}$ under a model (or even under the null hypothesis), analysis of each resimulated data set is needed: computationally very intensive. Monte Carlo within Monte Carlo.

An approach to solving these difficulties

- Objective:

(1) A new way to assess significance in such problems

(2) A way to express the uncertainty about this significance

where uncertainty derived from uncertainty about latent variables (not model mis-specification, etc.)

- First we introduce a simple example, without latent variables.

Testing association in a 2×2 table

- Suppose we have pairs of n binary (0/1) independent identically distributed observations (X_i, Y_i).

(For now, X_i is not latent: later it will be.)

Testing association in a 2×2 table

- The model. Under H_0: $\mathrm{P}(X = Y = 1) = \mathrm{P}(X = 1)\mathrm{P}(Y = 1)$

	Y=1	Y=0	
X = 1	$\mathrm{P}(X = Y = 1)$		$\mathrm{P}(X = 1)$
X = 0			$\mathrm{P}(X = 0)$
	$\mathrm{P}(Y = 1)$	$\mathrm{P}(Y = 0)$	1

- The data: independent pairs (X_i, Y_i), $i = 1, ..., n$.

	Y=1	Y=0	
X=1	$\sum_i X_i Y_i$	$\sum_i X_i(1 - Y_i)$	$\sum_i X_i$
X=0	$\sum_i (1 - X_i) Y_i$	$\sum_i (1 - X_i)(1 - Y_i)$	$\sum_i (1 - X_i)$
	$\sum_i Y_i$	$\sum_i (1 - Y_i)$	

Four possible ways to test

(1) Condition on $\sum X_i$, and $\sum Y_i$;
hypergeometric: one-fish, two-fish; tag-fish, new-fish.
Robust to the marginal models for $\mathsf{P}(X = 1)$ and $\mathsf{P}(Y = 1)$.

(2) If we have a model for Y_i; condition on $\mathbf{X} = (X_i)$.
Robust to marginal model for $\mathbf{X}$.

(3) If we have a model for X_i; condition on $\mathbf{Y} = (Y_i)$.
Robust to marginal model for $\mathbf{Y}$.

(4) Full model: unconditional: uses model for both $\mathbf{X}$ and $\mathbf{Y}$
Least robust, but most powerful.

A particular case of interest

- Suppose we know $\mathsf{P}(X_i = 0) = \mathsf{P}(X_i = 1) = 1/2$.

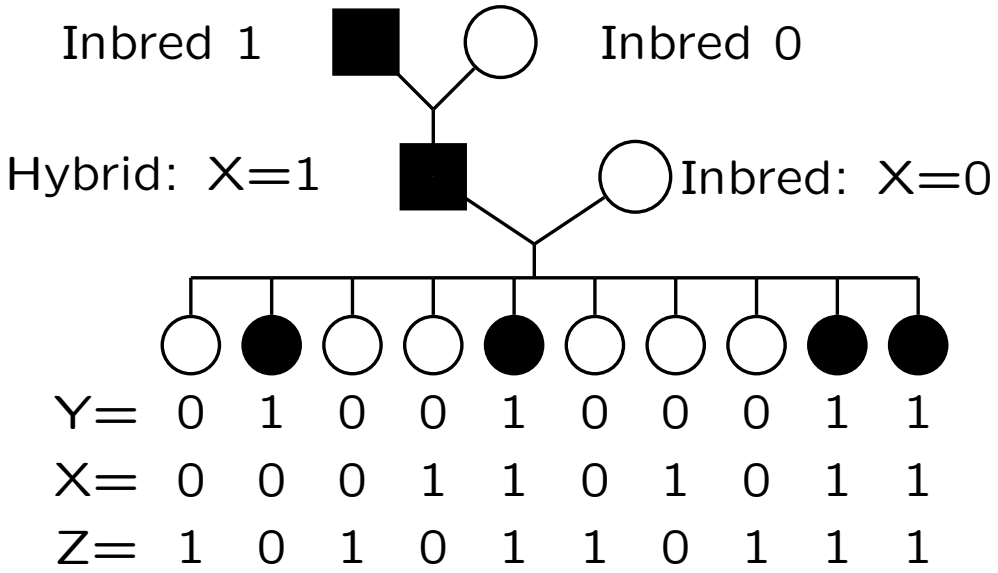
- Let $Z_i = 1$ if $X_i = Y_i$. (Agreement of X_i and Y_i.)

	Y=1	Y=0	
X = 1	Z=1	Z=0	1/2
X = 0	Z=0	Z=1	1/2
	p_1	p_0	1

Under the null hypothesis, $\mathsf{P}(Z_i = 1) = 1/2$, regardless of the distribution of Y_i (regardless of p_1 and p_0).

- Thus in this case, a good test statistic is
$T = \sum_i Z_i$ where $Z_i = (X_i Y_i + (1 - X_i)(1 - Y_i))$.

The backcross linkage design



Y_i denotes some trait value. X_i denotes DNA marker type.

- Mendelian genetics says $P(X_i = 1) = 1/2$.
- $Z_i = (X_i Y_i + (1 - X_i)(1 - Y_i))$. $Z_i = 1$ is $X_i = Y_i$.

Example data

In our example data

$T = \sum_i Z_i = 7$

	Y=1	Y=0	
X = 1	3	2	5
X = 0	1	4	5
	4	6	10

The traditional binomial test

- $H_0 : P(Z_i = 1) = 1/2$ vs. $P(Z_i = 1) = \theta > 1/2$
- Observe n outcomes: $T = \sum_{i=1}^{n} Z_i$.
- P-value: $P = \mathrm{P}_0(T \geq t_{\mathrm{obs}})$
- In our example, $n = 10$, and $T = 7$. P $= 0.172$
- Another example: $n = 30$, $\mathrm{P}_0(T \geq 20) = 0.049 \approx 0.05$
- For a test size 0.05 (Type 1 error): reject H_0 if $T > 19$.
- Due to discreteness of binomial, usually need a randomized test, (and our examples do), but this is not point of talk.

Four ways to condition

- 1. Condition on $\sum_i X_i = 5$ and $\sum_i Y_i = 4$.

In binary case, this is just hypergeometric distribution.
In general, it is a permutation test: permute $\mathbf{X}$ against $\mathbf{Y}$.
It is robust to the marginal distributions of X_i and Y_i.

- 2. Condition on $\mathbf{X}$, resimulate $\mathbf{Y}$ under H_0.

Requires knowledge of the marginal distribution of $\mathbf{Y}$.

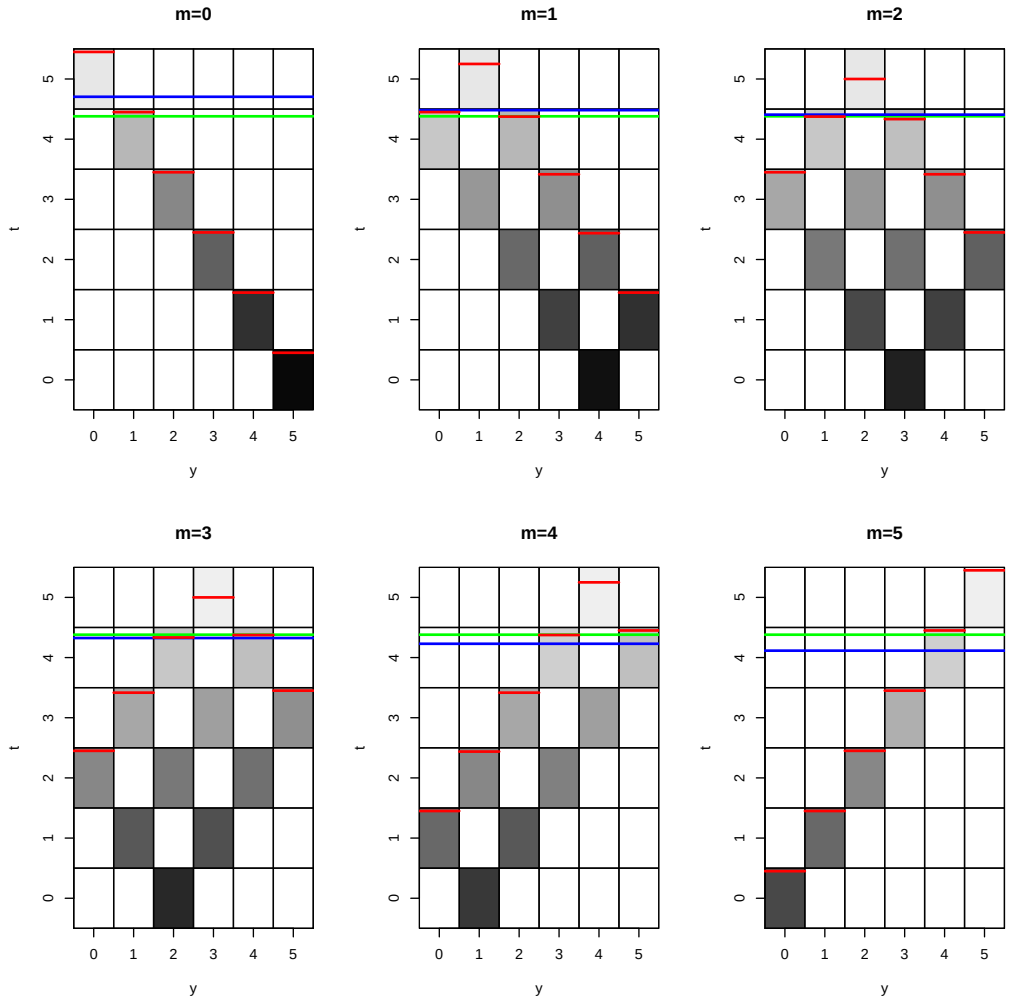
- 3. Condition on $\mathbf{Y}$, resimulate $\mathbf{X}$ under H_0.

Under H_0: $P(Z_i = 1 \mid Y_i = 1) = P(X_i = 1) = 1/2$
$P(Z_i = 1 \mid Y_i = 0) = P(X_i = 0) = 1/2$.

- 4. Resimulate both Y_i and X_i: the traditional binomial test.

In this case, the test is equivalent to (3), but note in general it is not robust to the marginal distribution of Y_i.

Comparison of rejection regions: case n=5



Explanation of figure of previous page

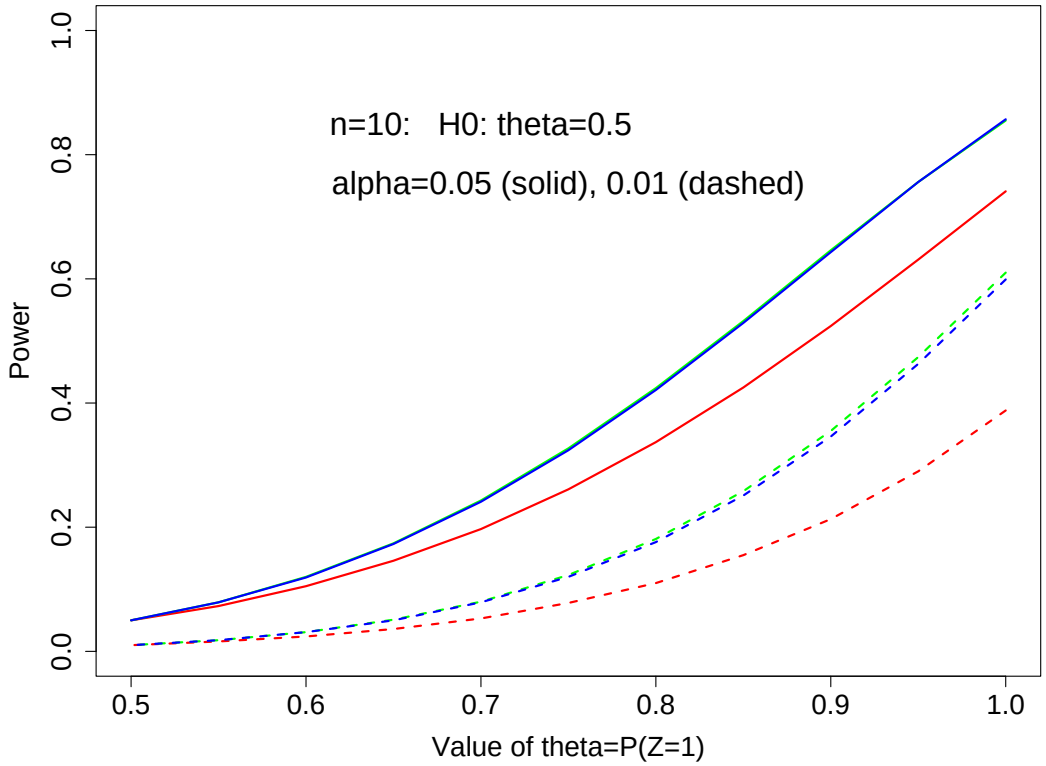
- Six grids: $m = \sum_i X_i$ = 0,1,2, 3,4,5
  Each grid, $y = \sum_i Y_i$ = 0,1,2...,5, on horizontal axis
  Each grid, t (agreements of X_i and Y_i) on vertical axis.

- White squares are impossible
  Shading denotes the LR − reject light squares first
  Note each grid square also has a probability (not shown)

- Green: unconditional, depends only on t
  Blue: conditional on m − reject the right portion of each grid
  Red: conditional on m and y: permutation test.
  − reject the right portion of every column.

Power of the tests

- The more we condition, the more robust the test is.
  The more we condition, the less powerful the test is.



Latent variables: Uncertainty in X_i or Z_i

- In the binary 2×2 table case,
  conditional on Y_i or unconditionally:
  $H_0 : P(Z_i = 1) \ = \ 1/2$ vs. $P(Z_i = 1) \ > \ 1/2$

- Suppose we do not observe Z_i but only V_i where
  $P(V_i = 0|Z_i = 1) = q_1$, $P(V_i = 1|Z_i = 0) = q_0$,
  where q_0 and q_1 are known.

- Under H_0: $P(V_i = 1) = (q_0 + (1 - q_1))/2 \ = \ q^*$
  $P(V_i = 0) = (q_1 + (1 - q_0))/2 \ = \ (1 - q^*)$.

- The Z_i (or X_i) are now latent variables.

One standard approach

- Standard approach: compute a statistic
  $W(\mathbf{V}, \mathbf{Y}) = \mathsf{E}_{\mathbf{0}}(\mathbf{T}(\mathbf{X}, \mathbf{Y}) \mid \mathbf{V}, \mathbf{Y}) = \mathsf{E}_{\mathbf{0}}(\mathbf{T}(\mathbf{Z}) \mid \mathbf{V}, \mathbf{Y})$,
in complex cases, by simulating $\mathbf{Z}$ given $\mathbf{V}$ and $\mathbf{Y}$.

- That is, if $T = \sum_i Z_i$, $W = \sum_i W_i$ where

$$W_i = \mathsf{P}(Z_i = 1|V_i) = \frac{V_i(1-q_1)}{2q^*} + \frac{(1-V_i)q_1}{2(1-q^*)}$$

- Now for a P-value, we need a distribution for $W(\mathbf{Y})$??
On complex data structures, the permutation test is often not an option. Also, will lose power, relative to using model for $\mathbf{X}$.

- Note that, under H_0,
  $\mathsf{E}(W_i) = \mathsf{E}(\mathsf{E}(Z_i|V_i)) = \mathsf{E}(Z_i) = 1/2$,
  However, $\mathsf{var}(W_i) = \mathsf{var}(\mathsf{E}(Z_i|V_i)) < \mathsf{var}(Z_i)$.

Example: Z observed with error/uncertainty

$P(V = 0 \mid Z = 1) = q_1 = 0.3$. $P(V = 1 \mid Z = 0) = q_0 = 0.2$; assume we know q_1 and q_0.

	$Z = 0$	$Z = 1$	$P(V)$	Under $\theta = 1/2$
$V = 0$	$0.8(1-\theta)$	0.3θ	$0.8(1-\theta) + 0.3\theta$	$1.1/2$
$V = 1$	$0.2(1-\theta)$	0.7θ	$0.2(1-\theta) + 0.7\theta$	$0.9/2$
	$1-\theta$	θ	1	1

Under H_0: $P(Z = 1|V = 1) = .7/.9$, $P(Z = 1|V = 0) = 0.3/1.1$

$\mathrm{E}(Z \mid V) = (7/9)V + (3/11)(1-V) = (3/11) + (50/99)V$.

Standard test is based on
$W = \mathrm{E}(T|V_1, \ldots, V_n) = (3n/11) + (50/99)V$ where $V = \sum_i V_i$ and $T = \sum_i Z_i$.

Testing based on $W = \mathrm{E}(T|\{V_i\})$

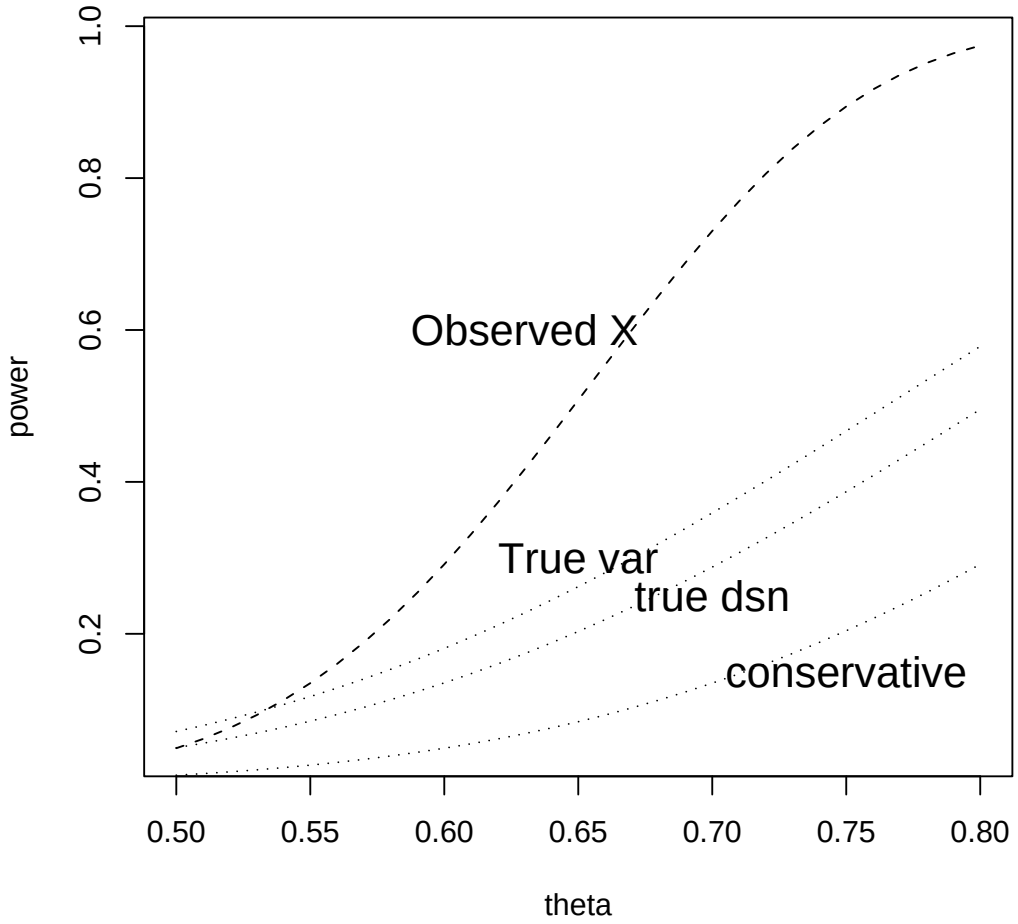
Under H_0: $\mathrm{E}(W) = n*((3/11)+(50/99)*(0.45)) = n/2 = \mathrm{E}(T)$
$\mathrm{var}(W) = n*(50/99)^2*0.45*0.55 \approx \mathrm{var}(T)/4$

Three possible tests (example is $n = 30$):
(1) If we can compute it use the correct distribution of W;
Critical value is $V = 18$: corresponds to $W = \mathrm{E}(T|V) = 17.27$
Reject H_0 with prob 0.43 if W=17.27 ($\sum_i V_i = 18$),
and if $W \geq 17.78$ ($\sum_i V_i \geq 19$)

(2) Or use a normal approximation with the correct H_0 variance (if we can compute it). Anti-conservative.

(3) Or we can use a normal approximation with the larger variance, which under H_0we do know; $\mathrm{var}(T)$. Conservative .

The powers of tests based on W and T: ($n = 30$)



Problems with this standard approach

- The distribution of $W(\mathbf{V}, \mathbf{Y})$ may be unknown (not here).

- Can always, in principle, simulate under H_0 to obtain an empirical p-value: computationally intensive.

- May need to simulate $\mathbf{Y}$ and/or $\mathbf{V}$ under H_0: want a test robust to marginal models of $\mathbf{Y}$ and $\mathbf{V}$.

- If we can simulate $T = \sum_i Z_i$ given $(\mathbf{V}, \mathbf{Y})$, we have an empirical distribution of T given $(\mathbf{V}, \mathbf{Y})$, not just $W = \mathsf{E}(T|\mathbf{V}, \mathbf{Y})$.

- Uncertainty in what V_i says about Z_i is confounded with evidence Z_i provide about H_0.

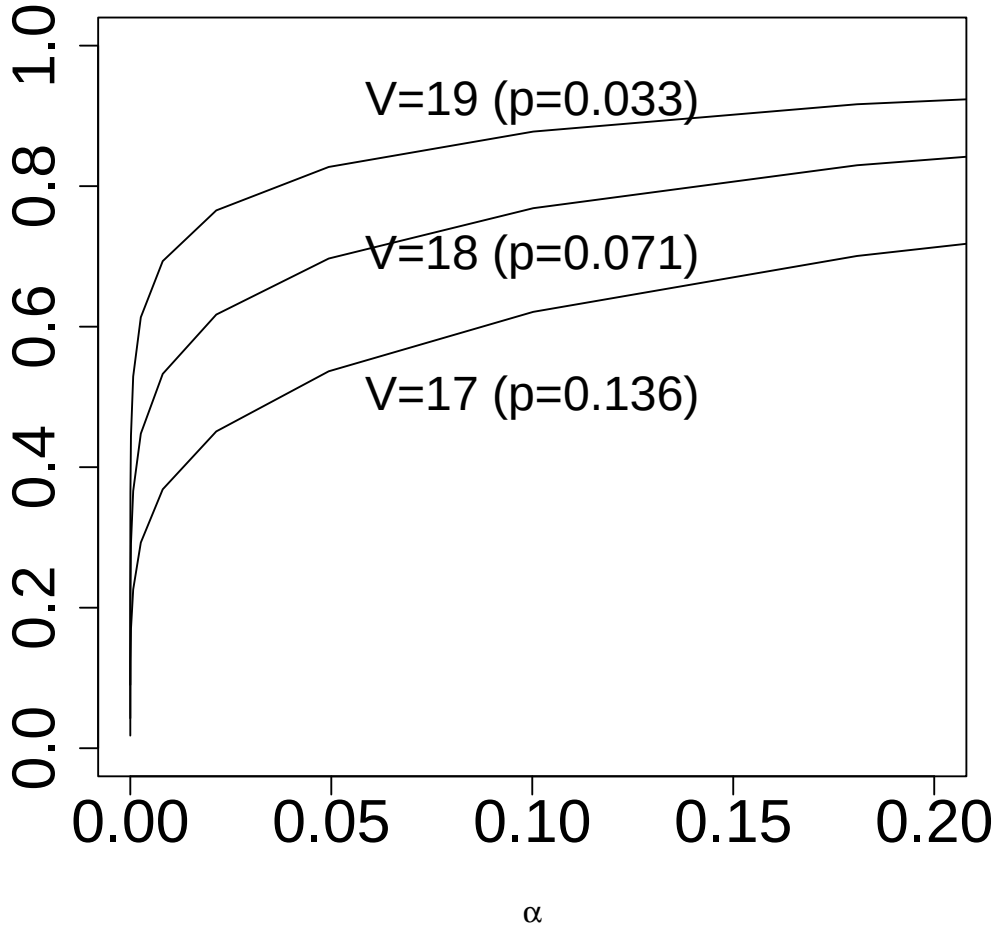
A fuzzy p-value; definition and computation.

- Let $\pi(t)$ be the p-value if we observe $T = t$.

- The fuzzy p-value is a RANDOM VARIABLE which has the probability distribution of $\pi(T)$ where T has the prob dsn of T given we observe $V = v$.

- Requires only simulation of T under H_0, to get $\pi(T)$,
and simulation of T given V, to get required fuzzy-p distribution.
No simulation of data variables V or Y is required.
Everything is conditional on the observed values of (V, Y).

Fuzzy-p dsns for observed V values; V=17, 18, 19



A fuzzy p-value; interpretation

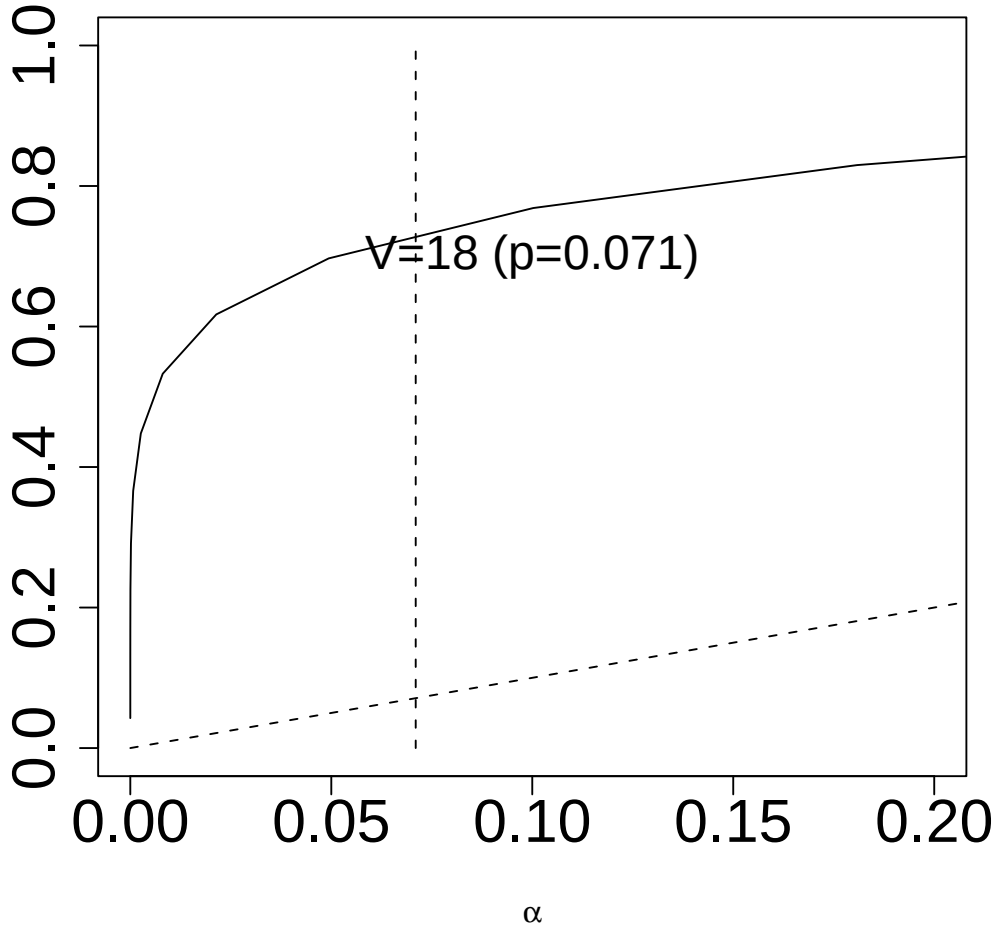
• If V specified T exactly, it would be concentrated at a single value (like a "regular" p-value).

• If V says nothing about T it is spread uniformly on $(0, 1)$, with cdf $F(q) = q$ – see graph.

• The fuzzy p-value expresses both the strength of evidence about H_0 and the uncertainty about the evidence (due to uncertainty about T).

• The uncertainty is put directly onto the p-value scale.

Interpretation of the fuzzy-p distribution



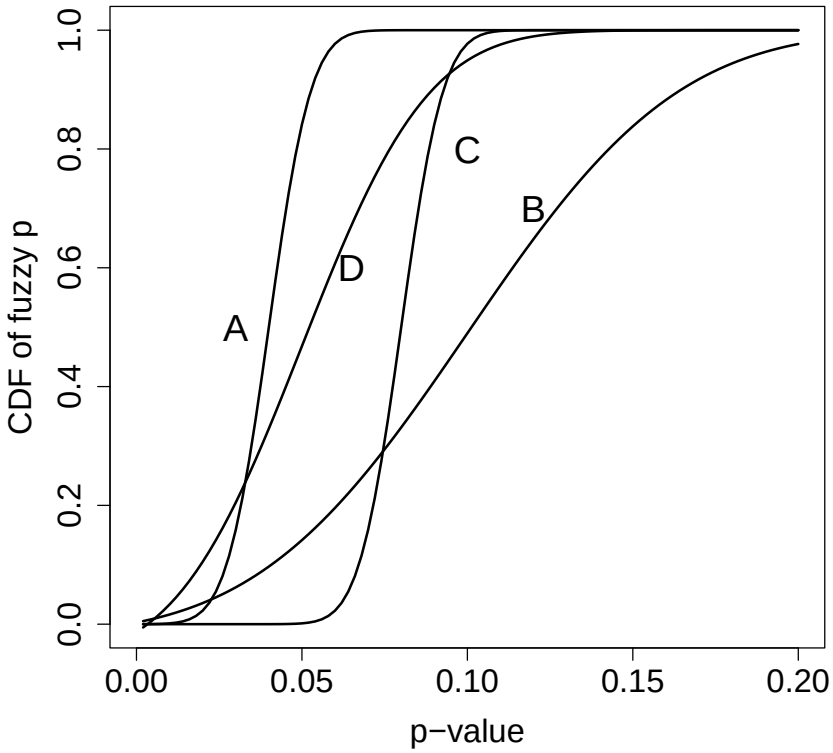
Test based on the fuzzy p-value

Given data $V = v$, and wanting a test of size 0.05 say, we reject H_0 with the probability that the fuzzy-p random variable is less than 0.05 (about 0.65 for $V = 18$ in our example).

This test is not very powerful – the power curve is not too different from the "conservative" test in our example. (But Type-I error is correct.)

This is because we have taken into account our uncertainty about T. It is misleading to have a powerful test, that does not reflect this uncertainty.

Reducing uncertainty, with additional data



- Suppose there is some way to reduce the uncertainty in Z_i or X_i, given V_i ($q_0 \approx q_1 \approx 0$).
- The fuzzy-p dsn can guide collection of such potential additional data.
- A: No need.
- B: No hope.
- C: Not on these structures.
- D: Yes, on these structures!!
- The fuzzy-p dsn puts current uncertainty directly onto p-value (evidence) scale.

CONCLUSION: Conditioning

• In genetic examples, we may have confidence in a model for DNA inheritance $\mathbf{X}$, but not for trait data $\mathbf{Y}$.

• We want tests that are robust to the model for $\mathbf{Y}$.

• In case of bivariate binary data (X_i, Y_i) a permutation test is robust to marginal distributions of both X_i and Y_i, but on more complex data structures, permutation is not an option. Also, permutation test loses information, unnecessarily.

• An alternative is to re-simulate simulate $\mathbf{X}$, under H_0, to obtain an empirical p-value for a test, conditional on $\mathbf{Y}$.

CONCLUSION: Expressing Uncertainty

• $\mathbf{X}$ may be latent: it is often the latent variables that would provide the evidence for scientific hypotheses.

• The data (V_i, Y_i) may be a very imperfect reflection of $Z_i(X_i, Y_i)$.

• Basing p-values on statistics constructed from data $(\mathbf{V}, \mathbf{Y})$ is very computationally intensive, and may not be robust.

• Evidence in $\{Z_i\}$ is confounded with uncertainty about $\{Z_i\}$.

• Fuzzy p-values address these issues, putting uncertainty in $\{Z_i\}$, (i.e. $\mathbf{X}$) directly on evidence scale, and can thus guide collection of additional data.