

Hypotheses testing as a  
fuzzy set estimation problem

Glen Meeden  
University of Minnesota

<http://www.stat.umn.edu/glen/talks>

Joint work with Siamak Noorbaloochi

## *p*-values

As introduced by Fisher a *p*-value measures the strength evidence in the data against the null hypothesis.

They are often easy to compute after the null hypothesis has been chosen.

They are widely used and perhaps widely misunderstood.

## Problems with $p$ -values

The usual formal justification comes from the Neyman-Pearson theory of hypotheses testing which assumes a sharp break between the null and alternative hypothesis.

When the true parameter point is close to the boundary and the sample size is large then there is high probability the outcome will be statistically significant although most observers would agree that the result is of no practical importance.

Over emphasis in the scientific community on only publishing results which have a  $p$ -value smaller than 0.05 or 0.01.

## Alternative interpretation of a $p$ -value

Depending on the situation the null hypothesis can either be the set of **good** or **bad** parameter values. Here we will assume they are the **good** values.

A  $p$ -value is an unbiased estimator of its expectation.

Its expectation can be interpreted as the fuzzy membership function of the parameter values which belong to the set of good or interesting values.

Statisticians do not worry enough about what function the  $p$ -value is estimating.

## Hypotheses testing as fuzzy set estimation

Alternatively one can define a fuzzy membership function on the parameter space which represents the degree of membership of each point in the set of **good** parameter values.

This membership function should be chosen carefully to reflect the realities of the problem.

Then one must find a good estimator of the selected membership function. Standard theory can be helpful here.

Selecting the membership function can be less straightforward than selecting a null hypothesis but if done well the resulting analysis can be more useful.

## The botox example

In a clinical trial 22 patients with chronic refractory shoulder pain were injected with a mixture of Botox and lidocaine. After a month the patients were checked to see how many of them had experienced a meaningful reduction in their pain and 10 of the 22 responded that it did. Do these data support the conclusion that Botox could be useful in such situations?

Let  $\theta$  denote the probability that a patient responds to the Botox treatment. A classical analysis might select a  $\theta_0$  and compute a  $p$ -value for testing

$H : \theta \geq \theta_0$  against  $K : \theta < \theta_0$

How to choose  $\theta_0$ ?

## Choosing $\theta_0$

There is not a good treatment for this condition.

In such clinical trials it is known that as many as 25% of the patients can experience a placebo effect.

Little is known about the efficacy of Botox as a pain reliever and its possible side effects.

Our choice is  $\theta_0 = 0.35$ .

## Choosing a fuzzy membership function

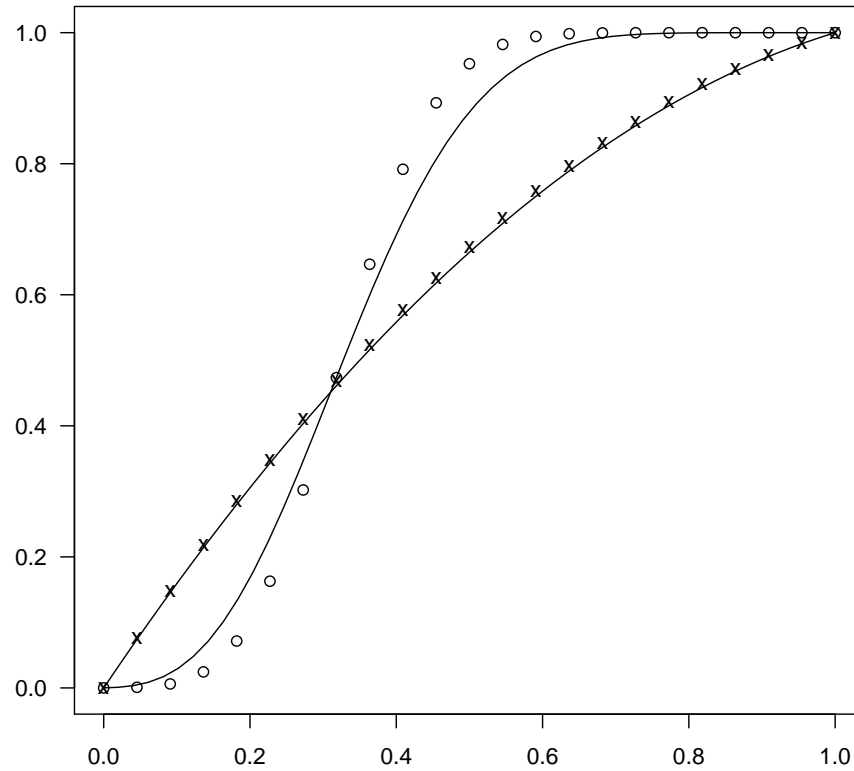
We want our fuzzy membership function to take on the value 0.5 at  $\theta_0 = 0.35$ .

For a positive integer  $m < n$  let  $\phi_m$  denote the UMP level 0.5 test of  $H$  against  $K$  based on  $Y_m$  a binomial( $m, \theta$ ) random variable.

Let  $\lambda_m(\theta) = 1 - E_\theta \phi_m(Y_m)$ . Then  $\lambda$  is a strictly increasing function on the unit interval whose range is also the unit interval and it takes on the value  $1/2$  at  $\theta_0$ .

So any convex combination of such functions is a possible fuzzy membership function and each has an unbiased estimator.

We selected  $\lambda_2(\theta)$  as our fuzzy membership function (fmf) since we judged that small differences in the neighborhood of  $\theta_0 = 0.35$  are not so important.



Plots of the expected value of the  $p$ -value and our fmf. The  $o$ 's and  $x$ 's are the  $p$ -values and our estimates when  $n = 22$ .

## Some remarks

For the data the  $p$ -value is more optimistic about the usefulness of Botox than our estimate.

If the sample size were increased the curve of the expected value of the  $p$ -value would change, getting steeper and steeper in the neighborhood of  $\theta_0 = 0.35$ .

Our fuzzy membership function was selected to reflect the realities of the problem and does not depend on the sample size.

The dependence of the  $p$ -value on the sample size seems to us to be a flaw.

## Reproductive success of Chimpanzees

Chimpanzees mate promiscuously and a female may (and typically does) mate with many males within their group.

In a group more than one female can become fertile at the same time.

Higher ranking males tend to have more access to fertile females than lower ranking males.

Data were collected in Gombe National Park in Tanzania by Emily Wroblewski and Anne Pusey of the U of Minnesota during one study period.

They have a theory that relates rank of a male to number of offspring he sires.

rank	expected	observed
1	0.3738	0.3030
2	0.2829	0.1515
3	0.1465	0.0606
4	0.0758	0.0000
5	0.0607	0.1515
6	0.0304	0.0606
7	0.0152	0.0303
8	0.0065	0.0606
9	0.0028	0.0909
10	0.0028	0.0606
11	0.0028	0.0000
12	0.0000	0.0303

The expected and observed frequencies for the 12 males along with their ranks. The data is based on 33 births.

## A standard analysis

One possible statistical analysis would be to use a chi-squared test to test the null hypothesis using their expected frequencies as the “truth” .

This test rejects the null hypothesis because a cell with zero probability under the null contains an observation.

This seems unfair to the theory since any theory is only an approximation to reality.

We now consider how fuzzy membership functions can be used to evaluate the reasonableness of the theory in light of the data.

## Choosing a parameter space

Let  $p = (p_1, \dots, p_{12})$  represent the true, average proportion of offspring that would be sired by the 12 males in this group.

We assume their rank decreases as their labels increase.

Let  $\mathcal{P}$ , a subset of the simplex, contain the  $p$  where higher ranked males tend to have more access to fertile females than lower ranked males.

The  $p$  with  $p_1 = 1$  could belong to this set but is not a realistic possibility for our problem.

We want  $\mathcal{P}$  to exclude such silly vectors but without assuming too much prior information.

## Definition of $\mathcal{P}$

We take  $\mathcal{P}$  to be the subset of the simplex which satisfies the following 3 constraints.

- $\max_{\{i=1,\dots,6\}} p_i \leq 0.5$
- $\max_{\{i=7,\dots,12\}} p_i \leq 0.25$
- $\sum_{i=1}^6 p_i \geq 0.5$

## A family of fuzzy membership functions

$$f(p) = f(p_1, \dots, p_{12}) = \sum_{i=1}^{11} a_i(p_i - p_{i+1}) + c$$

where  $c$  and the  $a_i$ 's are constants which satisfy

- The  $a_i$ 's are decreasing in  $i$ .
- $f((0.5, 0.5, 0, \dots, 0)) = 1$ .
- The range of  $f$  on  $\mathcal{P}$  is  $[0, 1]$ .

## Our fmf

After some experimentation we settled on  $f^*$  where  $c = 0.20$  and  $a_1 = 1.75$ ,  $a_2 = 1.60$ ,  $a_3 = 1.35$ ,  $a_4 = 1.20$ ,  $a_5 = 1.05$ ,  $a_6 = 0.90$ ,  $a_7 = 0.75$ ,  $a_8 = 0.60$ ,  $a_9 = 0.45$ ,  $a_{10} = 0.30$  and  $a_{11} = 0.15$ .

The value of our fuzzy membership function  $f^*$  at the expected is 0.746 which is considerably higher than 0.620, its value at the observed.

This suggests that their model somewhat over states the importance of male rank in the mating process.

## Two possible objections

Choosing the fmf  $f^*$  was too hard and the final estimate is somewhat arbitrary.

We believe that there is some force to the first objection. Selecting a good fuzzy membership function for a problem can require some thought. But once one has been found it makes the interpretation of its estimate more meaningful. We think that this is a strength and not a weakness of this approach.

Some additional analysis shows that our answer for this problem is quite robust against a large class of choices for  $c$  and the  $a_i$ 's.

## Final remarks

When computing a  $p$ -value it is the **ordering** of the data values in strength of evidence against the null which is crucial.

In many problems, once the null hypothesis has been selected, there is only one sensible ordering of the data and hence there is only one level of significance for a given data point.

This suggests that the usual theory of  $p$ -values is too crude, since it does not allow for a more nuanced measure of the evidence which takes into account more of the realities of the problem.