Comparison of various methods for meta–analysis of p–values

Michael Dewey

May 13, 2016

1 Introduction

1.1 What is this document for?

This document compares some methods for the meta–analysis of p–values (significance values) and their implementation in the package `metap` which is available from CRAN here:

1.2 Notation

The k studies give rise to p–values, $p_i, i = 1, \ldots, k$. These are assumed to be independent. We shall also need weights $w_i, i = 1, \ldots, k$. Logarithms are natural. A function for combining p–values is denoted g.

The methods are referred to by the name of the function in `metap`. Table 1 shows other descriptions of each method.

2 Theoretical results

There have been various attempts to clarify the problem and to discuss optimality of the various methods. A detailed account was provided by Lipták

Function name	Description
`logitp`	
`meanp`	
`minimump`	
`sumlog`	Fisher’s method
`sump`	Edgington’s method
`sumz`	Stouffer’s method
`votep`	

Table 1: Methods considered

(1958) although the readers is cautioned that this requires a certain familiaruty with the methods of probability theory.

Birnbaum (1954) considered the property of admisibility. A method is admissible if when it rejects H_0 for a set of p_i it will also reject H_0 for P_i^* wherre $p_i^* \leq p_i$ for all i. He also points out two classes ofalternative hypothesis H_a. The first is that all p_i have the same (unknown) non–uniform, non–increasing density, the second is that at least one p_i has an (unknown) non–uniform, non–increasing density.

See also Owen (2009).

An annotated bibliography is provided by Cousins (2008)

3 Weighting

It is possible to weight the p–values. At the moment this is only provided in `sumz` as this is the only method for which a published example is accessible. According to the account by Zaykin (2011) following Lipták (1958) the preferred weights are the effect sizes but failing them the square root of the sample sizes may be used. In fact when effect sizes are available the usual meta–analytic arsenal is available and the methods in this package are not needed.

4 Comparison of methods

The problem which the methods are designed to solve is poorly specified. This may account for the number of methods available and their differing behaviour. In this section we present a number of rather extreme situations to study the behaviour of the different methods.

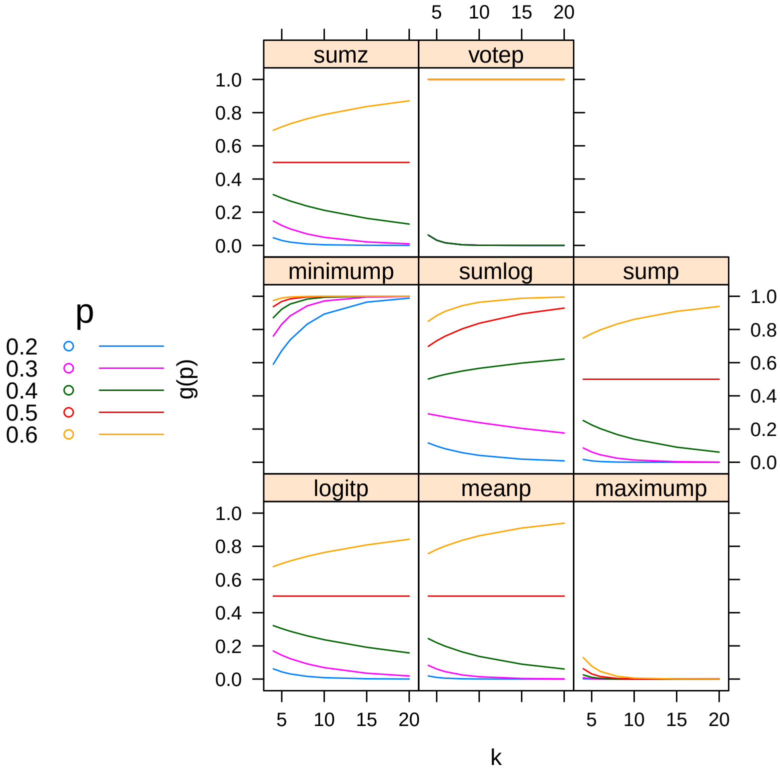

Figure 1: Behaviour of the methods for k values of $p = p_i$

4.1 What if all $p_i = p$?

The methods do not all perform in the same way here. We classify them by how $g(p_1, \ldots, p_i, \ldots, p_k)$ varies with p_i and k. Figure 1 shows the result for various values of p_i and k.

Decrease with k when p below 0.5, increase above In this category are `logitp`, `meanp`, `sump`, `sumz`

Decrease with k when p below x, increase above This is true for `sumlog` and the cut–off x varies with k, is somewhere between 0.3 and 0.5

Always increase with k This is true for `minimump`

Always decrease with k This is true for `maximump`

Invariant with p below 0.5 `votep` is 1 above 0.5 and otherwise invariant with p but decreases with k.

4.2 Exactly two values of p

All the methods have the property that $g(p_1, p_2) = g(p_2, p_1)$.

Figure 2 shows the behaviour of four of the methods. The x– and y– axes are the values of p_1 and p_2 and the z–axis is the values of $g(p_1, p_2)$. Figure 2 only shows `logitp` as `sump` and `sumz` have similar properties.

4.2.1 `logitp`, `sump` and `sumz`

These have similar properties. Note that $g(p, 1-p) = 0.5$ for all p so the line joining the point $(0, 1, 0.5)$ and $(1, 0, 0.5)$ is a straight line. Note how when one p is small the value of g still increase as the other one tends to 1.

4.2.2 `sumlog`

Dominated by smaller p_i. Note how the value of g remains small for small p even as the other one tends to 1.

4.2.3 `minimump` and `maximump`

Completely dominated by smaller (larger) p (obviously). $g(p_1, 1-p_1) \simeq 0.5$ for $p_1 \simeq 0.3$ (Work out exactly).

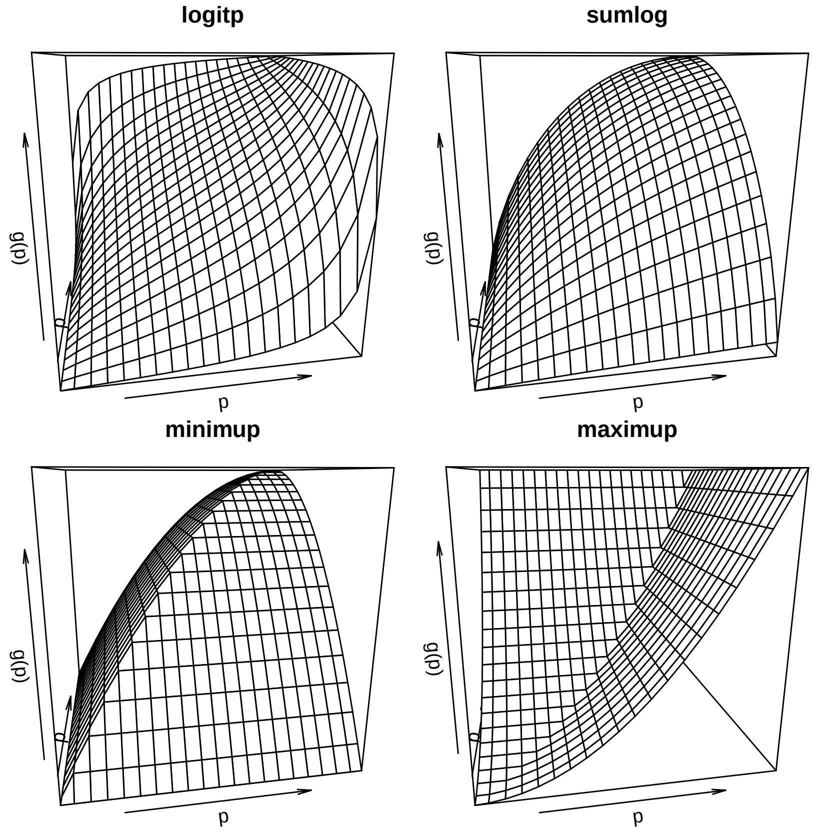

Figure 2: Two values of p_i

4.2.4 `votep`

0.25 when $p_1 < 0.5$ and $p_2 < 0.5$

0.5 when $p_1 = 0.5$ and $p_2 = 0.5$

1 when $p_1 > 0.5$ and $p_2 > 0.5$

0.75 when $p_1 < 0.5$ and $p_2 > 0.5$

4.3 Cancellation

What happens when the collection of primary studies contains a number of values significant in both directions? An example might be four studies having p–values 0.001, 0.001, 0.999, 0.999. If the intention of the synthesis is to examine a directional hypothesis one would want a method where these cancelled out. Note that of the methods considered here the method of the sum of logs and Wilkinson's method (and its special case minimum p) do not cancel out and report a significant result for this example. This is a consequence of the behaviour we have seen in a more limited setting in Section 4.1.

As an example we use `sumlog` and `sumz`.

```
> pvals <- c(0.001, 0.001, 0.999, 0.999)
> sumlog(pvals)
```

```
chisq =  27.63502  with df =  8  p =  0.0005488615
```

```
> sumz(pvals)
```

```
sumz =  0 p =  0.5
```

Clearly the choice should be made on scientific grounds not on the baiss of the outcome.

4.4 Effect of one extreme value

The alternative hypothesis is either that all the tests rejected the null, or that at least one of them did. We can investigate this by using the situation where all but one of the values are drawn under the null with just one extreme one.

In order to show the effect of one extreme value we resort to simulation. Each run generates nine uniform random numbers in $(0, 1)$. Each method is then applied to those numbers with the addition of a tenth number which is 0.001. The whole thing is repeated 10000 times and the result shown in the Figure 3.

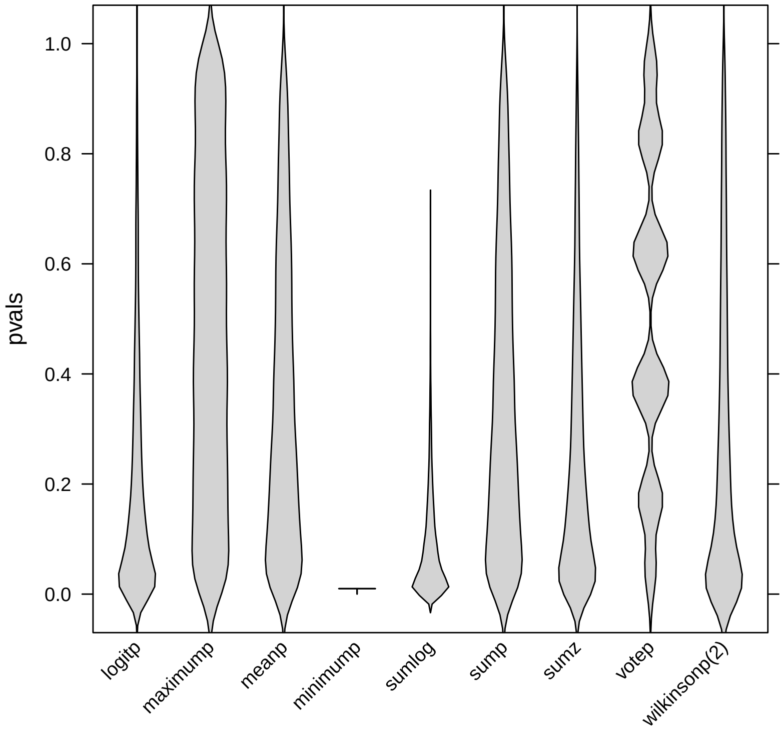

Figure 3: One extreme p and nine uniform on (0, 1)

As can be seen the methods give quite different results. The most extreme examples are `minimump` which is clearly not affected by the other values, `maximump` which is not affected by the one extreme p, and `votep` which clearly shows the discrete nature of the values returned. Both `logitp` and `sumlog` have more of their density towards the lower end and `meanp` and `sump` have similar densities. For illustration we also show `wilkinsonp` with $r = 2$.

5 Feedback

I aim to include any method for which there exists a published example against which I can test the code. I welcome feedback about such sources and any other comments about either the documentation or the code.

References

A Birnbaum. Combining independent tests of significance. *Journal of the American Statistical Association*, 49:559–574, 1954.

R D Cousins. Annotated bibliography of some papers on combining significances or p–values, 2008. arXiv:0705.2209.

T Lipták. On the combination of independent tests. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményi*, 3:171–197, 1958.

A B Owen. Karl Pearson's meta–analysis revisited. *Annals of Statistics*, 37: 3867–3892, 2009.

D V Zaykin. Optimally weighted z–test is a powerful method for combining probabilities in meta–analysis. *Journal of Evolutionary Biology*, 24:1836–1841, 2011.