The metap package

Michael Dewey

January 22, 2017

1 Introduction

1.1 What is this document for?

This document describes some methods for the meta–analysis of p–values (significance values) and their implementation in the package metap. It also contains some commentary on the performance of the various algorithms under a small number of different scenrios with some hints on the choice of method.

The problem of meta–analysis of p–values is of course not completely unconnected with the more general issue of simultaneous statistical inference.

1.2 Why and when to meta–analyse significance values

The canonical way to meta–analyse a number of primary studies is to combine estimates of effect sizes from each of them. There are a large number of packages for this purpose available from CRAN and described in the task view http://CRAN.R-project.org/view=MetaAnalysis. However sometimes the only available information may be p–values especially when some of the primary studies were published a long time ago or were published in sources which were less rigorous about insisting on effect sizes. The methods outlined here are designed for this eventuality. The situation may also arise that some of the studies can be combined in a conventional meta–analysis using effect sizes but there are many others which cannot and in that case the conventional meta–analysis of the subset of studies which do have effect sizes may usefully be supplemented by an overall analysis of the p–values.

Just for the avoidance of doubt I should point out that if each study has produced a proportion and the goal is to synthesise them to a common estimate or analyse the differences between them then the standard methods are appropriate not the ones outlined here. The p–values in this document are significance levels.

1.3 Notation

The k studies give rise to p–values, $p_i, i = 1, \dots, k$. These are assumed to be independent We shall also need the ordered p–values: $p_{[1]} \leq p_{[2]}, \dots, \leq p_{[k]}$ and weights $w_i, i = 1, \dots, k$. 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.

Function name	Description(s)	
`logitp`		Logistic
`meanp`		
`maximump`		
`minimump`	Tippett's method	
`sumlog`	Fisher's method	Chi square (2 df)
`sump`	Edgington's method	Uniform
`sumz`	Stouffer's method	Normal
`votep`		
`wilkinsonp`	Wilkinson's method	

Table 1: Methods considered in this document

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

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

Birnbaum (1954) considered the property of admissibility. 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^* where $p_i^* \leq p_i$ for all i. He considered that Fisher's and Tippett's method were admissible. See also Owen (2009).

He also points out the problem is poorly specified. This may account for the number of methods available and their differing behaviour. The null hypothesis H_0 is well defined, that all p_i have a uniform distribution on the unit interval. There are two classes of alternative hypothesis

- H_A: all p_i have the same (unknown) non–uniform, non–increasing density,
- H_B: at least one p_i has an (unknown) non–uniform, non–increasing density.

If all the tests being combined come from what are basically replicates then H_A is appropriate whereas if they are of different kinds of test or different conditions then H_B is appropriate. Note that Birnbaum specifically considers the possibility that the tests being combined may be very different for instance some tests of means, some of variances, and so on.

3 Preparation for meta–analysis of p–values

3.1 Preliminaries

I assume you have installed R and `metap`. You then need to load the package.

```
> library(metap)
```

3.2 Directionality

It is usual to have a directional hypothesis, for instance that treatment is better than control. For the methods described here a necessary preliminary is to ensure that all the p–values refer to the same directional hypothesis. If the value from the primary study is two–sided it needs to be converted. This is not simply a matter of halving the quoted p–value as values in the opposite direction need to be reversed. A convenience function `two2one` is provided for this.

```
> pvals <- c(0.1, 0.1, 0.9, 0.9, 0.9, 0.9)
> istwo <- c(TRUE,  FALSE, TRUE, FALSE, TRUE, FALSE)
> toinvert <- c(FALSE, TRUE, FALSE, FALSE, TRUE, TRUE)
> two2one(pvals, two = istwo, invert = toinvert)
```

```
[1] 0.05 0.90 0.45 0.90 0.55 0.10
```

Note in particular the way in which 0.9 is converted under the different scenarios.

3.3 Plotting

```
> print(validity)
```

```
 [1] 0.015223 0.005117 0.224837 0.000669 0.004063 0.549106 0.052925 0.024674
 [9] 0.004618 0.287803 0.738475 0.009563 0.071971 0.000003 0.001040 0.031221
[17] 0.005274 0.098791 0.067441 0.250210
```

It would be a wise precaution to examine the p–values graphically or otherwise before subjecting them to further analysis. A function `schweder` is provided for this purpose. This plots the ordered p–values, $p_{[i]}$, against i. Although the original motivation for the plot is Schweder and Spjøtvoll (1982) the function uses a different choice of axes due to Benjamini and Hochberg (2000). We will use an example dataset on the validity of student ratings quoted in Becker (1994). Figure 1 shows the plot from `schweder`.

`schweder` also offers the possibility of drawing one of a number of straight line summaries. The three possible straight line summaries are shown in Figure 2 and are:

- the lowest slope line of Benjaimin and Hochberg which is drawn by default as solid,
- a least squares line drawn passing through the point $k + 1, 1$ and using a specified fraction of the points which is drawn by default as dotted,

```
> par(pin = c(3, 3))
> schweder(validity)
```

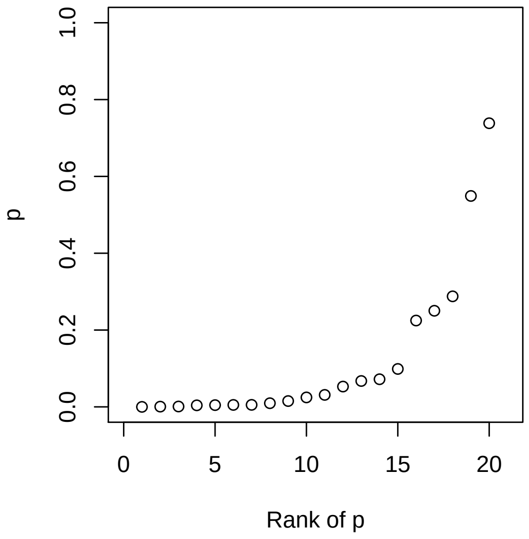

Figure 1: Simple example of plot using `schweder`

```
> par(pin = c(3, 3))
> schweder(validity, drawline = c("bh", "ls", "ab"),
+    ls.control = list(frac = 0.5), ab.control = list(a = 0, b = 0.01))
```

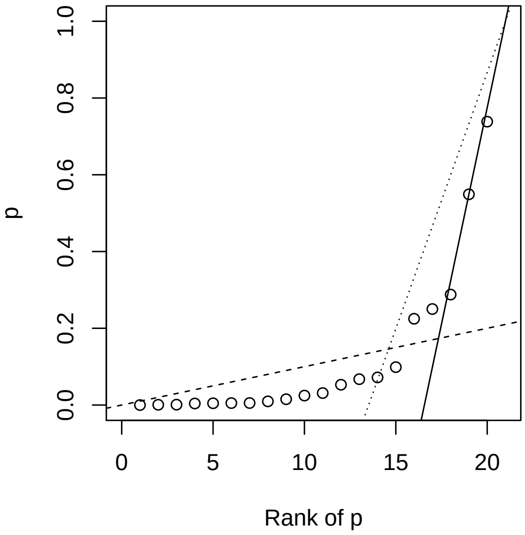

Figure 2: Example of plot with lines added

- a line with user specified intercept and slope which is drawn by default as dashed.

3.4 Reporting problems in the primary studies

Another issue is what to do with studies which have simply reported on whether a conventional level of significance like 0.05 was achieved or not. If the exact associated *p* cannot be derived from the statistics quoted in the primary source then the value of the level achieved, in this case 0.05, can be used although this may be conservative. Studies which simply report not significant could be included as having $p = 1$ (or $p = 0.5$ if it is known that the direction was right) although this is very conservative.

4 The methods

4.1 Comparison scenarios

To provide a standard of comparison we shall use the following two situations. Some authors have also used the case of exactly two p_i.

4.1.1 What if all $p_i = p$?

Perhaps surprisingly there are substantial differences here as we shall see when we look at each method. We describe how the returned value varies with p_i and k.

4.1.2 Cancellation

When the collection of primary studies contains a number of values significant in both directions for example four studies having p–values 0.001, 0.001, 0.999, 0.999 the methods can give very different results. If the intention of the synthesis is to examine a directional hypothesis one would want a method where these cancelled out. We shall use those four values as our example.

4.2 Methods using transformation of the p–values

One class of methods relies on transforming the p–values and then combining them.

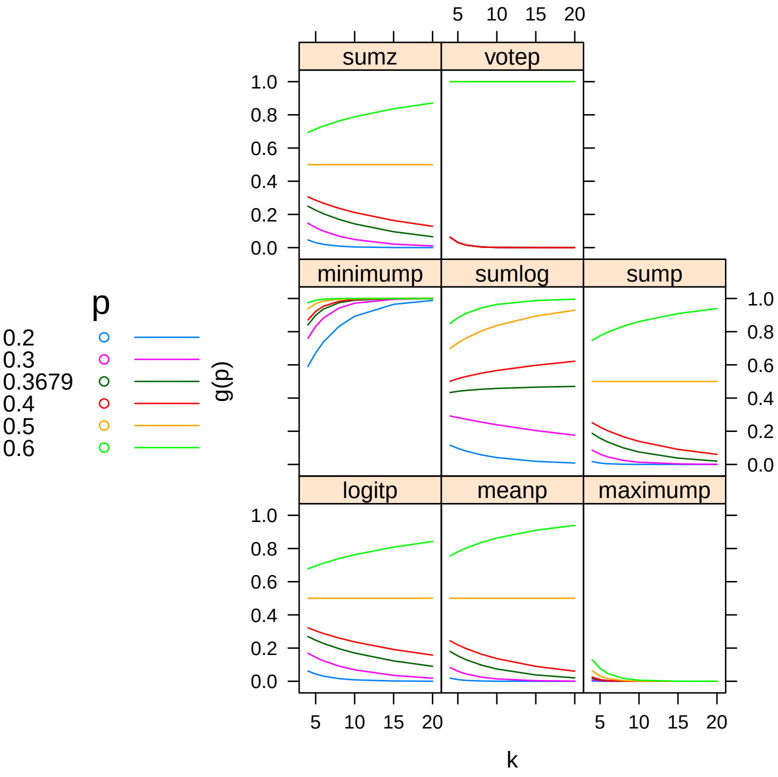

Figure 3: Behaviour of the methods for k values of $p = p_i$

4.2.1 The method of summation of logs, Fisher's method

The method relies on the fact that

$$\sum_{i=1}^{k} -2 \log p_i \tag{1}$$

is a chi-squared with $2k$ df. This works because $-2 \log p_i$ is a χ^2_2 and the sum of χ^2 is itself a χ^2 with degrees of freedom equal to the sum of the degrees of freedom of the individual χ^2. Of course the sum of the log of the p_i is also the log of the product of the p_i. Fisher's method is provided in `sumlog`.

When all the $p_i = p$ this method returns a value which decreases with k when $p < 0.32$, increases with k when $p > 0.37$, and in between increases with k and then decreases. Some detailed algebra provided in a post to stats.stackexchange.com by Christoph Hanck suggests that the breakpoint is $e^{-1} = 0.3679$ so that where the p_i are less than that then for a sufficiently large k the result will be significant and not if above that. Over the range of k we are plotting this bound is not yet closely approached. Hanck's plot suggests that k must be several hundred for this to happen.

This method does not cancel significant values in both direction and returns a significant result for our example.

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

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

It would of course be possible to generalise this to use transformation to χ^2 with any other number of degrees of freedom rather than 2. Lancaster (1961) suggests that this is highly correlated with `sumlog`.

4.2.2 The method of summation of z values, Stouffer's method

Defined as

$$\frac{\sum_{i=1}^{k} z(p_i)}{\sqrt{k}} \tag{2}$$

is a standard normal deviate where z is the quantile function of the normal distribution.

The method of summation of z values is provided in `sumz`. It returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

This method does cancel significant values in both directions.

```
> sumz(pvals)
```

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

A weighted version is available

$$\frac{\sum_{i=1}^{k} w_i z(p_i)}{\sqrt{\sum_{i=1}^{k} w_i^2}} \tag{3}$$

where w_i are the weights.

By default the weights are equal. In the absence of effect sizes (in which case a method for combining effect sizes would be more appropriate anyway) best results are believed to be obtained with weights proportional to the square root of the sample sizes (Zaykin, 2011) following Lipták (1958). At the moment weighting is only provided in `sumz` as this is the only method for which a published example is accessible.

4.2.3 The method of summation of logits

Defined as

$$-\frac{\sum_{i=1}^{k} \log \frac{p}{1-p}}{C} \tag{4}$$

is distributed as Student's t with $5k + 4$ df where

$$C = \sqrt{\frac{k\pi^2(5k+2)}{3(5k+4)}} \tag{5}$$

This method is provided in `logitp`. The constant was arrived at by equating skewness and kurtosis with that of the t–distribution (Loughin, 2004).

This method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

This method does cancel significant values in both directions.

```
> logitp(pvals)
t =  5.114524e-16  with df =  24  p =  0.5
```

4.2.4 Examples for `sumlog`, `sumz`, and `logitp`

Using the same example dataset which we have already plotted

```
> sumlog(validity)
chisq =  159.82  with df =  40  p =  2.989819e-16
> sumz(validity)
sumz =  8.186994 p =  1.339156e-16
> logitp(validity)
t =  9.521107  with df =  104  p =  3.954051e-16
```

As can be seen these are in quite good agreement.

4.3 Methods using untransformed p–values

4.3.1 The method of minimum p and Wilkinson's method

The minimum p method is usually described in terms of a rejection at the α_* level of the null hypothesis

$$p_{[1]} < 1 - (1 - \alpha_*)^{\frac{1}{k}} \tag{6}$$

The minimum p method is a special case of Wilkinson's method which uses $p_{[r]}$ where $1 \leq r \leq k$ (Wilkinson, 1951). Wilkinson's method is provided in `wilkinsonp` and a convenience function `minimump` with its own `print` method is provided for the minimum p method. It is also possible to use the method for the maximum p (that is $r = k$) and a convenience function `maximump` is provided for that purpose.

These methods return a value for our $p_i = p$ example which always increases with k which is true for `minimump` and which always decreases with k which is true for `maximump`

The minimum p method does not cancel significant values in both direction and returns a significant result for our example but the maximum p does cancel.

```
> minimump(pvals)
```

```
p =  0.003994004  using minimum p
```

```
> maximump(pvals)
```

```
p =  0.996006  using maximum p
```

4.3.2 The method of summation of p–values, Edgington's method

Define

$$S = \sum_{i=1}^{k} p_i \tag{7}$$

then this method is defined as

$$\frac{(S)^k}{k!} - \binom{k-1}{1}\frac{(S-1)^k}{k!} + \binom{k-2}{2}\frac{(S-2)^k}{k!} - \dots \tag{8}$$

where there are k studies and the series continues until the term in in the numerator $(S - i)$ becomes negative (Edgington, 1972a). This method is provided in `sump`.

This method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

This method does cancel significant values in both directions.

```
> sump(pvals)
```

```
psum =  0.5
```

Some authors use a simpler version, for instance Rosenthal (1978) in the text although compare his Table 4.

$$\frac{(\sum p)^k}{k!} \tag{9}$$

where there are k studies but this can be very conservative when $\sum p > 1$ There seems no particular need to use this method but it is returned by `sump` as the value of `conservativep` for use in checking published values.

Note also that there can be numerical problems for extreme values of S and in that case recourse might be made to `sumz` or `logitp` which have similar properties.

4.3.3 The mean p method

This is defined as

$$
\begin{aligned}
z &= (0.5 - \bar{p})\sqrt{12k} \\
\bar{p} &= \frac{\sum_{i=1}^{k} p_i}{k}
\end{aligned}
\tag{10}
$$

which is a standard normal (Edgington, 1972b) and where . Although this method is attributed to Edgington when the phrase Edgington's method is used it refers to the method of summation of p–values described above in Section 4.3.2.

This method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

This method does cancel significant values in both directions.

```
> meanp(pvals)

z =  0  p =  0.5
```

4.3.4 Examples for `minimump`, `maximump`, `sump`, and `meanp`

```
> minimump(validity)

p =  5.999829e-05  using minimum p

> maximump(validity)

p =  0.002326569  using maximum p

> sump(validity)

psum =  2.356122e-11
```

```
> meanp(validity)
```

```
z =  5.853608  p =  2.405102e-09
```

Agreement here is not so good especially for the maximump method.

4.4 Other methods

4.4.1 The method of vote–counting

A simple way of looking at the problem is vote counting. Strictly speaking this is not a method which combines p–values in the same sense as the other method. If most of the studies have produced results in favour of the alternative hypothesis irrespective of whether any of them is individually significant then that might be regarded as evidence for that alternative. The numbers for and against may be compared with what would be expected under the null using the binomial distribution. A variation on this would allow for a neutral zone of studies which are considered neither for nor against. For instance one might only count studies which have reached some conventional level of statistical significance in the two different directions.

This method returns a value for our $p_i = p$ example which is 1 above 0.5 and otherwise invariant with p but decreases with k.

This method does cancel significant values in both directions.

```
> votep(pvals)
```

```
p =  0.6875
```

4.5 Examples of `votep`

```
> votep(validity)
```

```
p =  0.0002012253
```

5 Loughin's recommendations

In his simulation study Loughin (2004) carried out extensive comparisons. He bases his recommendations on criteria of structure and the arrangement of evidence against H_0.

Under structure he considers three cases with the following recommendations: emphasis on small p–values (`sumlog` and `minimump`), emphasis on large p–values (`maximump` and `sump`), and equal emphasis (`logitp` and `sumz`).

Under arrangement of evidence he considers where this is concentrated. His recommendations are summarised in Table 2.

Equal in all tests	$k < 10$ `sump`, `maximump`
	Any k `sumz`, `logitp`
Some in all tests	$k < 10$ `sump`, `maximump`
	Any k `sumz`, `logitp`
In majority of tests	`sumz`, `logitp`
In minority of tests	Moderate or strong evidence `sumlog`
	Any power `sumz`, `logitp`
In one test only	Strong total evidence `minimup`
	Moderate total evidence `sumlog`
	Weak total evidence `sumz`, `logitp`

Table 2: Loughin's recommendations for method choice

6 Other considerations

6.1 Directionality

When the collection of primary studies contains a number of values significant in both directions we have seen that the methods can give very different results. If the intention of the synthesis is to examine a directional hypothesis one would want a method where these cancelled out. Clearly the choice should be made on scientific grounds not on the baiss of the outcome.

6.2 Extractor functions

The standard `print` and `plot` methods are provided.

6.3 Legal values for p_i

	Valid for		
	$p = 0$	$p = 1$	Notes
`logitp`	N	N	
`meanp`	Y	Y	Requires at least four studies
`sumlog`	N	Y	
`sump`	Y	Y	
`sumz`	Y	Y	
`votep`	Y	Y	
`wilkinson`	Y	Y	

Table 3: Restrictions on values of p_i

Not all methods work with $p = 0$ or $p = 1$. See Table 3 for details. If these values occur in your dataset and you do not wish the functions to take their routine action of excluding that study then you need to decide what to do. If you believe that injudicious rounding is to blame you might wish to replace zero values by the least upper bound of the values which would still round to zero to the given number of decimal places. So you might replace 0.00 with 0.005, 0.000 with 0.0005 and so on. Similar action can be taken for values given as unity.

6.4 Reading

An annotated bibliography is provided by Cousins (2008)

7 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

B J Becker. Cambining significance levels. In H Cooper and L V Hedges, editors, *A handbook of research synthesis*, chapter 15, pages 215–235. Russell Sage, New York, 1994.

Y Benjamini and Y Hochberg. On the adaptive control of the false disovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25:60–83, 2000.

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.

E S Edgington. An additive method for combining probability values from independent experiments. *Journal of Psychology*, 80:351–363, 1972a.

E S Edgington. A normal curve method for combining probability values from independent experiments. *Journal of Psychology*, 82:85–89, 1972b.

H Lancaster. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3:20–33, 1961.

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

T M Loughin. A systematic comparison of methods for combining p–values from independent tests. *Computation Statistics and Data Analysis*, 47: 467–485, 2004.

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

R Rosenthal. Combining results of independent studies. *Psychological Bulletin*, 85:185–193, 1978.

T Schweder and E Spjøtvoll. Plots of p–values to evaluate many tests simultaneously. *Biometrika*, 69:493–502, 1982.

B Wilkinson. A statistical consideration in psychological research. *Psychological Bulletin*, 48:156–158, 1951.

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.