The `metap` package

Michael Dewey

July 25, 2018

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 comments on the performance of the various algorithms under a small number of different scenarios with hints on the choice of method. I welcome feedback about sources of published examples against which I can test the code and any other comments about either the documentation or the code.

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 uses 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, 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, $p_i,\ i = 1, \ldots, k$. These are assumed to be independent. We shall also need the ordered p–values: $p_{[1]} \leq p_{[2]}, \ldots, \leq p_{[k]}$ and weights $w_i,\ i = 1, \ldots, k$. Logarithms are natural. A function for combining p–values is denoted g. The size of the test is α. We may also need k degrees of freedom, ν_i.

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)	
	Eponym	
`invchisq`	Lancaster's method	Inverse chi square
`invt`		Inverse t
`logitp`		Logistic
`meanp`		
`meanz`		
`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 methods. A detailed account was provided by Lipták (1958).

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 (a) shows the plot from schweder.

```
> schweder(validity)
```

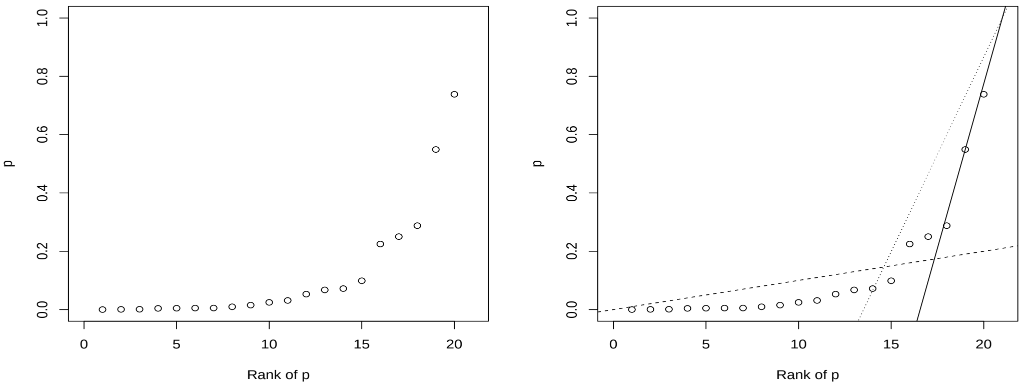
schweder also offers the possibility of drawing one of a number of straight line summaries. The three possible straight line summaries are shown in Figure (b) and are:

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

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

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



(a) Simple graph

(b) With lines

Figure 1: Ouput from schweder

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. The theory of handling p–values which have been truncated like this has been developed by Zaykin et al. (2002) and is available in the CRAN package `TFisher`.

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.

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

Cancellation When the collection of primary studies contains a number of values significant in both directions 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. The decision between methods should be made on theoretical grounds of course. We shall use the following four values as our example.

```
> cancel <- c(0.001, 0.001, 0.999, 0.999)
```

4.2 Methods using transformation of the p–values

One class of methods relies on transforming the p–values first.

Function name	Definition	Critical value
`invchisq`	$\sum_{i=1}^{k} \chi^2_{\nu_i}(p_i)$	$\chi^2_{\sum \nu_i}(\alpha)$
`invt`	$\frac{\sum_{i=1}^{k} t_{\nu_i}(p_i)}{\sqrt{\sum_{i=1}^{k} \frac{\nu_i}{\nu_i-2}}}$	$z(\alpha)$
`logitp`	$\frac{\sum_{i=1}^{k} \log \frac{p}{1-p}}{C}$ $C = \sqrt{\frac{k\pi^2(5k+2)}{3(5k+4)}}$	t_{5k+4}
`meanz`	$\frac{\bar{z}}{s_{\bar{z}}}$ $\bar{z} = \sum_{i=1}^{k} \frac{z(p_i)}{k}$ $s_{\bar{z}} = \frac{s_z}{\sqrt{k}}$	$t_{k-1}(\alpha)$
`sumlog`	$\sum_{i=1}^{k} -2 \log p_i$	$\chi_{2k}(\alpha)$
`sumz`	$\frac{\sum_{i=1}^{k} z(p_i)}{\sqrt{k}}$	$z(\alpha)$

Table 2: Definitions of methods using transformation of the p values

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

See Table 2 for the definition. 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 (Fisher, 1925) is provided in `sumlog`. 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`. Lancaster's method is provided in `invchisq`. In fact the resemblance to `sumlog` becomes less as the number of degrees of freedom increases.

As can be seen in Figure 2 when all the $p_i = p$ `sumlog` 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 https://stats.stackexchange.com/questions/243003 by Christoph Hanck suggests that the breakpoint is $e^{-1} = 0.3679$. Where the p_i are less than this then for a sufficiently large k (several hundred) 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.

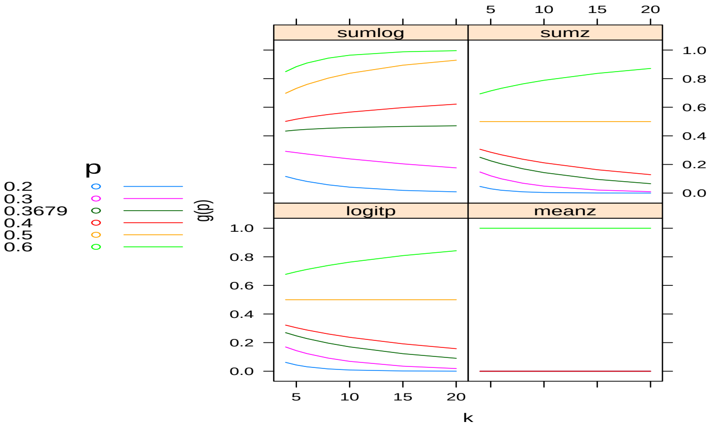


Figure 2: Behaviour of the methods using transformed p values for k values of $p = p_i$

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

The method of summation of z values is provided in `sumz` (Stouffer et al., 1949). See Table 2 for the definition. As can be seen in Figure 2 it returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above. There is also a closely related method using the mean of normals provided in `meanz` also defined in Table 2 which has very similar properties except that when all the p_i are equal it either gives 0 or 1 as can be seen in Figure 2.

A weighted version of Stouffer's method is available $\frac{\sum_{i=1}^{k} w_i z(p_i)}{\sqrt{\sum_{i=1}^{k} w_i^2}}$ where w_i are the weights. In the absence of effect sizes (in which case a method using 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).

4.2.3 The inverse t method

A closely related method is the inverse t method. See Table 2 for the definition. This method is provided in `invt`. As is clear from the definition this method tends to Stouffer's method as $\nu_i \to \infty$.

4.2.4 The method of summation of logits

See Table 2 for the definition. This method is provided in `logitp`. The constant C was arrived at by equating skewness and kurtosis with that of the t–distribution (Loughin, 2004). As can be seen in Figure 2 this method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

4.2.5 Examples for methods using transformations of the p values

Function name	validity	cancel
`logitp`	3.95405066641324e-16	0.5
`meanz`	8.09658964578493e-12	0.5
`sumlog`	2.98981918888483e-16	0.000548861519997557
`sumz`	1.33915623099787e-16	0.5

Table 3: Examples of methods using transformation of the p values

Using the same example dataset which we have already plotted and our cancellation dataset we have the values in Table 3. As can be seen all the methods cancel except for `sumlog`. The agreement for the validity dataset is close. Lancaster's method and inverse t are not shown as they are both infinite families of possible methods.

4.3 Methods using untransformed p–values

Function name	Definition	Critical value
`meanp`	$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k}$	
	$z = (0.5 - \bar{p})\sqrt{12k}$	$z(\alpha)$
`minimump`	$p_{[1]}$	$1 - (1 - \alpha)^{\frac{1}{k}}$
`maximump`	$p_{[k]}$	α^k
`wilkinsonp`	$p_{[r]}$	$\sum_{s=r}^{k} \binom{k}{s} \alpha^s (1 - \alpha)^{k-s}$
`sump`	$\frac{(S)^k}{k!} - \binom{k-1}{1}\frac{(S-1)^k}{k!} + \binom{k-2}{2}\frac{(S-2)^k}{k!} - \ldots$	α
	$S = \sum_{i=1}^{k} p_i$	

Table 4: Definitions of methods not using transformation of the p values,the summation in the numerator of `sump` continues until the term in in the numerator $(S - i)$ becomes negative

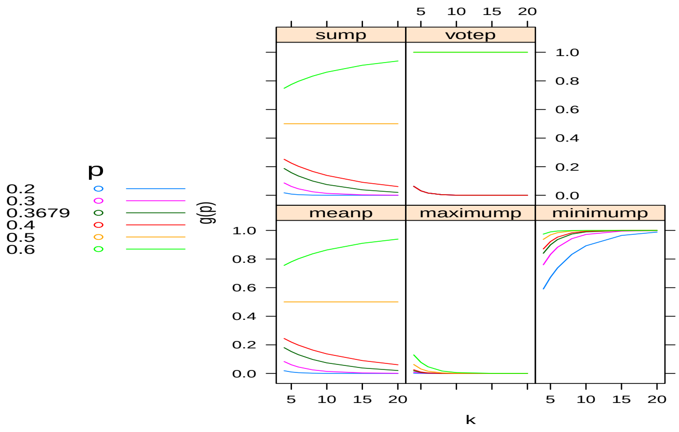


Figure 3: Behaviour of the methods using untransformed p values for k values of $p = p_i$

4.3.1 The method of minimum p, maximum p, and Wilkinson's method

The methods of minimum p (Tippett, 1931), maximum p and Wilkinson (Wilkinson, 1951) are defined in Table 4. Wilkinson's method depends on which value (the rth) of $p_{[i]}$ is selected. Wilkinson's method is provided in `wilkinsonp` and a convenience function `minimump` with its own `print` method is provided for the minimum p method ($r = 1$). 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.

As can be seen in Figure 3 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`

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

Defined in Table 4 (Edgington, 1972a). This method is provided in `sump`. As can be seen in Figure 3 this method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

Some authors use a simpler version, $\frac{(\sum p)^k}{k!}$, for instance Rosenthal (1978) in the text although compare his Table 4. 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

Defined in Table 4. Although this method is attributed to Edgington (Edgington, 1972b) 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. As can be seen in Figure 3 this method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

4.3.4 Examples for methods using untransformed p–values

Using the same example dataset which we have already plotted and our cancellation dataset we have the values in Table 5. As can be seen `meanp` and `sump` cancel but the other two do not. Agreement here is not so good especially for the maximump method. Wilkinson's method not shown as it depends on the value of r.

Function name	validity	cancel
`minimump`	5.99982900307796e-05	0.003994003999
`maximump`	0.00232656906767947	0.996005996001
`meanp`	2.40510184300166e-09	0.5
`sump`	2.3561224666017e-11	0.5

Table 5: Examples for methods using the untransformed p values

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.

Function name	validity	cancel
`votep`	0.000201225280761719	0.6875

Table 6: Examples for vote counting

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 7.

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 7: Loughin's recommendations for method choice

6 Miscellanea

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

Reading An annotated bibliography is provided by Cousins (2008)

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.

R A Fisher. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh, 1925.

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.

S A Stouffer, E A Suchman, L C DeVinney, S A Star, and R M Jnr Williams. *The American soldier, vol 1: Adjustment during army life.* Princeton University Press, Princeton, 1949.

L H C Tippett. *The methods of statistics.* Williams and Norgate, London, 1931.

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.

D V Zaykin, L A Zhivotovsky, P H Westfall, and B S Weir. Truncated product method for combining p–values. *Genetic Epidemiology*, 22:170–185, 2002.