Comparison of methods in the `metap` package

Michael Dewey

April 29, 2026

1 Introduction

1.1 What is this document for?

This document describes some methods for the meta–analysis of p–values (significance values) contained in the package `metap` and contains comments on the performance of the various algorithms under a small number of different scenarios with hints on the choice of method.

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 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.

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).

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
`truncated`	Truncated Fisher	
`truncated`		rank–truncated
`votep`		
`wilkinsonp`	Wilkinson's method	

Table 1: Methods considered in this document

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 The methods

3.1 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)
```

3.2 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`	$\frac{\sum_{i=1}^{k} \log \frac{p}{1-p}}{C}$ where $C = \sqrt{\frac{k\pi^2(5k+2)}{3(5k+4)}}$	t_{5k+4}
`meanz`	$\frac{\bar{z}}{s_{\bar{z}}}$ where $\bar{z} = \sum_{i=1}^{k} \frac{z(p_i)}{k}$ and $s_{\bar{z}} = \frac{s_z}{\sqrt{k}}$	$t_{k-1}(\alpha)$
`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

3.2.1 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`.

As can be seen in Figure 1 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.

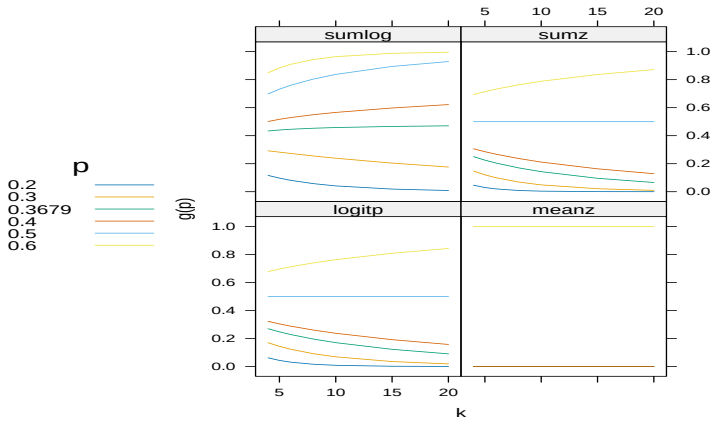


Figure 1: Behaviour of the methods using transformed p values for k values of $p = p_i$

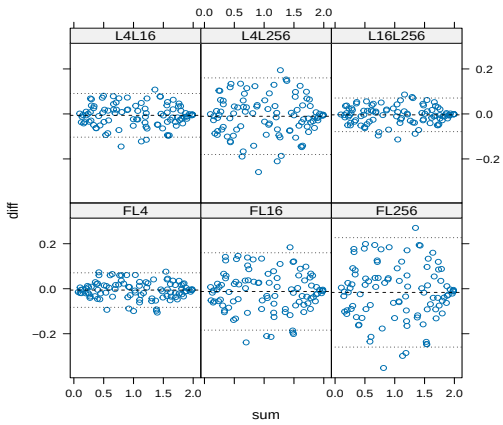
3.2.2 Inverse χ^2 Lancaster's method

It would of course be possible to generalise Fisher's method 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. Figure 2a shows for a small number of selected degrees of freedom how it compares to Fisher's method.

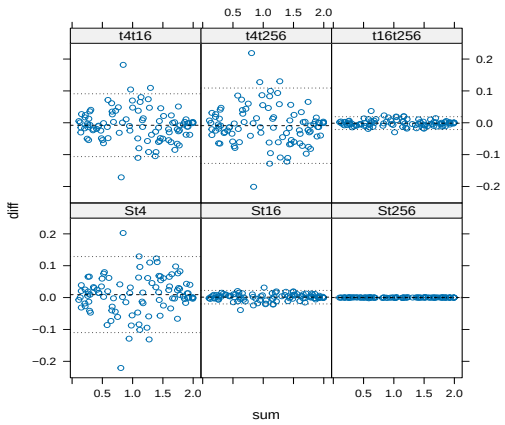
3.2.3 The method of summation of z values, Stouffer's method

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

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



(a) Fisher's method and Lancaster's method



(b) Stouffer's method and inverse t

Figure 2: Sum and difference plots of Fisher v Lancaster and Stouffer v inverse t

be obtained with weights proportional to the square root of the sample sizes (Zaykin, 2011) following Lipták (1958).

3.2.4 Mean of normals method

There is also a method closely related to Stouffer's 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 1.

```
> meanz(c(0.3, 0.31))$p
```

```
[1] 5.581505e-280
```

```
> meanz(c(0.1, 0.2))$p
```

```
[1] 6.959644e-07
```

The method of `meanz` also has the unusual property that a set of p–values which are all less than those in another set can still give rise to a larger overall p. See example above. This is the only method considered here which has this property so if it is a desirable one then that is the only method to consider.

3.2.5 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 $\nu_i \to \infty$. Figure 2b shows this for selected degrees of freedom.

3.2.6 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 1 this method returns a value for our $p_i = p$ example which decreases with k when p below 0.5 and increases above.

3.2.7 Examples for methods using transformations of the p values

Function name	validity value expressed as $-\log_{10} p$	cancel
`logitp`	15.4	0.5
`meanz`	11.09	0.5
`sumlog`	15.52	0.00055
`sumz`	15.87	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 except for `meanz` whoch gives a value several orders of magnitude greater than the other three. Lancaster's method and inverse t are not shown as they are both infinite families of possible methods and in any event are similar to Fisher's method and Stouffer's method respectively.

3.3 Methods using untransformed p–values

3.3.1 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

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`	$\frac{(S)^k}{k!} - \binom{k}{1}\frac{(S-1)^k}{k!} + \binom{k}{2}\frac{(S-2)^k}{k!} - \ldots$ where $S = \sum_{i=1}^{k} p_i$	α

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

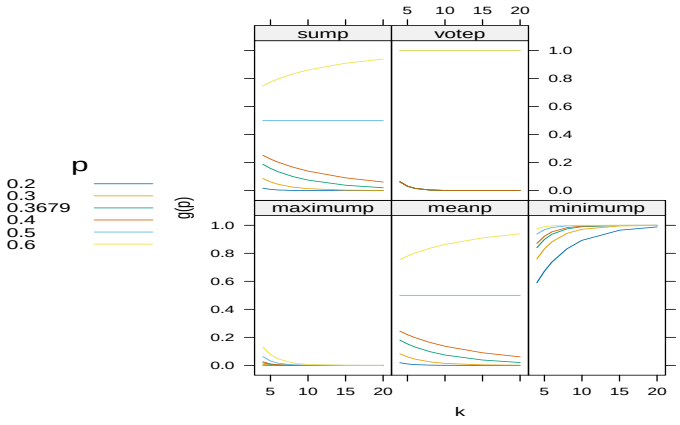


Figure 3: Behaviour of the methods using untransformed p values for k values of $p = p_i$

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`

3.3.2 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 `meanp` which has similar properties.

3.3.3 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 3.3.2. 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.

Not surprisingly this method gives very similar results to Edington's other method implemented in `sump` and since it does not have the numerical problems of that method it might perhaps be preferred.

3.3.4 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 maximum p method. Wilkinson's method not shown as it depends on the value of r.

Function name	validity value expressed as $-\log_{10} p$	cancel
`minimump`	4.22	0.00399
`maximump`	2.63	0.99601
`meanp`	8.62	0.5
`sump`	10.63	0.5

Table 5: Examples for methods using the untransformed p values

3.4 Other methods

3.4.1 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 methods. 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 for p values 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`	0.000201	0.6875

Table 6: Examples for vote counting

3.4.2 Methods not using all p–values

If there is a hypothesis that the signal will be concentrated in only a few p–values then alternative methods are available in `truncated`. This is a wrapper to two packages available on CRAN: `TFisher` which provides the truncated Fisher method (Zaykin et al., 2007; Zhang et al., 2018) and `mutoss` which provides the rank–truncated Fisher method (Dudbridge and Koeleman, 2003). Note that Table 7 only shows results for the validity data–set

as, since the methods explicitly only consider results in one direction the cancellation issue does not arise.

Function name	truncated at $p = 0.5$	truncated at rank = 5
`truncated`	15.48	8.06

Table 7: Examples for truncated using the validity data–set expressed as $-\log_{10} p$

Dudbridge and Koeleman (2003) compare these two methods. They comment that in a meta–analysis the method using truncated Fisher may be preferred particularly if reporting bias is suspected. When the interest is in a small set of signals in the presence of much noise as occurs in genome wide association scanning then the rank truncation method is recommended. Their article provides more details about the methods.

4 Loughin's recommendations

In his simulation study Loughin (2004) carried out extensive comparisons. Note that he did not consider all the methods implemented here. These omissions are not too important for our purposes. The methods implemented here as `invchisq`, `invt`, `meanp` and `meanz` are all very similar to ones which he did study. The truncation methods appeared about the same time as his work but in any case are fundamentally different. Vote counting is arguably not a method of the same sort.

As Loughin points out the first thing to consider is whether large p–values should cancel small ones. If this is not desired then the only methods to consider are those in `sumlog` (Fisher), `minimump` (Tippett) and `maximump`.

He bases his recommendations on criteria of structure and the arrangement of evidence against H_0. Figure 4 shows a summary of his recommendations about the structure of the evidence.

Figure 5 summarise his recommendations about the arrangement of evidence.

Overall he considered the choice to lie between Stouffer's method, Fisher's method and the logistic method implemented in `logitp`. As has already been mentioned Fisher's method cancels whereas the other two do not so if the weak evidence in a small number of p–values is not to be over–whelmed by the others then Fisher is the best choice. However where the evidence is more evenly spread Stouffer's method may be preferred. The logistic method represents a compromise between them and is perhaps best suited where

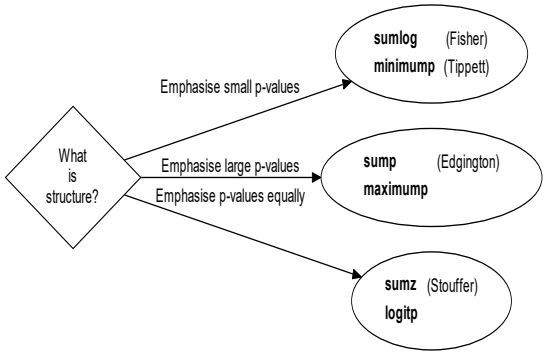


Figure 4: Loughin's recommendations based on structure

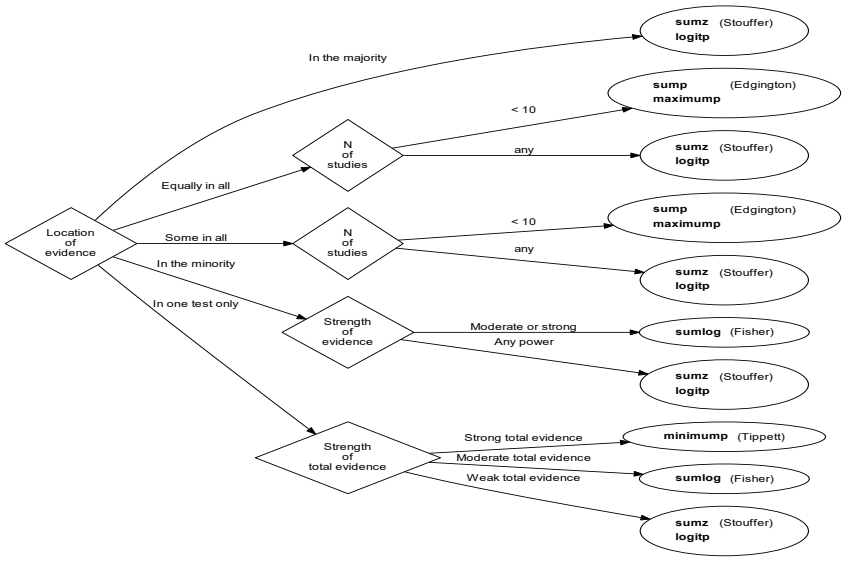


Figure 5: Loughin's recommendations based on where the strength of the evidence is located

the pattern of evidence is not clear in advance. The other methods are not universally ruled out and may be helpful in the specific circumstance outlined in his summaries.

References

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

F Dudbridge and B P C Koeleman. Rank truncated product of p–values, with application to genomewide association scans. *Genetic Epidemiology*, 25:360–366, 2003.

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.

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, W Czika, S Shao, and R D Wolfinger. Combining p–values in large–scale genomics experiments. *Pharmaceutical Statistics*, 6:217–236, 2007.

H Zhang, T Tong, J Landers, and Z Wu. TFisher tests: optimal and adaptive thresholding for combining p–values. *arXiv*, 2018. URL `arXiv:1801.04309`.