Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.
https://simfit.org.uk
https://simfit.silverfrost.com

Meta analysis is widely used in areas such as evidence based medicine in order to examine several studies of the same problem by different analysts, then extract the most plausible and objective overall conclusions. One common situation is where there are k alternative 2 by 2 contingency tables available, and worked examples to demonstrate the options available in SIMFIT to analyze this type of data set will now be presented.

Open the SIMFIT main menu, choose [Statistics], [Analysis of proportions], then [Meta Analysis] and examine the default test file `meta.tf1` which is formatted as follows.

y	N	x
126	226	1
35	96	1
908	1596	2
497	1304	2
913	1660	3
336	934	3
235	407	4
58	179	4
402	710	5
121	336	5
182	338	6
72	170	6
60	159	7
11	54	7
104	193	8
21	57	8

The format for SIMFIT meta analysis data files must be exactly as now summarized.

- The number of rows in the data matrix must be an even number.
- Distinct 2 by 2 contingency tables are included as sequential pairs of adjacent rows.
- Column 1 at row i must contain the number of critical outcomes $y_i \geq 0$, e.g. successful recovery.
- Column 2 at row i must contain the total number of observations $N_i \geq y_i$, and not $N_i - y_i$ which would be the complement of y_i, i.e. the number of failures to respond to treatment.
- Column 3 at row i must contain the control variable x for use in plotting.

Note that control variable x is not used in subsequent calculations, and it is only used for identifying the adjacent 2 by 2 contingency tables, and as a coordinate for plotting, which will be explained subsequently. Obviously, the value of x in rows j and $j + 1$ must be the same for $j = 1, 3, \ldots, k - 1$.

For instance, the first 2 by 2 contingency table that can be constructed from the data set is

126	100	*and not*	126	226
35	61		35	96

so we would have the probability estimates $\hat{p}_1 = 126/226$ and $\hat{p}_2 = 35/96$ and the odds ratio for this 2 by 2 contingency table would then be

$$2.196 = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}.$$

Reading in this data set produces the following summary table.

To test H_0: equal binomial p-values

Number of 2 by 2 tables	8	
Overall sum of Y	4081	
Overall sum of N	8419	
Overall estimate of p	0.4847	95% confidence limits = (0.4740,0.4955)
$-2\log\lambda \quad (-2LL)$	310.9	$NDOF = 15$
$P(\chi^2 \ge -2LL)$	0.0000	*Reject H_0 at 1% significance level*
Chi-square test statistic (C)	306.9	$NDOF = 15$
$P(\chi^2 \ge C)$	0.0000	*Reject H_0 at 1% significance level*

Subsequent analysis leads to these results

Cochran-Mantel-Haenszel 2 by 2 by k Meta Analysis

y	N	Odds Ratio	$E[n(1,1)]$	$Var[n(1,1)]$
126	226	2.19600	113.00000	16.89720
35	96			
908	1596	2.14296	773.23448	179.30144
497	1304			
913	1660	2.17526	799.28296	149.27849
336	934			
235	407	2.85034	203.50000	31.13376
58	179			
402	710	2.31915	355.00000	57.07177
121	336			
182	338	1.58796	169.00000	28.33333
72	170			
60	159	2.36915	53.00000	9.00000
11	54			
104	193	2.00321	96.50000	11.04518
21	57			

H_0: conditional independence (all odds ratios = 1)
CMH Test Statistic = 279.4
$P(\chi^2 \ge CMH)$ = 0.0000 *Reject H_0 at 1% significance level*
Common Odds Ratio = 2.174, 95% confidence limits = (1.914,2.471)

Overall 2 by 2 contingency table

y	$N-y$
2930	2359
1151	1979

Overall Odds Ratio = 2.1360, 95% confidence limits = (1.950, 2.338)

The default log-odds plot for these 2 by 2 contingency tables can be easily viewed but to perform the editing necessary to create the next plot the following procedure has to be used.

1. Read in data and perform the meta analysis.

2. Display the default log odds plot using logarithms to base e or 10 as required.

3. Choose the [Advanced] option.

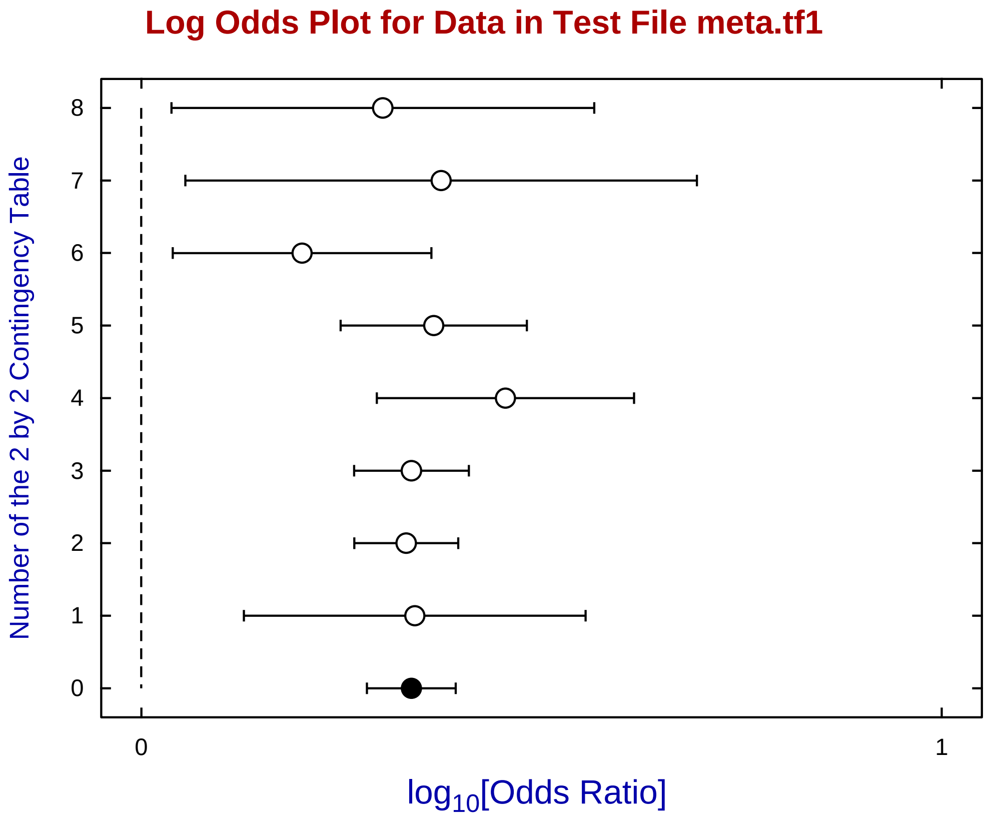
4. Select the [Avanced editing] option to transfer the data into the **simplot** procedure.

5. Note: this always transfers data into **simplot** in original not transformed coordinates.

6. Select the [Tansform] option, then the reverse y-semilog transformation.

7. The [Titles], [Labels], and [Legends] options can then be used for fine tuning as required.

Note that the sold circle represents the overall log odds ratio, while the dotted vertical line represents the reference position corresponding to the special case $p_1 = p_2$, which serves to indicate orders of magnitude deviation of the odds from the ideal case where the Odds = 1. As the Odds are all greater than 1 with these data, the points displayed all lie to the right of this reference line.



Various other tables can be displayed, such as the next one which summarizes the differences and also calculates NNT, the approximate number needed to treat in order to score another success, along with very approximate 95% confidence limits.

$d_{i,j} = \hat{p}_i - \hat{p}_j, \quad NNT = 1/|d_{i,j}|$

Row(i)	Row(j)	$d_{i,j}$	lower-95%	upper-95%	Conclusion	$Var(d_{i,j})$	NNT	(95%c.l.)
1	2	0.19294	0.07691	0.30897	$p_1 > p_2$	0.00350	6	(3,14)
3	4	0.18779	0.15194	0.22364	$p_3 > p_4$	0.00033	6	(4,7)
5	6	0.19026	0.15127	0.22924	$p_5 > p_6$	0.00040	6	(4,7)
7	8	0.25337	0.16969	0.33706	$p_7 > p_8$	0.00182	4	(2,6)
9	10	0.20608	0.14312	0.26903	$p_9 > p_{10}$	0.00103	5	(3,7)
11	12	0.11493	0.02360	0.20626	$p_{11} > p_{12}$	0.00217	9	(4,43)
13	14	0.17365	0.04245	0.30486	$p_{13} > p_{14}$	0.00448	6	(3,24)
15	16	0.17044	0.02682	0.31406	$p_{15} > p_{16}$	0.00537	6	(3,38)

Zero cells

Contingency table analysis is compromised when cells have zero frequencies, as many of the usual summary statistics become undefined. Structural zeros can be handled by applying loglinear GLM analysis but sampling zeros presumably arise from small samples with extreme probabilities. Such tables can be analyzed by exact methods, but usually a positive constant is added to all the frequencies to avoid the problems.

The next table illustrates how this problem is handled in SimFIT when analyzing data in the test file `meta.tf4`; the correction of adding 0.01 to all contingency tables frequencies being indicated.

Values ranging from 0.00000001 to 0.5 have been suggested elsewhere for this purpose, but all such choices are a compromise and, if possible, sampling should be continued until all frequencies are nonzero.

```
Cochran-Mantel-Haenszel 2 x 2 x k Meta Analysis
 y       N      Odds Ratio    E[n(1,1)]          Var[n(1,1)]
---------------------------------------------------------------
*** 0.01 added to all cells for next calculation
 0       6        0.83361      0.01091              0.00544
 0       5
*** 0.01 added to all cells for next calculation
 3       6      601.00000      1.51000              0.61686
 0       6
*** 0.01 added to all cells for next calculation
 6       6     1199.00995      4.01000              0.73008
 2       6
*** 0.01 added to all cells for next calculation
 5       6        0.00825      5.51000              0.25454
 6       6
*** 0.01 added to all cells for next calculation
 2       2        0.40120      2.01426              0.00476
 5       5
H0: conditional independence (all odds ratios = 1)

CMH Test Statistic = 386.2
P(χ^2 ≥ CMH) = 0.0494, Reject H0 at 5% significance level
Common Odds Ratio = 6.749, 95% confidence limits = (1.144, 39.81)

Overall 2 by 2 table
 y     N - y
16      10
13      15
Overall Odds Ratio = 1.842, 95% confidence limits = (0.6241,5.435)
```

Creating composite log odds plots

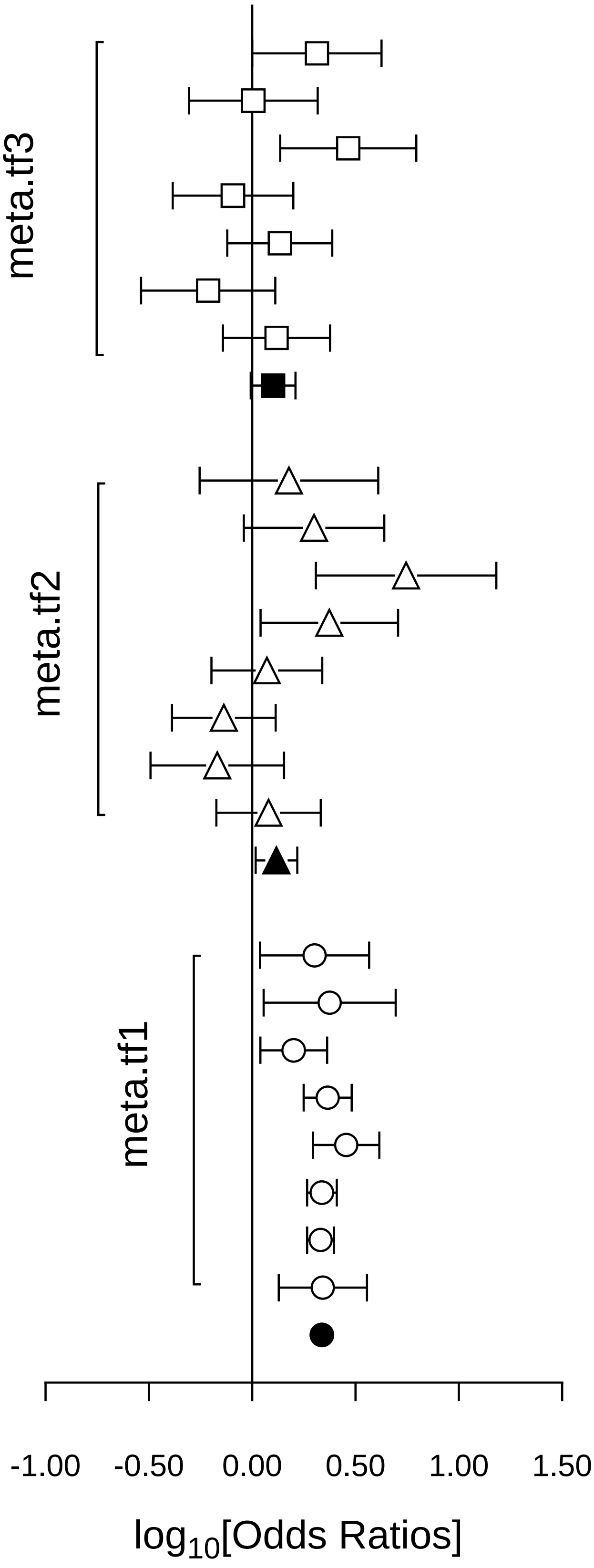
It is often necessary to create extensive log odds plots for three main reasons.

1. A single large data set is presented for analysis.
   This presents no problems if the control variables have been set correctly. However, if the graph becomes crowded it will need to be stretched.

2. Several data sets are available.
   These can be combined into a single data set by copying and pasting, or by using the SimFIT program **editmt**. However the control variables must already be consistent for this purpose or can be made so by editing at the same time.

3. Several individual log odds plots are available.
   In this case individual coordinate files can be saved then combined as a library file for SimFIT program **simplot** to make a composite plot. For this purpose the control variables on the individual data sets must be consistent to control spacing.

To illustrate these issues of spacing and stretching a worked example follows.



(1) **The data**

Test files `meta.tf1`, `meta.tf2`, and `meta.tf3` were analyzed in sequence using the SIMFIT Meta Analysis procedure. Note that, in these files, column 3 contains spacing coordinates so that data will be plotted consecutively.

(2) **The ASCII coordinate files**

During Meta Analysis, $100(1-\alpha)\%$ confidence limits on the Log-Odds-Ratio resulting from a 2 by 2 contingency tables with cell frequencies n_{ij} can be constructed from the approximation $\hat{e}$ where

$$\hat{e} = Z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

When Log-Odds-Ratios with error bars are displayed, the overall values (shown as filled symbols) with error bars are also plotted with a x coordinate one less than smallest x value on the input file. For this figure, error bar coordinates were transferred into the project archive using the [Advanced] option to save ASCII coordinate files.

(3) **Creating the composite plot**

Program **simplot**was opened and the six error bar coordinate files were retrieved from the project archive. Experienced users would do this more easily using a library file of course. Reverse y-semilog transformation was selected, symbols were chosen, axes, title, and legends were edited, then half bracket hooks identifying the data were added as arrows and extra text.

(4) **Creating the PostScript file**

Vertical format was chosen then, using the option to stretch PostScript files, the y coordinate was stretched by a factor of two.

(5) **Editing the PostScript file**

To create the final PostScript file for LaTeX a tighter bounding box was calculated using **gsview**then, using **notepad**, clipping coordinates at the top of the file were set equal to the BoundingBox coordinates, to suppress excess white space. This can also be done using the [Style] option to omit painting a white background, so that PostScript files are created with transparent backgrounds, i.e. no white space, and clipping is irrelevant.

Theory

A pair of success/failure classifications with y successes in N trials, i.e. with frequencies $n_{11} = y_1$, $n_{12} = N_1 - y_1$, $n_{21} = y_2$, and $n_{22} = N_2 - y_2$, results in a 2 by 2 contingency table, and meta analysis is used for exploring k sets of such 2 by 2 contingency tables. That is, each row of each table is a pair of numbers of successes and number of failures, so that the Odds ratio in contingency table k can be defined as

$$\begin{aligned}\text{Odds ratio}_k &= \frac{y_{1k}/(N_{1k} - y_{1k})}{y_{2k}/(N_{2k} - y_{2k})}\\ &= \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}.\end{aligned}$$

Typically, the individual contingency tables would be for partitioning of groups before and after treatment, and a common situation would be where the aim of the meta analysis would be to assess differences between the results summarized in the individual contingency tables, or to construct a best possible Odds ratio taking into account the sample sizes for appropriate weighting. Suppose, for instance, that contingency table number k is

n_{11k}	n_{12k}	n_{1+k}
n_{21k}	n_{22k}	n_{2+k}
n_{+1k}	n_{+2k}	n_{++k}

where the marginals are indicated by plus signs in the usual way. Then, assuming conditional independence and a hypergeometric distribution, the mean and variance of n_{11k} are given by

$$\begin{aligned}E(n_{11k}) &= n_{1+k}n_{+1k}/n_{++k}\\ V(n_{11k}) &= \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k} - 1)},\end{aligned}$$

and, to test for significant differences between m contingency tables, the Cochran-Mantel-Haenszel test statistic CMH, given by

$$CMH = \frac{\left\{\left|\sum_{k=1}^{m}(n_{11k} - E(n_{11k}))\right| - \frac{1}{2}\right)^2}{\sum_{k=1}^{m} V(n_{11k})}$$

can be regarded as an approximately chi-square variable with one degree of freedom. Some authors omit the continuity correction and sometimes the variance estimate is taken to be

$$\hat{V}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^3.$$

The estimated common odds ratio $\hat{\theta}_{MH}$ presented in the previous tables is calculated allowing for random effects using

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{m}(n_{11k}n_{22k}/n_{++k})}{\sum_{k=1}^{m}(n_{12k}n_{21k}/n_{++k})},$$

while the variance is used to construct the confidence limits from

$$\begin{aligned}\hat{\sigma}^2[\log(\hat{\theta}_{MH})] = {} & \frac{\displaystyle\sum_{k=1}^{m}(n_{11k}+n_{22k})n_{11k}n_{22k}/n_{++k}^2}{2\left(\displaystyle\sum_{k=1}^{m} n_{11k}n_{22k}/n_{++k}\right)^2} \\ & + \frac{\displaystyle\sum_{k=1}^{m}[(n_{11k}+n_{22k})n_{12k}n_{21k} + (n_{12k}+n_{n21k})n_{11k}n_{22k}]/n_{++k}^2}{2\left(\displaystyle\sum_{k=1}^{m} n_{11k}n_{22k}/n_{++k}\right)\left(\displaystyle\sum_{k=1}^{m} n_{12k}n_{21k}/n_{++k}\right)} \\ & + \frac{\displaystyle\sum_{k=1}^{m}(n_{12k}+n_{21k})n_{12k}n_{21k}/n_{++k}^2}{2\left(\displaystyle\sum_{k=1}^{m} n_{12k}n_{21k}/n_{++k}\right)^2}.\end{aligned}$$

Also, in these tables, the overall 2 by 2 contingency table using the pooled sample assuming a fixed effects model is listed for reference, along with the overall odds ratio and estimated confidence limits calculated using the expressions presented elsewhere for an arbitrary log odds ratio.

The table of differences illustrates another technique to study sets of 2 by 2 contingency tables. SIMFIT can calculate all the standard probability statistics for sets of paired experiments. In this case the pairwise differences are illustrated along with the number needed to treat i.e. $NNT = 1/|d|$, but it should be remembered that such estimates have to be interpreted with care. For instance, the differences and log ratios change sign when the rows are interchanged.

Again, it should be emphasized that SIMFIT outputs values and confidence limits both for the differences $d_{1,2} = \hat{p}_1 - \hat{p}_2$ and the calculated $NNT = 1/d_{1,2}$ values, but the choice between these quantities for data interpretation is controversial. To appreciate the reason why a value of NNT calculated from a sample is just a coarse estimate of the size of a sample needed to treat in order to obtain one additional cure, and could be very misleading, consider the situation of binomial trials with exactly known probabilities p_1 and p_2, and $p_1 > p_2$. The condition that the expectation of a binomial variable X_1 with probability p_1 should be one greater than than a binomial variable X_2 with probability p_2 given a sample size N is

$$\begin{aligned}E(X_1) &= E(X_2) + 1 \\ Np_1 &= Np_2 + 1, \text{ so that} \\ N &= \frac{1}{p_1 - p_2}.\end{aligned}$$

Of course NNT calculated from data is not the exact N as just derived but is given by the random function

$$NNT = \frac{1}{\hat{p}_1 - \hat{p}_2}$$

where there is experimental uncertainty in the parameter estimates. This is one reason why many experts recommend relying on conclusions based directly on the difference $d_{1,2}$, because this quantity is more robust for the purpose of hypothesis testing than NNT where reciprocation exaggerates random effects. Another reason is that it is possible to calculate accurate confidence limits for the difference $d_{1,2}$, but confidence limits calculated for NNT are unsymmetrical and much less intuitive. It just seems more informative to say, for instance, that with a possible error of up to 5%, a treatment improves the chance of cure from approximately 10 to 20%, or say from 60 to 70%, than to simply report NNT = 10 to cover all possible 10% improvements