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

If there only two categories, such as success or failure, male or female, dead or alive, etc., the data are referred to as dichotomous, and there is only one parameter to consider. So the analysis of two–category data is based on the binomial distribution which is required when y events (e.g., successes) have been recorded in N independent trials with constant probability of success (i.e., Bernoulli trials) and it is wished to explore possible variations in the binomial parameter estimate

$$\hat{p} = y/N,$$

and its unsymmetrical confidence limits, possibly as ordered by an indexing parameter x.

Analyzing binomial proportions

From the main SimFIT menu choose [Statistics], [Analysis of proportions], then [Binomial proportions], and examine the default test file `binomial.tf3` which has the following format.

y	N	x
23	84	1
12	78	2
31	111	3
65	92	4
71	93	5

The columns in this data format must be as follows.

- Column 1: The number of successes $0 \le y \le N$
- Column 2: The number of Bernoulli trials $N > 0$
- Column 3: An optional indexing parameter x

Note that the indexing parameter x is not used for any calculations, it is only required in order to identify, label, and space the data for subsequent plotting. If this third column is missing, as in `binomial.tf2`, SimFIT simply appends a third column of successive integers $1, 2, \ldots, N$. Typically x could be sample identifiers, concentrations of chemical, time from start of treatment, etc.

The SimFIT analysis of proportions procedure accepts a matrix of such y, N data then calculates the binomial parameters and derived parameters such as the Odds

$$\text{Odds } = \hat{p}/(1-\hat{p}), \text{ where } 0 < \hat{p} < 1,$$

and log(Odds), along with standard errors and confidence limits. It also performs a chi-square contingency table test and a likelihood ratio test for common binomial parameters as in the next table.

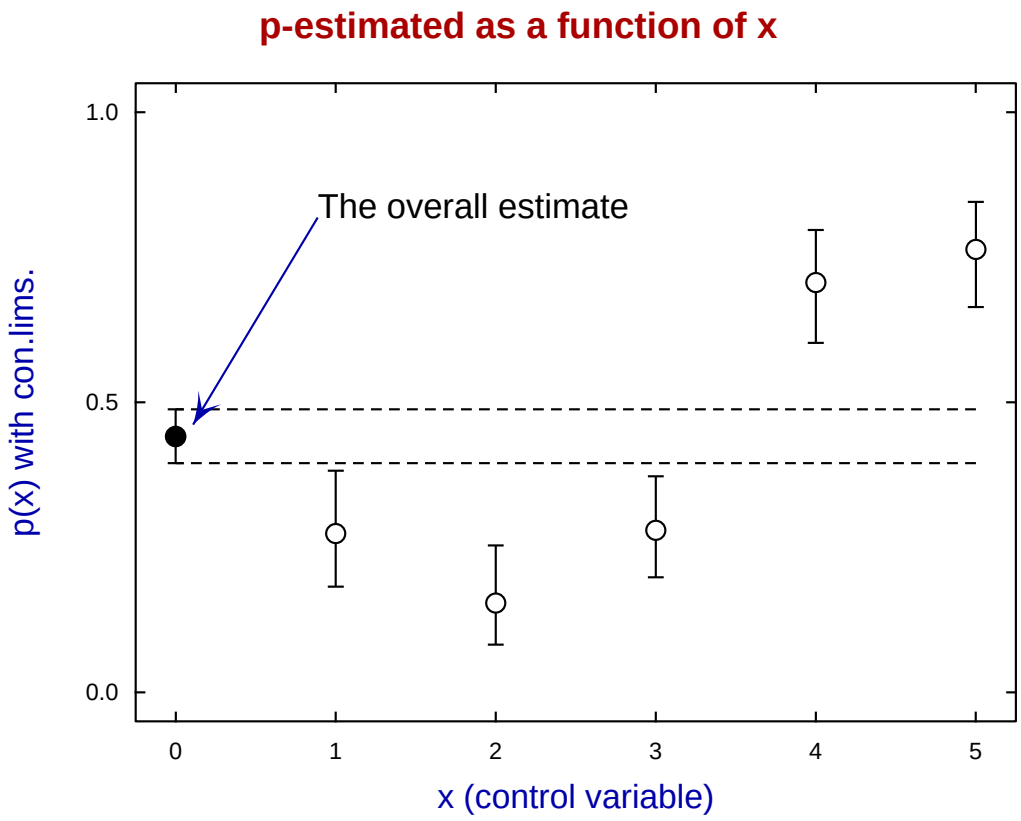
To test H_0 : equal binomial p-values for data in test file `binomial.tf3`		
Sample-size i.e. number of pairs	5	
Overall sum of y	202	
Overall sum of N	458	
Overall estimate of p	0.4410	
Lower 95% confidence limit	0.3950	
Upper 95% confidence limit	0.4879	
$-2\log\lambda$ $(-2LL)$	118.3	$NDOF = 4$
$P(\chi^2 \ge -2LL)$	0.0000	`Reject` H_0 `at 1% significance level`
Chi-square test statistic (C)	112.9	$NDOF = 4$
$P(\chi^2 \ge C)$	0.0000	`Reject` H_0 `at 1% significance level`

After choosing to analyze the parameter estimates, the next table with the data, p estimates, and exact 95% confidence limits is displayed.

y	N	lower-95%	$\hat{p}$	upper-95%
23	84	0.18214	0.27381	0.38201
12	78	0.08210	0.15385	0.25332
31	111	0.19829	0.27928	0.37241
65	92	0.60242	0.70652	0.79688
71	93	0.66404	0.76344	0.84542

Plotting binomial proportions

These results can then be plotted as individual sample estimates with 95% confidence limits, as in this graph where the overall estimate with overall confidence limits is also displayed.

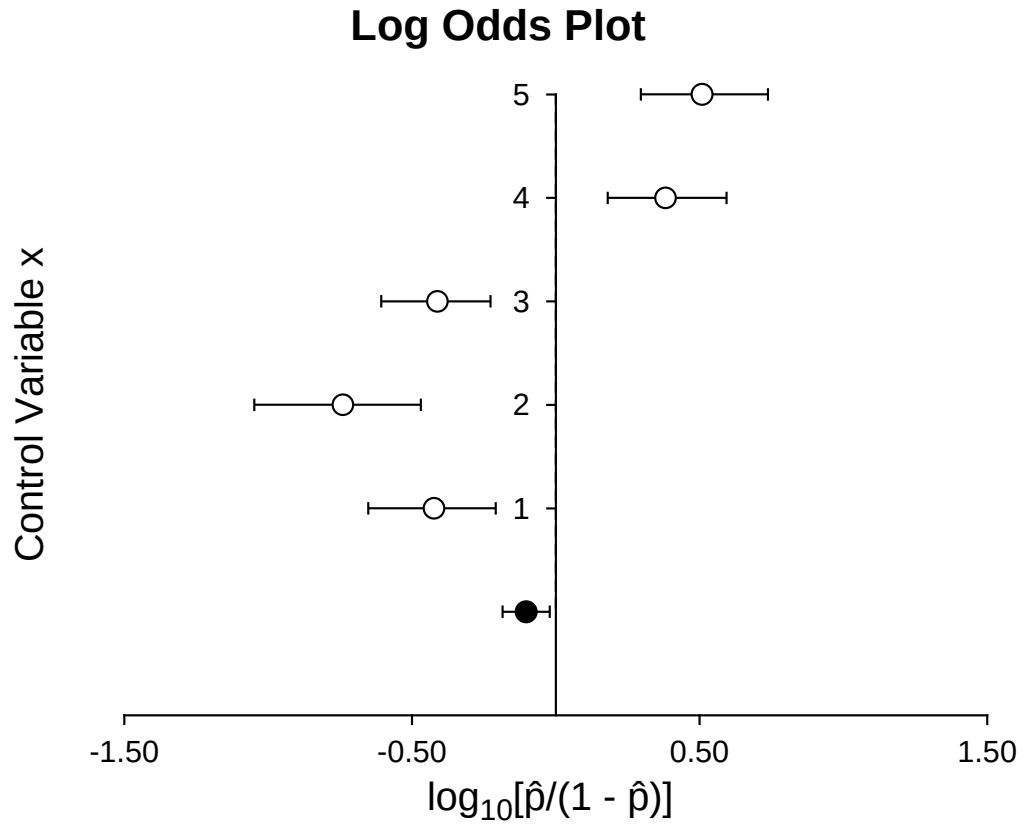


A useful rule of thumb is to analyze such plots for the position of confidence limits for the individual estimates with respect to the other individual confidence limits and the overall confidence limits shown as dotted lines. It is clear that samples 1, 2, and 3 lie below the overall confidence limits, while samples 4 and 5 lie above the overall confidence limits, confirming the previous conclusions from the χ^2 test.

Plotting log odds

It should be emphasized that everything that can be done with binomial proportions can be done by calculating and plotting parameter estimates $\hat{p}$ and confidence limits as just discussed. However, many experimentalists prefer to work with log odds in order to emphasize differences in order of magnitudes. For instance, the data can be plotted as log odds as in this next figure. However, to perform advanced graphics editing SIMFIT always transfers raw data not transformed data into the advanced editing, so it is necessary to transfer x, $y/(N-y)$ into

the advanced editing option first of all, followed by choosing a reverse y-semilog interactive transformation using logs to base ten. In addition, for finishing touches, the legends were edited and the y-axis moved to the central position indicated.



Binomial parameter confidence limits

It is obvious that a binomial parameter estimate $\hat{p} = y/N$ for the true population parameter p must satisfy

$$0 \leq \hat{p} \leq 1$$

and so the confidence limits should also be constrained to this range. Hence any accurate confidence limits cannot be symmetrical but must be skewed and so, when a binomial parameter is estimated, it is not possible to report the result in the usual way as $\hat{p} \pm \hat{s}$, or as $\hat{p}(\hat{p} - \hat{s}, \hat{p} + \hat{s})$, where $\hat{s}$ is estimated from the sample and percentiles of a standard normal distribution. Nevertheless, many users of computer packages do not understand this and prefer an approximate expression using the normal distribution because, as long as the sample is large and $p \approx 0.5$, a binomial distribution can be approximated by a normal distribution. For that reason a large sample 95% approximate central confidence range for the true population parameter p is often constructed using

$$\tilde{p} - \tilde{s} \leq p \leq \tilde{p} + \tilde{s}, \text{ where } \tilde{s} = Z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{N}}$$

with $\tilde{N} = N + 4$, and $\tilde{p} = (y + 2)/\tilde{N}$.

It is clear that for large samples with $y \approx N/2$ the normal approximation will be adequate but, in order to check the closeness of the approximate limits to the exact ones in any given case, SIMFIT provides tables to check the values. For instance, analysis of the test file `binomial.tf4` yields the following comparison.

$\hat{p} = (y/N)$ with exact unsymmetrical small sample limits

y	N	Lower-95%	$\hat{p}$	Upper-95%
23	84	0.182144	0.273810	0.382008
12	78	0.082102	0.153846	0.253321
31	111	0.198289	0.279279	0.372414
91	92	0.940922	0.989130	0.999725
1	93	0.000272	0.010753	0.058458

$\tilde{p} = (y + 2)/(N + 4)$ with approximate central limits $[\tilde{p} \pm \tilde{s}]$

y	N	Lower-95%	$\tilde{p}$	Upper-95%	$\tilde{s}$
23	84	0.189866	0.284091	0.378316	0.094225
12	78	0.089290	0.170732	0.252173	0.081442
31	111	0.204283	0.286957	0.369630	0.082673
91	92	0.933945	0.968750	1.003555	0.034805 ***
1	93	-0.003524	0.030928	0.065380	0.034452 ***

*** Indicates parameter limits outside range (0,1)

The column $\hat{s}$ indicates the amount $\hat{s}$ added to and subtracted from $\hat{p}$ to derive the limits so that the results can be reported as $\hat{p} \pm \hat{s}$. It will be seen that modifying the data in test file `binomial.tf3` to make test file `binomial.tf4` by editing in a couple of extreme values causes the approximate method to overflow or underflow as indicated by ***. Actually the numerical calculation to estimate the exact confidence takes much longer than estimation of the normal approximation, so SIMFIT allows users to choose the method to use when analyzing large samples.

Differences between probability estimates

For cases where the number of samples is relatively small, it is also sometimes helpful to examine tables that highlight significant differences between estimates as follows, using the test file `binomial.tf4`.

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

i	j	Lower-95%	$d(i, j)$	Upper-95%	Result	$Var(d(i, j))$	NNT	(95%c.l.)
1	2	-0.00455	0.11996	0.24448	Not significant	0.00404	9	(...)
1	3	-0.13219	-0.00547	0.12125	Not significant	0.00418	183	(NNH)
1	4	-0.81300	-0.71532	-0.61764	p(1) < p(4)	0.00248	2	(NNH)
1	5	0.16542	0.26306	0.36069	p(1) > p(5)	0.00248	4	(3,6)
2	3	-0.24109	-0.12543	-0.00977	p(2) < p(3)	0.00348	8	(NNH)
2	4	-0.91811	-0.83528	-0.75246	p(2) < p(4)	0.00179	2	(NNH)
2	5	0.06033	0.14309	0.22586	p(2) > p(5)	0.00178	7	(4,17)
3	4	-0.79596	-0.70985	-0.62374	p(3) < p(4)	0.00193	2	(NNH)
3	5	0.18247	0.26853	0.35458	p(3) > p(5)	0.00193	4	(3,5)
4	5	0.94857	0.97838	1.00818	p(4) > p(5)	0.00023	2	(1,2)

Where $d(i, j) < 0$, number needed to harm (NNH) is shown instead of NNT confidence limits.

Note that when the lower limit is negative and the upper limit is positive the confidence range includes zero so that the difference between estimates is not significantly different from zero. When the parameters are listed as different, the result can be interpreted as stricter (since $\alpha/2$ is used) than a one–sided lower tail or upper tail test (where α would normally be used). A purist would argue that, as three tests are being done on the same data, the Bonferroni principle would require that significance levels should be divided by three anyway.

It should be noted that the number needed to treat (NNT) is simply the reciprocal of the absolute difference $d(i, j) = p_i - p_j$, except that, to avoid overflow, this is constrained to the range $1 \leq NNT \leq 10^6$. Where confidence limits for NNT cannot be estimated, this is indicated by (...), and when the probability difference is negative the number needed to harm is indicated by (NNH) instead of the confidence range, as will be seen in the above table.

Confidence limits for analysis of two proportions

Given two proportions p_i and p_j estimated as

$$\hat{p}_i = y_i/N_i$$
$$\hat{p}_j = y_j/N_j$$

it is often wished to estimate confidence limits for the relative risk RR_{ij}, the difference between proportions DP_{ij}, and the odds ratio OR_{ij}, defined as

$$RR_{ij} = \hat{p}_i/\hat{p}_j$$
$$DP_{ij} = \hat{p}_i - \hat{p}_j$$
$$OR_{ij} = \frac{\hat{p}_i/(1-\hat{p}_i)}{\hat{p}_j/(1-\hat{p}_j)}.$$

First of all note that, for small proportions, the odds ratios and relative risks are similar in magnitude. Further, unlike the case of single proportions, exact confidence limits for these derived parameters can not be calculated. However, approximate central $100(1-\alpha)\%$ confidence limits can be obtained using the large sample normal approximations

$$\log(RR_{ij}) \pm Z_{\alpha/2}\sqrt{\frac{1-\hat{p}_i}{N_i\hat{p}_i} + \frac{1-\hat{p}_j}{N_j\hat{p}_j}}$$
$$DP_{ij} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{N_i} + \frac{\hat{p}_j(1-\hat{p}_j)}{N_j}}$$
$$\log(OR_{ij}) \pm Z_{\alpha/2}\sqrt{\frac{1}{y_i} + \frac{1}{N_i-y_i} + \frac{1}{y_j} + \frac{1}{N_j-y_j}}$$

provided $\hat{p}_i$ and $\hat{p}_j$ are not too close to 0 or 1. Here $Z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point, i.e., the lower $100(1-\alpha/2)$ percentage point for the standard normal distribution, and confidence limits for RR_{ij} and OR_{ij} can be obtained using the exponential function.

If the confidence regions estimated by this procedure include zero the significance is reported in the table of differences as not significant. Otherwise only the relative magnitudes of the pair in question are indicated.

When the difference between two probabilities is positive, a very approximate estimate for the confidence limits for NNT can be obtained using the values for DP_{ij}.

As elsewhere in SimFiT the significance level can be set by the user, and either natural or base ten logarithms can be plotted.