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

A contingency table is an array of nonnegative frequencies with n rows and m columns, such as this table contained in SimFIT test file `chisqd.tf4`, for 15 observations carried out on two populations to test for equal probabilities of success.

	Success	Failure	
Sample 1	3	3	6
Sample 2	7	2	9
	10	5	15

Here, the cell frequencies are $(3, 3, 7, 2)$, the sum of row frequencies known as row marginals are $(6, 9)$, the sum of column frequencies known as column marginals are $(10, 5)$, and obviously the row and column marginals must separately both add up to the total number of frequencies (15).

To be precise, in the general case there will be frequencies f_{ij} where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$, and it is wished to test for homogeneity, i.e. independence, or no association between the variables, which can be stated as the null hypothesis

$$H_0 : \mu_{ij} = \mu_{i.}\mu_{.j}, \text{ for } i = 1, 2, \ldots, n, \text{ and } j = 1, 2, \ldots, m$$

where each cell probability μ_{ij} is completely determined by the corresponding row marginal $\mu_{i.}$, and the column marginal $\mu_{.j}$. To examine a given data set SimFIT provides the following three alternatives.

1. **The chi-square test.**
   This is the easiest to perform and and interpret, and is the test most generally used. However, it must be emphasized that the test statistic is only asymptotically distributed as chi-square with $(n-1)(m-1)$ degrees of freedom in the limit for large samples. Where there are small frequencies the option to combine cells should be considered, and note that the Yate's continuity correction may be used where appropriate.

2. **The Fisher exact test.**
   This is very powerful and widely used, but sometimes suffers from being difficult to interpret with large samples, which also may lead to computational problems.

3. **The loglinear contingency table analysis.**
   This uses general linear modeling assuming a Poisson error distribution and log link, but it does require some expertise on the part of users.

Choose `[A/Z]` from the main SimFIT menu, then open SimFIT program **chisqd**.

1 Chi-square test

For all tables, SimFIT calculates a chi-square test statistic C from the observed frequencies f_{ij}, and expected frequencies e_{ij}, and also a likelihood ratio test statistic L defined in terms of the expected values e_{ij} and

marginals $f_{i.}$ and $f_{.j}$ as follows

$$
\begin{aligned}
e_{ij} &= f_{i.} f_{.j} / N \\
C &= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \\
L &= -2 \log \lambda \\
&= 2 \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \log(f_{ij}/e_{ij})
\end{aligned}
$$

It is often recommended to combine cells where the expected values are small, say $e_{ij} < 0.5$, and this facility is provided.

Select chi-square contingency table analysis, then analyze the above data which leads to calculation of the approximate chi-square test statistic with the Yate's continuity correction

$$C = \frac{N(|f_{11} f_{22} - f_{12} f_{21}| - N/2)^2}{r_1 r_2 c_1 c_2}$$

for this 2 by 2 contingency table, where N is the sum of frequencies f_{ij}, r_i are the row marginals, and c_j are the column marginals, leading to the following results, which do not suggest rejecting H_0.

Number of rows	2
Number of columns	2
chi-square test statistic C	0.3125
Number of degrees of freedom	1
$P(\chi^2 \geq C)$	0.5762
Upper tail 5% point	3.841
Upper tail 1% point	6.635
$L = -2\log(\lambda)$	1.243
$P(\chi^2 \geq L)$	0.2649

2 The Fisher exact test

For 2 by 2 contingency tables, and $N \leq 100$, tables like the following are also displayed.

Observed		Rearranged so r_1 = smallest marginal, $c_2 \geq c_1$	
3	3	3	2
7	2	3	7

$p(r)$	p for $f_{11} = r$ after rearranging and adjusting	
$p(0)$	0.041958	
$p(1)$	0.251748	
$p(2)$	0.419580	
$p(3)$	0.239760	$p(*)$, observed frequencies
$p(4)$	0.044955	
$p(5)$	0.001998	

P_sums, 1-tail and 2-tail test statistics		
P_sum1	0.041958	sum of $p(r) \leq p(*)$ for $r < 3$
P_sum2	0.953047	sum of all $p(r)$ for $r \leq 3$
P_sum3	0.286713	sum of all $p(r)$ for $r \geq 3$
P_sum4	0.046953	sum of $p(r) \leq p(*)$ for $r > 3$
P_sum5	1.000000	P_sum2 + P_sum4
P_sum6	0.328671	P_sum1 + P_sum3

For convenience, this test starts by rearranging the data table until r_1 is the smallest marginal and $c_2 \geq c_1$. Then all hypothetical tables that are possible with the same marginals are considered, but now for $r = f_{11}$ for $r = 0, 1, \ldots, r_1$ as follows, where the observed frequencies are indicated by stars (*).

0	5	1	4	2	3	*3	*2	4	1	5	0
6	4	5	5	4	6	*3	*7	2	8	1	9

Assuming the null hypothesis, the probabilities $p(r)$ for tables with $f_{11} = r$ are then calculated for a hypergeometric distribution using

$$p(r) = \frac{r_1!r_2!c_1!c_2!}{f_{11}!f_{21}!f_{12}!f_{22}!N!}$$

With the tables under consideration it is clear that, had the outcome been as for the hypothetical tables indicated by $p(0)$, $p(4)$, or $p(5)$ then the possibility of rejecting H_0 would have to be considered. However, the current data $p(3)$, indicated by $p(*)$ would be accepted, as for the chi-square test on the same data. With less obvious results, various one-tailed and two-tailed tests can be based on considering probabilities for more extreme contingency tables, or sums of such probabilities. As an example consider the following data

	Boys	Girls	
Left-handed	6 (18%)	12 (22%)	18
Right-handed	28 (82%)	24 (67%)	52
	34	36	70

and possible hypotheses for this sample
H_0: left-handedness is not less common in boys than girls
H_A: left-handedness is less common in boys than girls.

$p(r)$	p for $f_{11} = r$ after rearranging and adjusting	
$p(0)$	0.000000	
$p(1)$	0.000013	
$p(2)$	0.000177	
$p(3)$	0.001436	
$p(4)$	0.007590	
$p(5)$	0.027720	
$p(6)$	0.072572	$p(*)$, observed frequencies
$p(7)$	0.139338	
$p(8)$	0.198959	
$p(9)$	0.212877	
$p(10)$	0.171062	
$p(11)$	0.102959	
$p(12)$	0.046046	
$p(13)$	0.015082	
$p(14)$	0.003535	
$p(15)$	0.000571	
$p(16)$	0.000060	
$p(17)$	0.000004	
$p(18)$	0.000000	
P_Sums, for 1-tail and 2-tail test statistics		
P_sum1	0.036936	sum of $p(r) \leq p(*)$ for $r < 6$
P_sum2	0.109508	sum of all $p(r)$ for $r \leq 6$ (one-tailed p)
P_sum3	0.963064	sum of all $p(r)$ for $r \geq 6$
P_sum4	0.065297	sum of $p(r) \leq p(*)$ for $r > 6$
P_sum5	0.174805	P_sum2 + P_sum4
P_sum6	1.000000	P_sum1 + P_sum3

Adding up the probabilities for the observed table $p(6) = p(*)$ and all the possible tables more extreme than this that would favor H_A against H_0 we see that the appropriate one-tailed p value is

$$p(0) + p(1) + p(2) + p(3) + p(4) + p(5) + p(6) = 0.109508$$

and so, for this sample with $\alpha = 0.05$ we would not consider rejecting H_0.

3 The loglinear contingency table analysis

The full details for this test will be found in the SimFIT reference manual, but meaningful interpretation of the results is possible without detailed understanding. Essentially, a statistical model is constructed for the contingency table with the following characteristics.

- Best-fit theoretical cell frequencies are calculated using a loglinear model.
- The parameter estimates are displayed along with standard errors and p values.
- Predicted cell frequencies are then compared with the observed data to generate differences, residuals, and leverages.
- The deviance is calculated, and the chi-square significance reported.

Here are the results for the SimFIT test data set.

Log-linear contingency table analysis
Data: Test file `chisqd.tf4`
number of rows = 2, number of columns = 2
Deviance (D) = 1.243, degrees of freedom = 1
$P(\chi^2 \geq D) = 0.2649$

Parameter	Estimate	Std.Err.	Lower 95%	Upper 95%	p
Constant	1.792	0.380	-3.04	6.62	0.1330 ***
Row 1	-0.4055	0.527	-7.10	6.29	0.5823 ***
Row 2	0.4055	0.527	-6.29	7.10	0.5823 ***
Col 1	-0.6931	0.547	-7.65	6.26	0.4254 ***
Col 2	0.6931	0.547	-6.26	7.65	0.4254 ***

Data	Model	Delta	Residual	Leverage
3	4	-1	-0.5234	0.7997
3	2	1	0.6579	0.6005
7	6	1	0.3976	0.8664
2	3	-1	-0.6149	0.7335

The model that is assumed expresses the theoretical cell probability μ_{ij} as a constant θ, plus row parameters α_i, column parameters β_j, and mixed row-column parameters γ_{ij} in the following way

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij}$$

where

$$\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 0.$$

The null hypothesis of homogeneity, that is $\mu_{ij} = \mu_{i.}\mu_{.j}$, can then be stated as

$$H_0 : \ \gamma_{ij} = 0 \text{ for } i = 1, 2, \ldots, n, \text{ and } j = 1, 2, \ldots, m$$

and the deviance measures the extent to which the hypothesis of homogeneity can be supported. Note that the purpose of starred parameter estimates is simply to warn users about suspiciously large ratios of standard errors to parameter estimates, i.e. where $p \geq 0.05$. Also, with large contingency tables, the ability to plot the residuals in a variety of ways to visualize goodness of fit is provided.

As before, this test provides no support for rejecting the null hypothesis of homogeneity with these data.