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

Given any data sample in the form of a rectangular table of values with no missing values it is useful to generate a summary of the all the parameters that can be estimated together with the ability to plot the data in alternative ways.

For instance, choose [Statistics] from the main SIMFIT menu then [Data exploration]. Open the [Exhaustive analysis of an arbitrary matrix] option and examine the data set contained in test file `cluster.tf1` which is the following 12 by 8 matrix.

1.0	4.0	2.0	11.0	6.0	4.0	3.0	9.0
8.0	5.0	1.0	14.0	19.0	7.0	13.0	21.0
3.0	1.0	3.0	1.0	3.0	6.0	23.0	37.0
9.0	0.0	7.0	7.0	1.0	2.0	21.0	2.0
7.0	12.0	9.0	5.0	14.0	9.0	12.0	14.0
2.0	13.0	15.0	2.0	23.0	6.0	34.0	8.0
11.0	7.0	2.0	1.0	4.0	17.0	11.0	4.0
6.0	3.0	7.0	12.0	11.0	8.0	8.0	0.0
8.0	21.0	1.0	10.0	31.0	9.0	3.0	18.0
19.0	14.0	12.0	9.0	16.0	10.0	0.0	27.0
17.0	18.0	10.0	6.0	19.0	14.0	1.0	24.0
15.0	21.0	8.0	7.0	17.0	12.0	4.0	22.0

The possibilities for further analysis are now listed.

- Summarize all columns (or rows)
- Exhaustive analysis of any column (or row)
- Analyze/paired-test any two rows (or columns)
- Plot
  2D barchart or stack plot with rows as groups
  2D box and whisker plot or bars and error bars
  2D scattergrams with symbols (and lines if requested)
  3D barchart or cylinder plot
- Display/file Sum–of–Squares, covariance, or correlation matrix

Summarizing all rows or columns

For instance, the option to summarize all columns results in this analysis.

Column	Mean	Variance	St.Dev.	Coeff.Var.
1	8.83333	33.4242	5.78137	65.45%
2	9.91667	57.7197	7.59735	76.61%
3	6.41667	21.5379	4.64089	72.33%
4	7.08333	18.6288	4.31611	60.93%
5	13.6667	81.3333	9.01850	65.99%
6	8.66667	17.6970	42.0678	48.54%
7	11.0833	107.720	10.3788	93.64%
8	15.5000	127.364	11.2855	72.81%

Here, for each column, the summary statistics are calculated downwards for all rows, and a similar table can be generated for rows calculated across all columns.

Pairwise statistical tests between rows or columns

Next consider pairwise statistical tests between selected rows or columns. If you choose to apply more than one statistical test this would be to use the absolutely forbidden technique of multiple tests on the same data.

In such a situation you can either use the Bonferroni method or similar with a factor related to the actual number of tests applied as explained in the SimFIT reference manual, or just use commonsense and regard this as a preliminary examination where the p values are simply being regarded as indicators of the differences between paired or columns and not being used for hypothesis tests.

Analysis and two-tail tests for:	
$N = 12$, $X =$ column 1, $Y =$ column 2	
Unpaired t test:	
t	-0.39309
p	0.69804
Paired t test:	
t	-0.56175
p	0.58555
Kolmogorov-Smirnov 2-sample test:	
d	0.25000
z	0.10206
p	0.53610
Mann-Whitney U test:	
u	68.5000
z	-0.17339
p	0.85377
Wilcoxon signed rank test:	
w	33.5000
z	-0.39299
p	0.70752
Run test:	
+	6 (no. $x > y$)
-	6 (no. $x < y$)
p	0.60823
Sign test: N for non-tied pairs	
N	12
-	6 (no. $x < y$)
p	1.00000

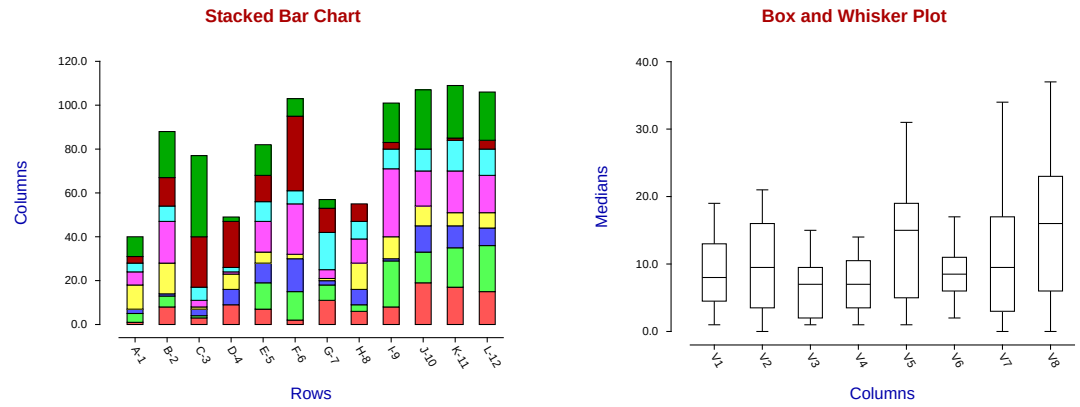
After an analysis like the above has been carried out, less controversial results are calculated, that is, the inner product of the two selected rows or columns regarded as vectors, leading to the angle between them and Euclidean distance between them, i.e. the square root of the sum of squared differences.

n	dot product	x size	y size	distance	$\cos(\theta)$	radians	degrees
12	1307.0	36.1109	42.6028	22.4722	0.849569	0.5556	31.835

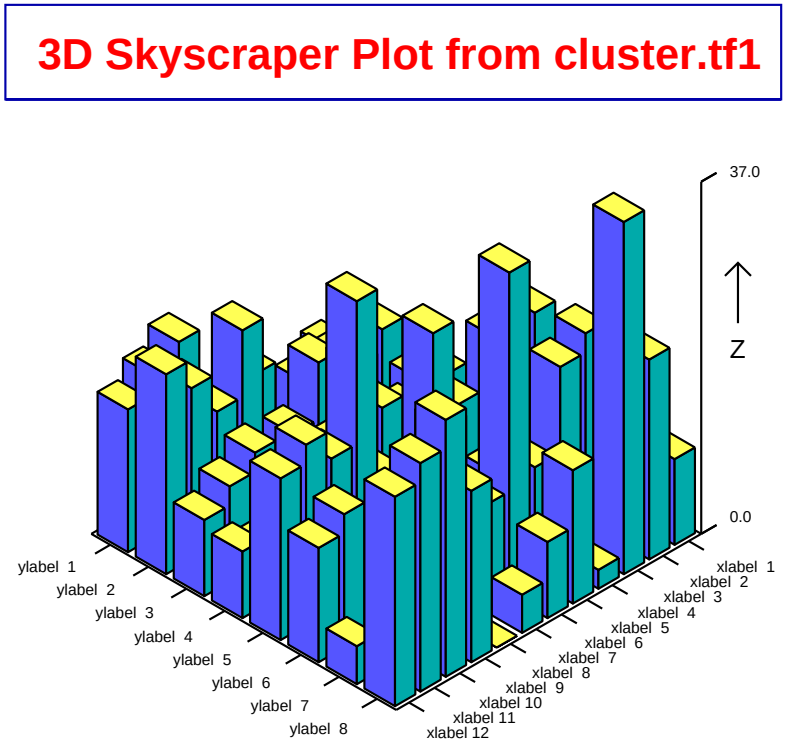
Plotting a matrix

As long as the number of rows and columns is fairly small, say < 20, and for some procedures the matrix contains only positive values, several graphs can be drawn to visualize the relative magnitude of column values across row.

For instance, the left-hand figure below plots a bar for each row with a stacked bar of segments each proportional to the column values for the corresponding rows. On the right is a box and whisker plot to illustrate the quartiles for each column calculated for all rows. Of course it is easy to interpret the row and column effects illustrated when it is realized that the data set has 12 rows and 8 columns.



The next plot illustrates a 3D skyscraper plot where, for each value in the data matrix, say x_{ij}, the vertical height of the bars is proportional to the x_{ij} values.



Numerous other graphs are available where the sign of x_{ij} is irrelevant, for instance clusters and 95% confidence ellipses for the data means or for the overall data ranges, and also linear regression according to all three conventions is possible for selected pairs of rows and/or columns.

Lower triangles of the covariance and correlation matrices

The exhaustive analysis of an arbitrary matrix can also calculate several symmetrical matrices from the data. For instance the sum of squares matrix or the covariance matrix as this will give some idea if the columns are independent.

Variance-Covariance matrix

33.4242							
23.2576	57.7197						
7.71212	11.5833	21.5379					
1.65152	-0.71970	-5.67424	18.6288				
10.1212	54.3333	9.06061	10.8485	81.3333			
15.2121	17.0606	0.51515	-5.15152	7.69697	17.6970		
-35.2576	-33.3561	11.1439	-23.4621	-18.2424	-19.7879	107.720	
19.6364	25.7727	-1.59091	-5.68182	21.8182	6.45455	-19.8636	127.364

An easier matrix to visualize for correlations in the data is the correlation matrix, which is sometimes given, as below, with unit diagonals to avoid confusion.

Pearson product-moment correlations

1							
0.529507	1						
0.287436	0.328526	1					
0.066185	-0.021948	-0.283279	1				
0.194119	0.792994	0.216482	0.278704	1			
0.625474	0.533805	0.026387	-0.283722	0.202879	1		
-0.587590	-0.423024	0.231361	-0.523754	-0.194895	-0.453213	1	
0.300959	0.300591	-0.030375	-0.116647	0.214369	0.135954	-0.169585	1

For a n by m matrix $\boldsymbol{X}$ with values x_{ij}, the sample column means $\bar{x}_j$, vector of column means $\bar{\boldsymbol{x}}$, variance of the j'th variable s_{jj}, covariance between the j and k'th variable s_{jk}, correlation between the j and k'th variable c_{jk}, and covariance matrix $\boldsymbol{S}$ are defined as follows.

$$\begin{aligned}
\bar{x}_j &= \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\
\bar{\boldsymbol{x}}^T &= (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m) \\
s_{jj} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \\
s_{jk} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \\
c_{jk} &= \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \\
\boldsymbol{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T
\end{aligned}$$

Alternatively, if $\hat{\boldsymbol{X}}$ is the matrix centered by subtracting the sample column means and $\tilde{\boldsymbol{X}}$ is the centered matrix scaled by dividing by the column standard deviations, then the sample covariance $\boldsymbol{S}$ and correlation matrices $\boldsymbol{C}$ are

$$\boldsymbol{S} = \frac{1}{n-1}\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} \text{ and } \boldsymbol{C} = \frac{1}{n-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}.$$