Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.

https://simfit.org.uk
https://simfit.silverfrost.com

Given a sample from a known distribution it is generally easy to estimate the population parameters using the sample estimates, but it is not always so easy to determine the confidence limits, such as a 95% confidence interval. From the main SIMFIT menu you can select [Statistics] then the option to perform statistical calculations. Here you can choose the distribution required and the significance level of interest, then input the estimates and sample sizes required. Note that the well-known case of a normal distribution leads many to believe that a confidence interval is always symmetrical about a parameter estimate, but many confidence intervals will be asymmetric for those distributions (Poisson, binomial) where exact methods are used, not calculations based on the normal approximation.

Confidence limits for a Poisson parameter

Given a sample $x_1, x_2, \ldots, x_n$ of n non-negative integers from a Poisson distribution with parameter λ, the parameter estimate $\hat{\lambda}$, i.e., the sample mean, and confidence limits λ_1, λ_2 are calculated as follows

$$K = \sum_{i=1}^{n} x_i,$$

$$\hat{\lambda} = K/n,$$

$$\lambda_1 = \frac{1}{2n}\chi^2_{2K,\alpha/2},$$

$$\lambda_2 = \frac{1}{2n}\chi^2_{2K+2,1-\alpha/2},$$

$$\text{so that } \exp(-n\lambda_1)\sum_{x=K}^{\infty}\frac{(n\lambda_1)^x}{x!} = \frac{\alpha}{2},$$

$$\exp(-n\lambda_2)\sum_{x=0}^{K}\frac{(n\lambda_2)^x}{x!} = \frac{\alpha}{2},$$

$$\text{and } P(\lambda_1 \le \lambda \le \lambda_2) = 1 - \alpha,$$

using the lower tail critical points of the chi-square distribution. The following very approximate rule-of-thumb can be used to get a quick idea of the range of a Poisson mean λ given a single count x and exploiting the fact that the Poisson variance equals the mean

$$P(x - 2\sqrt{x} \le \lambda \le x + 2\sqrt{x}) \approx 0.95.$$

Example

The number of weed seeds in 98 samples of meadow grass yielded these counts with a mean of 3.0204.

Number	0	1	2	3	4	5	6	7	8	9	10
Frequency	3	17	26	16	18	9	3	5	0	1	0

The 95% and 99% noncentral confidence intervals from the estimate were found to be as follows.

Sample size	Mean	Level	Interval
98	3.0204	95%	$2.68608 \le \lambda \le 3.38483$
98	3.0204	99%	$2.58737 \le \lambda \le 3.50272$

Confidence limits for a binomial parameter

For k successes in n trials, the binomial parameter estimate $\hat{p}$ is k/n and three methods are used to calculate confidence limits p_1 and p_2 so that

$$\sum_{x=k}^{n} \binom{n}{x} p_1^x (1-p_1)^{n-x} = \alpha/2,$$
$$\text{and } \sum_{x=0}^{k} \binom{n}{x} p_2^x (1-p_2)^{n-x} = \alpha/2.$$

- If $\max(k, n-k) < 10^6$, the lower tail probabilities of the beta distribution are used as follows

$$p_1 = \beta_{k,n-k+1,\alpha/2},$$
$$\text{and } p_2 = \beta_{k+1,n-k,1-\alpha/2}.$$

- If $\max(k, n-k) \geq 10^6$ and $\min(k, n-k) \leq 1000$, the Poisson approximation with $\lambda = np$ and the chi-square distribution are used, leading to

$$p_1 = \frac{1}{2n}\chi^2_{2k,\alpha/2},$$
$$\text{and } p_2 = \frac{1}{2n}\chi^2_{2k+2,1-\alpha/2}.$$

- If $\max(k, n-k) > 10^6$ and $\min(k, n-k) > 1000$, the normal approximation with mean np and variance $np(1-p)$ is used, along with the lower tail normal deviates $Z_{1-\alpha/2}$ and $Z_{\alpha/2}$, to obtain approximate confidence limits by solving

$$\frac{k - np_1}{\sqrt{np_1(1-p_1)}} = Z_{1-\alpha/2},$$
$$\text{and } \frac{k - np_2}{\sqrt{np_2(1-p_2)}} = Z_{\alpha/2}.$$

The following very approximate rule-of-thumb can be used to get a quick idea of the range of a binomial mean np given x and exploiting the fact that the binomial variance variance equals $np(1-p)$

$$P(x - 2\sqrt{x} \leq np \leq x + 2\sqrt{x}) \approx 0.95.$$

Example

In a study the number of deaths among pensioners in a six year period were as follows.

	Sample size	Deaths	Probability	95% confidence interval
Non-smokers	1067	117	0.109653	$0.091533 \leq p \leq 0.129957$
Smokers	402	54	0.134328	$0.102548 \leq p \leq 0.171609$

Again, note the noncentral 95% confidence intervals for the probability estimates $\hat{p}$ as summarized below.

Deaths/Subjects	$\hat{p}$	95% Confidence Interval	Group
117/1067	0.1097	0.1097 - 0.0182, 0.1097 + 0.0203	Non-smokers
54/402	0.1343	0.1343 - 0.0318, 0.1343 + 0.0373	Smokers

Confidence limits for a normal mean and variance

If the sample mean is $\bar{x}$, and the sample variance is s^2, with a sample of size n from a normal distribution having mean μ and variance σ^2, the confidence limits are defined by

$$P(\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}) = 1 - \alpha,$$
$$\text{and } P((n-1)s^2/\chi^2_{\alpha/2,n-1} \le \sigma^2 \le (n-1)s^2/\chi_{1-\alpha/2,n-1}) = 1 - \alpha$$

where the upper tail probabilities of the t and chi-square distribution are used.

Example

The body temperature of 25 intertidal crabs was recorded in ℃ as follows: 24.3, 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4. The sample mean, variance and standard deviation were $\bar{x} = 25.03$, $s^2 = 1.8$, and $s = 1.3416408$ leading to the following central confidence intervals for the mean and unsymmetrical confidence limits for the variance.

Sample size	Level	Parameter	Estimate	Interval
25	95%	Mean	25.03	$24.4762 \le \mu \le 25.5838$
25	99%	Mean	25.03	$24.2795 \le \mu \le 25.7805$
25	95%	Variance	1.8	$1.09745 \le \sigma^2 \le 3.48355$
25	99%	Variance	1.8	$0.948231 \le \sigma^2 \le 4.36971$

Confidence limits for a correlation coefficient

If a Pearson product-moment correlation coefficient r is calculated from two samples of size n that are jointly distributed as a bivariate normal distribution, the confidence limits for the population parameter ρ are given by

$$P\left(\frac{r - r_c}{1 - rr_c} \le \rho \le \frac{r + r_c}{1 + rr_c}\right) = 1 - \alpha,$$
$$\text{where } r_c = \sqrt{\frac{t^2_{\alpha/2,n-2}}{t^2_{\alpha/2,n-2} + n - 2}}.$$

Example

The wing and tail lengths in cm for 12 birds were as in this next table.

Wing	10.4	10.8	11.1	10.2	10.3	10.2	10.7	10.5	10.8	11.2	10.6	11.4
Tail	7.4	7.6	7.9	7.2	7.4	7.1	7.4	7.2	7.8	7.7	7.8	8.3

This gives a correlation coefficient of $r = 0.87$ with a sample size of $n = 12$, leading to the nonsymmetrical 95% confidence interval.

$$0.589337 \le \rho \le 0.963279$$

Confidence limits for trinomial parameters

If, in a trinomial distribution, the probability of category i is p_i for $i = 1, 2, 3$, then the probability P of observing n_i in category i in a sample of size $N = n_1 + n_2 + n_3$ from a homogeneous population is given by

$$P = \frac{N!}{n_1!n_2!n_3!}p_1^{n_1}p_2^{n_2}p_3^{n_3}$$

and the maximum likelihood estimates, of which only two are independent, are

$$\begin{aligned}\hat{p_1} &= n_1/N,\\ \hat{p_2} &= n_2/N,\\ \text{and } \hat{p_3} &= 1-\hat{p_1}-\hat{p_2}.\end{aligned}$$

The bivariate estimator is approximately normally distributed, when N is large, so that

$$\begin{bmatrix}\hat{p_1}\\ \hat{p_2}\end{bmatrix} \sim MN_2\left(\begin{bmatrix}p_1\\ p_2\end{bmatrix}, \begin{bmatrix}p_1(1-p_1)/N & -p_1p_2/N\\ -p_1p_2/N & p_2(1-p_2)/N\end{bmatrix}\right)$$

where MN_2 signifies the bivariate normal distribution. Consequently

$$((\hat{p_1}-p_1),(\hat{p_2}-p_2))\begin{bmatrix}p_1(1-p_1)/N & -p_1p_2/N\\ -p_1p_2/N & p_2(1-p_2)/N\end{bmatrix}^{-1}\begin{pmatrix}\hat{p_1}-p_1\\ \hat{p_2}-p_2\end{pmatrix} \sim \chi^2_2$$

and hence, with probability 95%,

$$\frac{(\hat{p_1}-p_1)^2}{p_1(1-p_1)}+\frac{(\hat{p_2}-p_2)^2}{p_2(1-p_2)}+\frac{2(\hat{p_1}-p_1)(\hat{p_2}-p_2)}{(1-p_1)(1-p_2)} \le \frac{(1-p_1-p_2)}{N(1-p_1)(1-p_2)}\chi^2_{2;0.05}.$$

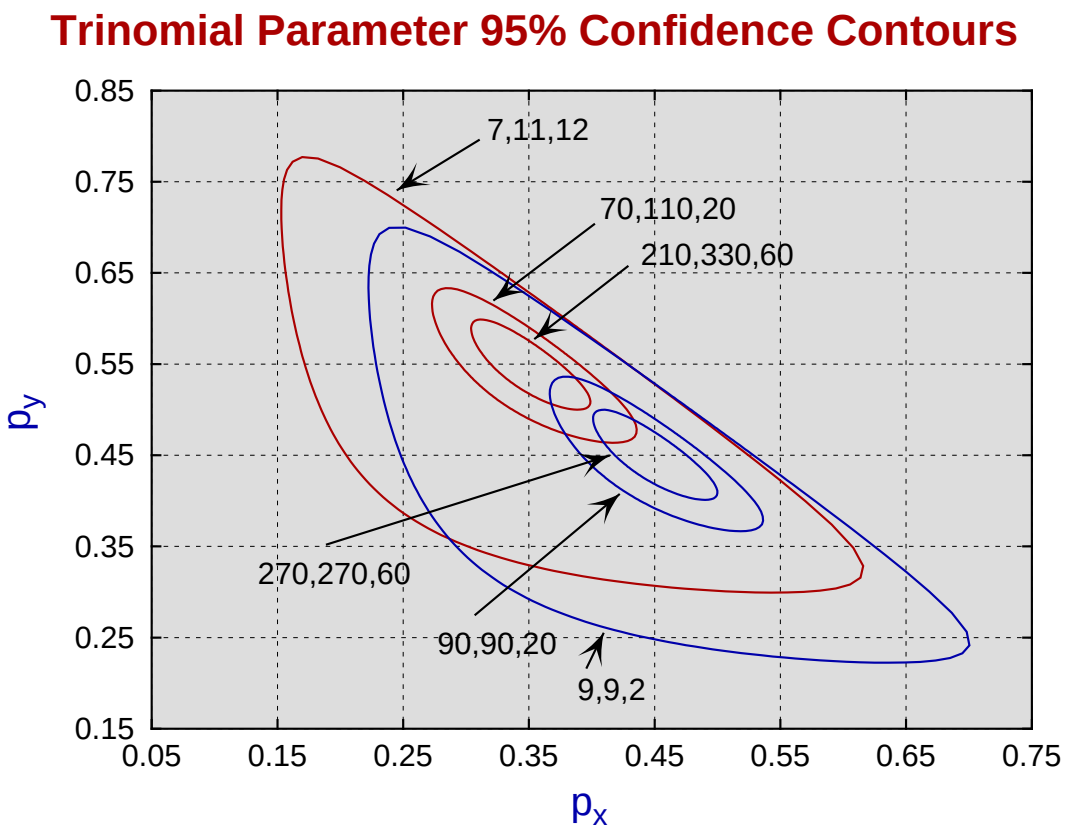
Such inequalities define regions in the (p_1, p_2) parameter space which can be examined for statistically significant differences between $p_{i(j)}$ in samples from populations subjected to treatment j. Where regions are clearly disjoint, parameters have been significantly affected by the treatments, as illustrated next.

Plotting trinomial parameter joint confidence regions

A useful rule of thumb to see if parameter estimates differ significantly is to check their approximate central 95% confidence regions. If the regions are disjoint it indicates that the parameters differ significantly and, in fact, parameters can differ significantly even with limited overlap. If two or more parameters are estimated, it is valuable to inspect the joint confidence regions defined by the estimated covariance matrix and appropriate chi-square critical value. Consider, for example, this figure generated by the contour plotting function of **binomial**. Data triples x, y, z can be any partitions, such as number of male, female or dead hatchlings from a batch of eggs where it is hoped to determine a shift from equi-probable sexes. The contours are defined by

$$((\hat{p_x}-p_x),(\hat{p_y}-p_y))\begin{bmatrix}p_x(1-p_x)/N & -p_xp_y/N\\ -p_xp_y/N & p_y(1-p_y)/N\end{bmatrix}^{-1}\begin{pmatrix}\hat{p_x}-p_x\\ \hat{p_y}-p_y\end{pmatrix} = \chi^2_{2:0.05}$$

where $N = x + y + z$, $\hat{p_x} = x/N$ and $\hat{p_y} = y/N$ as discussed in connection with the trinomial distribution. When $N = 20$ the triples 9,9,2 and 7,11,2 cannot be distinguished, but when $N = 200$ the orbits are becoming elliptical and converging to asymptotic values. By the time $N = 600$ the triples 210,330,60 and 270,270,60 can be seen to differ significantly.


Trinomial Parameter 95% Confidence Contours
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
p_y
p_x
7,11,12
70,110,20
210,330,60
270,270,60
90,90,20
9,9,2