Statistics menu

Univariate

This function computes a number of basic descriptive statistics for one or more samples of univariate data. Each sample must have at least 3 values, and occupies one column in the spreadsheet. The columns do not have to contain the same number of values. The example below uses two samples: the sizes in mm of the skulls of 30 female and 29 male gorillas. To run the analysis, the two columns (or the whole spreadsheet) must be selected.

![Example of Univariate Analysis](image)

The following numbers are shown for each sample:

- **N:** The number of values $n$ in the sample
- **Min:** The minimum value
- **Max:** The maximum value
- **Sum:** The sum
- **Mean:** The estimate of the mean, calculated as $\bar{x} = \frac{\sum x_i}{n}$
- **Std. error:** The standard error of the estimate of the mean, calculated as $SE_\bar{x} = \frac{s}{\sqrt{n}}$ where $s$ is the estimate of the standard deviation (see below).
- **Variance:** The sample variance, calculated as $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$.
Stand. dev.: The sample standard deviation, calculated as 
\[ s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} . \]

Median: The median of the sample. For \( n \) odd, the given value such that there are equally many values above and below. For \( n \) even, the average of the two central values.

25 prcntil: The 25th percentile, i.e. the given value such that 25% of the sample is below, 75% above. Linear interpolation is used when necessary.

75 prcntil: The 75th percentile, i.e. the given value such that 75% of the sample is below, 25% above. Linear interpolation is used when necessary.

Skewness: The sample skewness, zero for a normal distribution, positive for a tail to the right.

Calculated as 
\[ G_1 = \frac{n}{(n-1)(n-2)} \left( \frac{1}{n-1} \sum (x_i - \bar{x})^3 \right) . \]

Note there are several versions of this around – Past uses the same equation as SPSS and Excel. Slightly different results may occur using other programs, especially for small sample sizes.

Kurtosis: The sample kurtosis, zero for a normal distribution. Calculated as 
\[ G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \left( \frac{1}{n-1} \sum (x_i - \bar{x})^4 \right) - 3 \left( \frac{n-1}{n-2(n-3)} \right) . \]

Again Past uses the same equation as SPSS and Excel.

Geom. mean: The geometric mean, calculated as 
\[ \left( x_1 x_2 \cdots x_n \right)^{1/n} . \]

Missing data: Supported by deletion.
Similarity and distance indices

Computes a number of similarity or distance measures between all pairs of rows. The data can be univariate or (more commonly) multivariate, with variables in columns. The results are given as a symmetric similarity/distance matrix. This module is rarely used, because similarity/distance matrices are usually computed automatically from primary data in modules such as PCO, NMDS, cluster analysis and ANOSIM in Past.

Gower

A distance measure that averages the difference over all variables, each term normalized for the range of that variable:

\[
d_{jk} = \frac{1}{n} \sum_{i} \frac{|x_{ji} - x_{ki}|}{\max_{y} x_{y} - \min_{y} x_{y}}.
\]

The Gower measure is similar to Manhattan distance (see below) but with range normalization. When using mixed data types (see below), this is the default measure for continuous and ordinal data.

Euclidean

Basic Euclidean distance. In early versions of Past, this was normalized for the number of variables (the value is still adjusted for missing data).
\[ d_{jk} = \sqrt{\sum_i (x_{ji} - x_{ki})^2}. \]

**Correlation**

The complement 1 - r of Pearson’s r correlation across the variables:

\[ d_{jk} = 1 - \frac{\sum_i (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_i (x_{ji} - \bar{x}_j)^2} \sqrt{\sum_i (x_{ki} - \bar{x}_k)^2}}. \]

Taking the complement makes this a distance measure. See also the Correlation module, where Pearson’s r is given directly and with significance tests.

**Rho**

The complement 1 - r_s of Spearman’s rho, which is the correlation coefficient of ranks. See also the Correlation module, where rho is given directly and with significance tests.

**Dice**

Also known as the Sorensen coefficient. For binary (absence-presence) data, coded as 0 or 1 (any positive number is treated as 1). The Dice similarity puts more weight on joint occurrences than on mismatches.

When comparing two rows, a match is counted for all columns with presences in both rows. Using \( M \) for the number of matches and \( N \) for the total number of columns with presence in just one row, we have

\[ d_{jk} = \frac{2M}{2M + N}. \]

**Jaccard**

A similarity index for binary data. With the same notation as given for Dice similarity above, we have

\[ d_{jk} = \frac{M}{M + N}. \]

**Kulczynski**

A similarity index for binary data. With the same notation as given for Dice similarity above (with \( N_1 \) and \( N_2 \) referring to the two rows), we have
\[d_{jk} = \frac{M}{M + N_1} + \frac{M}{M + N_2} .\]

**Ochiai**

A similarity index for binary data, comparable to the cosine similarity for other data types:

\[d_{jk} = \sqrt{\frac{M}{M + N_1} \frac{M}{M + N_2}} .\]

**Simpson**

The Simpson index is defined simply as \(M / N_{\text{min}}\), where \(N_{\text{min}}\) is the smaller of the numbers of presences in the two rows. This index treats two rows as identical if one is a subset of the other, making it useful for fragmentary data.

**Bray-Curtis**

Bray-Curtis is becoming the standard distance index for abundance data in e.g. marine biology.

\[d_{jk} = \frac{\sum |x_{ji} \cdot x_{ki}|}{\sum (x_{ji} + x_{ki})} .\]

**Cosine**

The inner product of abundances each normalised to unit norm, i.e. the cosine of the angle between the vectors.

\[d_{jk} = \frac{\sum x_{ji} \cdot x_{ki}}{\sqrt{\sum x_{ji}^2} \cdot \sqrt{\sum x_{ki}^2}} .\]

**Morisita**

For abundance data.
\[
\lambda_1 = \frac{\sum_i x_{ji}(x_{ji} - 1)}{\sum_j x_{ji} \left( \sum_i x_{ji} - 1 \right)}
\]

\[
\lambda_2 = \frac{\sum_i x_{ki}(x_{ki} - 1)}{\sum_i x_{ki} \left( \sum_i x_{ki} - 1 \right)}
\]

\[
d_{jk} = \frac{2\sum_i x_{ji}x_{ki}}{(\lambda_1 + \lambda_2)\sum_i x_{ji}\sum_i x_{ki}}.
\]

**Raup-Crick**

Raup-Crick index for absence-presence data. This index (Raup & Crick 1979) uses a randomization (Monte Carlo) procedure, comparing the observed number of species occurring in both associations with the distribution of co-occurrences from 200 random replicates.

**Horn**

Horn’s overlap index for abundance data (Horn 1966).

\[
N_j = \sum_i x_{ji}
\]

\[
N_k = \sum_i x_{ki}
\]

\[
d_{jk} = \frac{\sum_i [(x_{ji} + x_{ki})\ln(x_{ji} + x_{ki})] - \sum_j x_{ji}\ln x_{ji} - \sum_k x_{ki}\ln x_{ki}}{(N_j + N_k)\ln(N_j + N_k) - N_j \ln N_j - N_k \ln N_k}.
\]

**Hamming**

Hamming distance for categorical data as coded with integers (or sequence data coded as CAGT). The Hamming distance is the number of differences (mismatches), so that the distance between (3,5,1,2) and (3,7,0,2) equals 2. In PAST, this is normalised to the range [0,1], which is known to geneticists as "p-distance".

**Chord**
Euclidean distance between normalized vectors. Commonly used for abundance data. Can be written as

\[ d_{jk} = \sqrt{2 - 2 \frac{\sum x_{ji} x_{ki}}{\sqrt{\sum x_{ji}^2 \sum x_{ki}^2}}}. \]

**Manhattan**

The sum of differences in each variable:

\[ d_{jk} = \sum_i |x_{ji} - x_{ki}|. \]

**Jukes-Cantor**

Distance measure for genetic sequence data (CAGT). Similar to \( p \) (or Hamming) distance, but takes into account probability of reversals:

\[ d = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} p \right) \]

**Kimura**

The Kimura 2-parameter distance measure for genetic sequence data (CAGT). Similar to Jukes-Cantor distance, but takes into account different probabilities of nucleotide transitions vs. transversions (Kimura 1980). With \( P \) the observed proportion of transitions and \( Q \) the observed number of transversions, we have

\[ d = -\frac{1}{2} \ln(1 - 2P - Q) - \frac{1}{4} \ln(1 - 2Q). \]

**Tajima-Nei**

Distance measure for genetic sequence data (CAGT). Similar to Jukes-Cantor distance, but does not assume equal nucleotide frequencies.

**User-defined similarity**

Expects a symmetric similarity matrix rather than original data. No error checking!
User-defined distance

Expects a symmetric distance matrix rather than original data. No error checking!

Mixed

This option requires that data types have been assigned to columns (see Entering and manipulating data). A pop-up window will ask for the similarity/distance measure to use for each datatype. These will be combined using an average weighted by the number of variates of each type. The default choices correspond to those suggested by Gower, but other combinations may well work better. The "Gower" option is a range-normalised Manhattan distance.

All-zeros rows: Some similarity measures (Dice, Jaccard, Simpson etc.) are undefined when comparing two all-zero rows. To avoid errors, especially when bootstrapping sparse data sets, the similarity is set to zero in such cases.

Missing data: Most of these measures treat missing data (coded as '?') by pairwise deletion, meaning that if a value is missing in one of the variables in a pair of rows, that variable is omitted from the computation of the distance between those two rows. The exceptions are rho distance, using column average substitution, and Raup-Crick, which does not accept missing data.

References


Correlation

A matrix is presented with the correlations between all pairs of columns. Correlation values are given in the lower triangle of the matrix, and the probabilities that the columns are uncorrelated are given in the upper. Both parametric (Pearson) and non-parametric (Spearman and Kendall) coefficients and tests are available. Algorithms follow Press et al. (1992).
Var-covar

A symmetric matrix is presented with the variances and covariances between all pairs of columns.

*Missing data:* Supported by pairwise deletion.
F and t tests (two samples)

A number of classical, parametric statistics and tests for comparing the means and variances of two univariate samples (in two columns). Normal distribution is assumed.

Sample statistics

Means and variances are estimated as described above under Univariate statistics. The 95% confidence interval for the mean is based on the standard error for the estimate of the mean, and the t distribution. With $s$ the estimate of the standard deviation, the confidence interval is

$$\left[ \bar{x} - t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}} \right].$$

Here, $t$ has $n-1$ degrees of freedom, and $1-\alpha = 0.95$ for a 95% confidence interval.

The 95% confidence interval for the difference between the means accepts unequal sample sizes:

$$\left[ x - y - t_{(\alpha/2, df)} S_D, \quad x - y + t_{(\alpha/2, df)} S_D \right],$$

where

$$\text{SSE} = \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2$$

$$df = (n_1 - 1) + (n_2 - 1)$$

$$\text{MSE} = \frac{\text{SSE}}{df}$$
\[ n_h = \frac{2}{1/n_1 + 1/n_2} \]
\[ s_D = \sqrt{\frac{2\text{MSE}}{n_h}}. \]

**F test**

The F test has null hypothesis

\[ H_0: \text{The two samples are taken from populations with equal variance.} \]

The F statistic is the ratio of the larger variance to the smaller. The significance is two-tailed, with \( n_1 \) and \( n_2 \) degrees of freedom.

**t test**

The t test has null hypothesis

\[ H_0: \text{The two samples are taken from populations with equal means.} \]

From the standard error \( s_D \) of the difference of the means given above, the test statistic is

\[ t = \frac{\bar{x} - \bar{y}}{s_D}. \]

**Unequal variance t test**

The unequal variance t test is also known as the Welch test. It can be used as an alternative to the basic t test when variances are very different, although it can be argued that testing for difference in the means in this case is questionable. The test statistic is

\[ t = \frac{\bar{x} - \bar{y}}{\sqrt{\text{Var}(x)/n_1 + \text{Var}(y)/n_2}}. \]

The number of degrees of freedom is

\[ df = \frac{\left[ \frac{\text{Var}(x)}{n_1} + \frac{\text{Var}(y)}{n_2} \right]^2}{\left[ \frac{\text{Var}(x)/n_1}{n_1 - 1} + \frac{\text{Var}(y)/n_2}{n_2 - 1} \right]}. \]
Permutation t test

The permutation test for equality of means uses the $t$ statistic, but is non-parametric test with few assumptions. The number of permutations can be set by the user. The power of the test is limited by the sample size – significance at the $p<0.05$ level can only be achieved for $n>3$ in each sample.

**T test (one sample)**

The one-sample $t$ test is used to investigate whether the sample is likely to have been taken from a population with a given (theoretical) mean.

The 95% confidence interval for the mean is calculated using the $t$ distribution.

*Missing data:* Supported by deletion.

**F and t tests from parameters**

Sometimes publications give not the data, but values for sample size, mean and variance for two samples. These can be entered manually using the 'F and t from parameters' option in the menu. This module does not use any data from the spreadsheet.
Paired tests (t, sign, Wilcoxon)

Three statistical tests (one parametric, two non-parametric) for two samples (columns) of univariate data. The data points are paired, meaning that the two values in each row are associated. For example, the test could be for length of the left vs. the right arm in a number of people, or the diversity in summer vs. winter at a number of sites. Controlling for a “nuisance factor” (person, site) in this way increases the power of the test. The null hypothesis is:

\( \text{H}_0: \) The mean (t test) or median (sign test, Wilcoxon test) of the difference is zero.

All reported \( p \) values are two-tailed.

\( t \) test

Testing for mean difference equal to zero using the normal one-sample \( t \) test. With \( d = x_i - y_i \), we have

\[
s = \sqrt{\frac{1}{n-1} \sum (d_i - \bar{d})^2},
\]

\[
t = \frac{\bar{d}}{s/\sqrt{n}}.
\]

There are \( n-1 \) degrees of freedom. This test assumes normal distribution of the differences.
**Sign test**

The sign (binomial) test simply counts the number of cases \( n_1 \) where \( x_i > y_i \), and \( n_2 \) where \( y_i > x_i \). The number \( \max(n_1, n_2) \) is reported. The \( p \) value is exact, computed from the binomial distribution. The sign test will typically have lower power than the other paired tests, but make few assumptions.

**Wilcoxon signed rank test**

A non-parametric rank test that does not assume normal distribution. The null hypothesis is no median shift (no difference).

All rows with zero difference are first removed by the program. Then the absolute values of the differences \( |d_i| \) are ranked \( (R_i) \), with mean ranks assigned for ties. The sum of ranks for pairs where \( d_i \) is positive is \( W^+ \). The sum of ranks for pairs where \( d_i \) is negative is \( W^- \). The reported test statistic is

\[
W = \max(W^+, W^-)
\]

(note that there are several other, equivalent versions of this test, reporting other statistics).

For large \( n \) (say \( n > 10 \)), the large-sample approximation to \( p \) can be used. This depends on the normal distribution of the test statistic \( W \):

\[
E(W) = \frac{n(n+1)}{4}
\]

\[
Var(W) = \frac{n(n+1)(2n+1)}{24} - \frac{\sum_g f_g^3 - f_g}{48}.
\]

The last term is a correction for ties, where \( f_g \) is the number of elements in tie \( g \). The resulting \( z \) is reported, together with the \( p \) value.

For smaller sample sizes (or in general), the Monte Carlo significance value should be used instead. It is based on 100,000 random reassignments of values to columns, within each pair. This value will be practically identical to the exact \( p \) value (which may be included in a future version).

*Missing data: Supported by deletion of the row.*
Normality (one sample)

Three statistical tests for normal distribution of one sample of univariate data, given in one column. The data below were generated by a random number generator with uniform distribution.

For all the three tests, the null hypothesis is

\[ H_0: \text{The sample was taken from a population with normal distribution.} \]

If the given \( p(\text{normal}) \) is less than 0.05, normal distribution can be rejected. Of the three given tests, the Shapiro-Wilk is considered to be the more exact, and the two other tests (Jarque-Bera and a chi-square test) are given for reference. There is a maximum sample size of \( n=5000 \), while the minimum sample size is 3 (the tests will of course have extremely small power for such small \( n \)).

**Shapiro-Wilk test**

The Shapiro-Wilk test (Shapiro & Wilk 1965) returns a test statistic \( W \), which is small for non-normal samples, and a \( p \) value. The implementation is based on the standard code “AS R94” (Royston 1995), correcting an inaccuracy in the previous algorithm “AS 181” for large sample sizes.

**Jarque-Bera test**

The Jarque-Bera test (Jarque & Bera 1987) is based on skewness \( S \) and kurtosis \( K \). The test statistic is

\[
JB = \frac{n}{6} \left( S^2 + \frac{(K-3)^2}{4} \right).
\]

In this context, the skewness and kurtosis used are
\[
S = \frac{1}{n} \frac{\sum (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum (x_i - \bar{x})^2 \right)^{3/2}} \]

\[
K = \frac{1}{n} \frac{\sum (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum (x_i - \bar{x})^2 \right)^{2}} \]

Note that these equations contain simpler estimators than the \( G_1 \) and \( G_2 \) given above, and that the kurtosis here will be 3, not zero, for a normal distribution.

Asymptotically (for large sample sizes), the test statistic has a chi-square distribution with two degrees of freedom, and this forms the basis for the \( p \) value given by Past. It is known that this approach works well only for large sample sizes, and a future version might include a method based on Monte Carlo simulation.

**Chi-square test**

The chi-square test uses an expected normal distribution in four bins, based on the mean and standard deviation estimated from the sample, and constructed to have equal expected frequencies in all bins. The upper limits of all bins, and the observed and expected frequencies, are displayed. A warning message is given if \( n<20 \), i.e. expected frequency less than 5 in each bin. There is 1 degree of freedom. This test is both theoretically questionable and has low power, and is not recommended. It is included for reference.

*Missing data: Supported by deletion.*

**References**


Chi^2

The Chi-square test is the one to use if your data consist of the numbers of elements in different bins (compartment). For example, this test can be used to compare two associations (columns) with the number of individuals in each taxon organized in the rows. You should be a little cautious about such comparisons if any of the bins contain less than five individuals.

There are two options that you should select or not for correct results. "Sample vs. expected" should be ticked if your second column consists of values from a theoretical distribution (expected values) with zero error bars. If your data are from two counted samples each with error bars, leave this box open. This is not a small-sample correction.

"One constraint" should be ticked if your expected values have been normalized in order to fit the total observed number of events, or if two counted samples necessarily have the same totals (for example because they are percentages). This will reduce the number of degrees of freedom by one.

When "one constraint" is selected, a permutation test is available, with 10000 random replicates. For "Sample vs. expected" these replicates are generated by keeping the expected values fixed, while the values in the first column are random with relative probabilities as specified by the expected values, and with constant sum. For two samples, all cells are random but with constant row and column sums.

With one constraint and a 2x2 table, the Fisher's exact test is also given (two-tailed). When available, the Fisher's exact test is far superior to the chi-square.

See Brown & Rothery (1993) or Davis (1986) for details.

Missing data: Supported by deletion.
Coefficient of variation

This module tests for equal coefficient of variation in two samples, given in two columns.

The coefficient of variation (or relative variation) is defined as the ratio of standard deviation to the mean in percent, and is computed as:

\[ CV = \frac{s}{\bar{x}} \cdot 100 = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \cdot \frac{1}{\bar{x}} \cdot 100. \]

The null hypothesis if the statistical test is:

\[ H_0: \text{The samples were taken from populations with the same coefficient of variation.} \]

If the given \( p(\text{normal}) \) is less than 0.05, normal distribution can be rejected. Donnelly & Kramer (1999) describe the coefficient of variation and review a number of statistical tests for the comparison of two samples. They recommend the Fligner-Killeen test (Fligner & Killeen 1976), as implemented in Past. This test is both powerful and is relatively insensitive to distribution. The following statistics are reported:

\( T \): The Fligner-Killeen test statistic, which is a sum of transformed ranked positions of the smaller sample within the pooled sample (see Donnelly & Kramer 1999 for details).

\( E(T) \): The expected value for \( T \).

\( z \): The z statistic, based on \( T \), \( \text{Var}(T) \) and \( E(T) \). Note this is a large-sample approximation.

\( p \): The \( p(H_0) \) value. Both the one-tailed and two-tailed values are given. For the alternative hypothesis of difference in either direction, the two-tailed value should be used. However,
the Fligner-Killeen test has been used to compare variation within a sample of fossils with variation within a closely related modern species, to test for multiple fossil species (Donnelly & Kramer 1999). In this case the alternative hypothesis might be that CV is larger in the fossil population, if so then a one-tailed test can be used for increased power.

The screenshot above reproduces the example of Donnelly & Kramer (1999), showing that the relative variation within Australopithecus afarensis is significantly larger than in Gorilla gorilla. This could indicate that A. afarensis represents several species.

References


Mann-Whitney test

The two-tailed (Wilcoxon) Mann-Whitney U test can be used to test whether the medians of two independent samples are different. It does not assume normal distribution, but does assume equal-shaped distribution in both groups. The null hypothesis is

\[ H_0: \text{The two samples are taken from populations with equal medians.} \]

This test is non-parametric, which means that the distributions can be of any shape.

For each value in sample 1, count number of values in sample 2 that are smaller than it (ties count 0.5). The total of these counts is the test statistic U (sometimes called T). If the value of U is smaller when reversing the order of samples, this value is chosen instead (it can be shown that \(U_1 + U_2 = n_1n_2\)).

In the left column is given an asymptotic approximation to \(p\) based on the normal distribution (two-tailed), which is only valid for large \(n\). It includes a continuity correction and a correction for ties:

\[
z = \frac{U - n_1n_2/2 + 0.5}{\sqrt{\frac{n_1n_2(n^3 - n - \sum_{g} f_g^2 - f_g)}{12n(n-1)}}}
\]

where \(n = n_1 + n_2\) and \(f_g\) is the number of elements in tie \(g\).

For \(n_1 + n_2 \leq 30\) (e.g. 15 values in each group), an exact \(p\) value is given, based on all possible combinations of group assignment. If available, always use this exact value. For larger samples, the asymptotic approximation is quite accurate. A Monte Carlo value based on 10 000 random assignments is also given – the purpose of this is mainly as a control on the asymptotic value.

\textit{Missing data: Supported by deletion.}
Kolmogorov-Smirnov

The Kolmogorov-Smirnov test is a nonparametric test for overall equal distribution of two univariate samples. In other words, it does not test specifically for equality of mean, variance or any other parameter. The null hypothesis is

\[ H_0: \text{The two samples are taken from populations with equal distribution.} \]

In the version of the test provided by Past, both columns must represent samples. You can not test a sample against a theoretical distribution (one-sample test).

The test statistic is the maximum absolute difference between the two empirical cumulative distribution functions:

\[ D = \max_x \left| S_{N_1}(x) - S_{N_2}(x) \right| \]

The algorithm is based on Press et al. (1992), with significance estimated after Stephens (1970).

Define the function

\[ Q_{KS}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2} \]

With \( N_e = N_1N_2/(N_1+N_2) \), the significance is computed as

\[ p = Q_{KS} \left( \sqrt{N_e} + 0.12 + 0.11\sqrt{N_e} D \right). \]

**Missing data:** Supported by deletion.

**References**


Spearmann/Kendall

These non-parametric rank-order tests are used to test for correlation between two variables. The permutation test is based on 1000 random replicates.

These statistics are also available through the “Correlation” module, but then without the permutation option.

*Missing data*: Supported by deletion.
Contingency table

A contingency table is input to this routine. Rows represent the different states of one nominal variable, columns represent the states of another nominal variable, and cells contain the counts of occurrences of that specific state (row, column) of the two variables. A measure and probability of association of the two variables (based on Chi-square) is then given.

For example, rows may represent taxa and columns samples as usual (with specimen counts in the cells). The contingency table analysis then gives information on whether the two variables of taxon and locality are associated. If not, the data matrix is not very informative. For details, see Press et al. (1992).
One-way ANOVA

One-way ANOVA (analysis of variance) is a statistical procedure for testing the null hypothesis that several univariate samples (in columns) are taken from populations with the same mean. The samples are assumed to be close to normally distributed and have similar variances. If the sample sizes are equal, these two assumptions are not critical. If the assumptions are strongly violated, the nonparametric Kruskal-Wallis test should be used instead.

ANOVA table

The between-groups sum of squares is given by:

$$SS_{bg} = \sum_g n_g (\bar{x}_g - \bar{x}_T)^2,$$

where $n_g$ is the size of group $g$, and the means are group and total means. The between-groups sum of squares has an associated df$_{bg}$, the number of groups minus one.

The within-groups sum of squares is

$$SS_{wg} = \sum_g \sum_i (x_{ig} - \bar{x}_g)^2,$$

where the $x_{ig}$ are those in group $g$. The within-groups sum of square has an associated df$_{wg}$, the total number of values minus the number of groups.

The mean squares between and within groups are given by

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}},$$

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}}.$$
Finally, the test statistic $F$ is computed as

$$F = \frac{MS_{bg}}{MS_{wg}}$$

The $p$ value is based on $F$ with $df_{bg}$ and $df_{wg}$ degrees of freedom.

**Levene’s test**

Levene’s test for homogeneity of variance (homoskedasticity), that is, whether variances are equal as assumed by ANOVA, is also given. Two versions of the test are included. The original Levene’s test is based on means. This version has more power if the distributions are normal or at least symmetric. The version based on medians has less power, but is more robust to non-normal distributions. Note that this test can be used also for only two samples, giving an alternative to the $F$ test for two samples described above.

**Unequal-variance (Welch) ANOVA**

If Levene’s test is significant, meaning that you have unequal variances, you can use the unequal-variance (Welch) version of ANOVA, with the $F$, $df$ and $p$ values given.

**Post-hoc pairwise tests**

If the ANOVA shows significant inequality of the means (small $p$), you can go on to study the given table of "post-hoc" pairwise comparisons, based on Tukey’s HSD (Honestly Significant Difference) test. The Studentized Range Statistic $Q$ is given in the lower left triangle of the array, and the probabilities $p(equal)$ in the upper right. Sample sizes do not have to be equal for the version of Tukey’s test used.
Two-way ANOVA

Two-way ANOVA (analysis of variance) is a statistical procedure for testing the null hypotheses that several univariate samples have the same mean across each of the two factors, and that there are no dependencies (interactions) between factors. The samples are assumed to be close to normally distributed and have similar variances. If the sample sizes are equal, these two assumptions are not critical.

Three columns are needed. First, a column with the levels for the first factor (coded as 1, 2, 3 etc.), then a column with the levels for the second factor, and finally the column of the corresponding measured values.

The algorithm uses weighted means for unbalanced designs.
Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric ANOVA, comparing the medians of several univariate groups (given in columns). It can also be regarded as a multiple-group extension of the Mann-Whitney test (Zar 1996). It does not assume normal distribution, but does assume equal-shaped distribution for all groups. The null hypothesis is

H₀: The samples are taken from populations with equal medians.

The test statistic \( H \) is computed as follows:

\[
H = \frac{12}{n(n+1)} \left( \sum_{g} \frac{T_g^2}{n_g} \right) - 3(n+1),
\]

where \( n_g \) is the number of elements in group \( g \), \( n \) is the total number of elements, and \( T_g \) is the sum of ranks in group \( g \).

The test statistic \( H_c \) is adjusted for ties:

\[
H_c = \frac{H}{\sum f_i^3 - f_i},
\]

\[
1 - \frac{1}{n^3 - n}
\]

where \( f_i \) is the number of elements in tie \( i \).

With \( G \) the number of groups, the \( p \) value is approximated from \( H_c \) using the chi-square distribution with \( G-1 \) degrees of freedom. This is less accurate if any \( n_g < 5 \).
Post-hoc pairwise tests

Mann-Whitney pairwise test $p$ values are given for all $N_p=G(G-1)/2$ pairs of groups, in the upper right triangle of the matrix. The lower right triangle gives the corresponding $p$ values, but multiplied with $N_p$ as a conservative correction for multiple testing (Bonferroni correction). The values use the asymptotic approximation described under the Mann-Whitney module. If samples are very small, it may be useful to run the exact test available in that module instead.

*Missing data: Supported by deletion.*

Reference

One-way ANCOVA

ANCOVA (Analysis of covariance) tests for equality of means for several univariate groups, adjusted for covariance with another variate. ANCOVA can be compared with ANOVA, but has the added feature that for each group, variance that can be explained by a specified "nuisance" covariate (x) is removed. This adjustment can increase the power of the test substantially.

The program expects two or more pairs of columns, where each pair (group) is a set of correlated x-y data (means are compared for y, while x is the covariate). The example below uses three pairs (groups).

The program presents a scatter plot and linear regression lines for all the groups. The ANOVA-like summary table contains sum-of-squares etc. for the adjusted means (between-groups effect) and adjusted error (within-groups), together with an F test for the adjusted means. An F test for the equality of regression slopes (as assumed by the ANCOVA) is also given. In the example, equal adjusted means in the three groups can be rejected at \( p < 0.005 \). Equality of slopes can not be rejected (\( p = 0.21 \)).

"View groups" gives the summary statistics for each group (mean, adjusted mean and regression slope).

Assumptions include similar linear regression slopes for all groups, normal distributions, similar variance and sample sizes.
Mixture analysis

Mixture analysis is a maximum-likelihood method for estimating the parameters (mean, standard deviation and proportion) of two or more univariate normal distributions, based on a pooled univariate sample. For example, the method can be used to study differences between sexes (two groups), or several species, or size classes, when no independent information about group membership is available.

The program expects one column of univariate data, assumed to be taken from a mixture of normally distributed populations. In the example below, sizes of male and female gorillas have been pooled in one sample. The means, standard deviations and proportions of the two original samples have been almost perfectly recovered (see “Univariate” above).

PAST uses the EM algorithm (Dempster et al. 1977), which can get stuck on a local optimum. The procedure is therefore automatically run 20 times, each time with new, random starting positions for the means. The starting values for standard deviation are set to $s/G$, where $s$ is the pooled standard deviation and $G$ is the number of groups. The starting values for proportions are set to $1/G$. The user is still recommended to run the program a few times to check for stability of the solution (“better” solutions have less negative log likelihood values).

The Akaike Information Criterion (AIC; Akaike 1974) is calculated with a small-sample correction:

$$ AICc = 2k - 2 \ln L + \frac{2k(k + 1)}{n - k - 1} $$

where $k$ is the number of parameters, $n$ the number of data points and $L$ the likelihood of the model given the data. A minimal value for AIC indicates that you have chosen the number of groups that produces the best fit without overfitting.

The program can also assign each of the data points to one of the groups. This can be used as a non-hierarchical clustering method for univariate data.
Missing data: Supported by deletion.

References


Genetic sequence stats

A number of simple statistics on genetic sequence (DNA or RNA) data. The module expects a number of rows, each with a sequence. The sequences are expected to be aligned and of equal length including gaps (coded as ‘?’). Some of these statistics are useful for selecting appropriate distance measures elsewhere in Past.

Total length: The total sequence length, including gaps, of one sequence

Average gap: The number of gap positions, averaged over all sequences

Average A, T/U, C, G: The average number of positions containing each nucleotide

Average p distance: The $p$ distance between two sequences, averaged over all pairs of sequences. The $p$ (or Hamming) distance is defined as the proportion of unequal positions

Average Jukes-Cantor $d$: The Jukes-Cantor $d$ distance between two sequences, averaged over all pairs of sequences. $d = -3\ln(1 - 4p/3)/4$, where $p$ is the $p$ distance

Maximal Jukes-Cantor $d$: Maximal Jukes-Cantor distance between any two sequences

Average transitions ($P$): Average number of transitions (a$\leftrightarrow$g, c$\leftrightarrow$t, i.e. within purines, pyrimidines)

Average transversions ($Q$): Average number of transversions (a$\leftrightarrow$t, a$\leftrightarrow$c, c$\leftrightarrow$g, t$\leftrightarrow$g, i.e. across purines, pyrimidines)

$R=P/Q$: The transition/transversion ratio

Missing data: Treated as gaps.