

Appendix E - The short R glossary

This very short glossary will help to find the corresponding R command for the most widespread statistical terms. This is similar to the “reverse index” which might be useful when you know what to do but do not know which R command to use.

Akaike’s Information Criterion, AIC – `AIC()` – criterion of the model optimality; the best model usually corresponds with minimal AIC.

analysis of variance, ANOVA – `aov()` – the family of parametric tests, used to compare multiple samples.

analysis of covariance, ANCOVA – `lm(response ~ influence*factor)` – just another variant of linear models, compares several regression lines.

“apply family” – `aggregate()`, `apply()`, `lapply()`, `sapply()`, `tapply()` and others — R functions which help to avoid *loops*, repeats of the same sequence of commands. Differences between most frequently used functions from this family (applied on data frame) are shown on Figure 1.

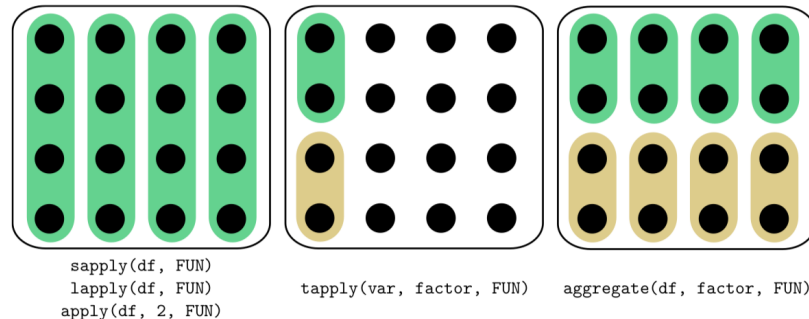


Figure 1 Five frequently used functions from “apply family”.

arithmetic mean, mean, average – `mean()` – sum of all sample values divides to their number.

bar plot – `barplot()` – the diagram to represent several numeric values (e.g., counts).

Bartlett test – `bartlett.test()` – checks the null if variances of samples are equal (ANOVA assumption).

bootstrap – `sample()` and many others – technique of sample sub-sampling to estimate population statistics.

boxplot – `boxplot()` – the diagram to represent main features of one or several samples.

Chi-squared test – `chisq.test()` – helps to check if there is a association between rows and columns in the contingency table.

cluster analysis, hierarchical – `hclust()` – visualization of objects’ dissimilarities as dendrogram (tree).

confidence interval – the range where some population value (mean, median *etc.*) might be located with given probability.

correlation analysis – `cor.test()` – group of methods which allow to describe the determination between several samples.

correlation matrix – `cor()` – returns correlation coefficients for all pairs of samples.

data types – there is a list (with synonyms):

- measurement:
 - continuous;
 - meristic, discrete, discontinuous;
- ranked, ordinal;
- categorical, nominal.

distance matrix – `dist()`, `daisy()`, `vegdist()` – calculates distance (dissimilarity) between objects.

distribution – the “layout”, the “shape” of data; *theoretical distribution* shows how data should look whereas *sample distribution* shows how data looks in reality.

F-test – `var.test()` – parametric test used to compare variations in two samples.

Fisher’s exact test – `fisher.test()` – similar to chi-squared but calculates (not estimates) p-value; recommended for small data.

generalized linear models – `glm()` – extension of linear models allowing (for example) the binary response; the latter is the logistic regression.

histogram – `hist()` – diagram to show frequencies of different values in the sample.

interquartile range – `IQR()` – the distance between second and fourth quartile, the robust method to show variability.

Kolmogorov-Smirnov test – `ks.test()` – used to compare two distributions, including comparison between sample distribution and normal distribution.

Kruskal-Wallis test – `kruskal.test()` – used to compare multiple samples, this is nonparametric replacement of ANOVA.

linear discriminant analysis – `lda()` – multivariate method, allows to create classification based on the training sample.

linear regression – `lm()` – researches linear relationship (linear regression) between objects.

long form – `stack()`; `unstack()` – the variant of data representation where group (feature) IDs and data are both vertical, in columns:

SEX SIZE

M 1

M 1

F 2

F 1

LOESS – `loess.smooth()` – Locally wEighted Scatterplot Smoothing.

McNemar’s test – `mcnemar.test()` – similar to chi-squared but allows to check association in case of paired observations.

Mann-Whitney test – `wilcox.test()` – see the Wilcoxon test.

median – `median()` – the value splitting sample in two halves.

model formulas – `formula()` – the way to describe the statistical model briefly:

- `response ~ influence`: analysis of the regression;
- `response ~ influence1 + influence2`: analysis of multiple regression, additive model;
- `response ~ factor`: one-factor ANOVA;
- `response ~ factor1 + factor2`: multi-factor ANOVA;
- `response ~ influence * factor`: analysis of covariation, model with interactions, expands into “`response ~ influence + influence : factor`”.

Operators used in formulas:

- all predictors (influences and factors) from the previous model (used together with `update()`);
- adds factor or influence;
- removes factor or influence;
- interaction;
- all logical combinations of factors and influences;
- inclusion, “`factor1 / factor2`” means that `factor2` is embedded in `factor1` (like street is “embedded” in district, and district in city);
- condition, “`factor1 | factor2`” means “split `factor1` by the levels of `factor2`”;
- intercept, so `response ~ influence - 1` means linear model without intercept;
- returns arithmetical values for everything in parentheses. It is also used in `data.frame()` command to skip conversion into factor for character columns.

multidimensional scaling, MDS – `cmdscale()` – builds something like a map from the distance matrix.

multiple comparisons – `p.adjust()` – see XKCD comic for the best explanation (Figure 2).

nonparametric – not related with a specific theoretical distribution, useful for the analysis of arbitrary data.

normal distribution plot – `plot(density(rnorm(1000000)))` – “bell”, “hat” (Figure 3).

normal distribution – `rnorm()` – the most important theoretical distribution, the basement of parametric methods; appears, for example if one will shot into the target for a long time and then measure all distances to the center (Figure 4):



Figure 2 Multiple comparisons (taken from XKCD, <http://xkcd.com/882/>).

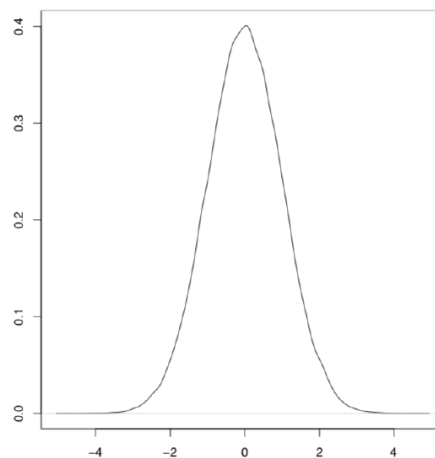


Figure 3 Normal distribution plot.

one-way test – `oneway.test()` – similar to simple ANOVA but omits the homogeneity of variances assumption.

pairwise t-test – `pairwise.t.test()` – parametric *post hoc* test with adjustment for multiple comparisons.

pairwise Wilcoxon test – `pairwise.wilcox.test()` – nonparametric *post hoc* test with adjustment for multiple comparisons.

parametric – corresponding with the known (in this book: normal, see) distribution, suitable to the analysis of the normally distributed data.

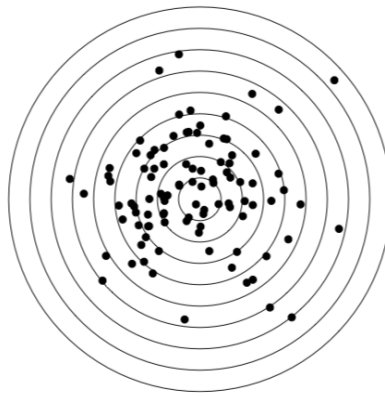


Figure 4 Similar to shooting practice results? But this is made in R using two normal distributions (see the code above)!

post hoc – tests which check all groups pairwise; contrary to the name, it is not necessary to run them after something else.

principal component analysis – `princomp()`, `prcomp()` – multivariate method “projected” multivariate cloud onto the plane of principal components.

proportion test – `prop.test()` – checks if proportions are equal.

p-value – probability to obtain the estimated value if the null hypothesis is true; if p-value is below the threshold then null hypothesis should be rejected (see the “two-dimensional data” chapter for the explanation about statistical hypotheses).

robust – not so sensitive to outliers, many robust methods are also nonparametric.

quantile – `quantile()` – returns values of quantiles (by default, values which cut off 0, 25, 50, 75 and 100% of the sample).

scatterplot – `plot(x, y)` – plot showing the correspondence between two variables.

Shapiro-Wilk test – `shapiro.test()` – test for checking the normality of the sample.

short form – `stack()`; `unstack()` – the variant of data representation where group IDs are horizontal (they are columns):

M.SIZE F.SIZE

1 2

1 1

standard deviation – `sd()` – square root of the variance.

standard error, SE – `sd(x)/sqrt(length(x))` – normalized variance.

stem-and-leaf plot – `stem()` – textual plot showing frequencies of values in the sample, alternative for histogram.

t-test – `t.test()` – the family of parametric tests which are used to estimate and/or compare mean values from one or two samples.

Tukey HSD – `TukeyHSD()` – parametric *post hoc* test for multiple comparisons which calculates Tukey Honest Significant Differences (confidence intervals).

Tukey’s line – `line()` – linear relation fit robustly, with medians of subgroups.

uniform distribution – `runif()` – distribution where every value has the same probability.

variance – `var()` – the averaged difference between mean and all other sample values.

Wilcoxon test – `wilcox.test()` – used to estimate and/or compare medians from one or two samples, this is the nonparametric replacement of the t-test.