http://babrone.edu.in/blog/?p=3119

BioNE
E-ZINE OF BIOLOGICAL SCIENCES
ISSN: 2456-7264 | Issue – 15 | Published On 07/09/2020

Bio-Statistics Using R Programming

Dr. Shantanu Tamuly, Assistant Professor, Department of Veterinary Biochemistry, College of Veterinary Science, Assam Agricultural University, Khanapara, Guwahati - 781022

In the modern world of biological science, the understanding of bio-statistics is vital for unearthing the information hidden inside the vast amount of biological data. It has been used by biological scientists to evaluate the results of their experiments and constructing various mathematical bio-models. In broad terms it is said that the statistics helps in making decisions in the face of uncertainty. The important data-based problems encountered in biological research that are addressed by bio-statistics are numerical observations made in a biological experiment and are generally found to be scattered. The experimenter is interested to know if the differences observed in various groups of experiment are due to certain cause or due to natural background variations. The experimenter is required to evaluate what do the numbers or values obtained in the experiment indicate.

The huge calculations and formulae used in various statistical methods look scary to students of biological sciences and appear to be highly time-consuming. The use of computer softwares is highly recommended for this purpose. There are scores of commercial softwares are available that can serve the purpose of a biological scientists such as SPSS, SAS, GraphPad, Stata, Minitab and so on. These softwares are user friendly and easy to handle and are recognized by high rated international scientific journals. Though, these softwares are highly useful to a biologist, these are highly expensive and are beyond the reach of many institutes. So, there is always a need of some statistical softwares that are available in open source. The R is the statistical programming language that is freely available and widely used for statistical computing. The users of R spans from fields like agricultural science, bioinformatics, medical science, animal science etc. to financial institutions, social sciences and so on. Switzerland ranks first in the world in terms of number of R users followed by New Zealand. India holds 119th rank in the world. As per the website rapporter.net, 60 downloads of the R and its different packages per 100,000 persons were reported in India till date. Globally, there are over 2 million users of R. Every year, the number of users of R increases manifold. The strong reasons behind the popularity of R are as follows:-

1. This software is free and hence many research and academic institutes with financial crunch can afford to use it.
2. It has excellent built-in help system that can educate the new user.
3. The commands used in various steps of statistical analysis are easy to comprehend and the people having no exposure to computer programming language do not find it hard.
4. The advanced users can create functions and packages for purpose of specific area of analysis.
5. The commands used for performing a particular analysis can be stored for future purposes. These commands can be used to carry out similar analysis in future.
6. The R has excellent graphical environment. The base package of R can make most of the graphical plots. However, the specific R package called “ggplot2” is used for high level graphics that are of the standard of international scientific journals.
7. The software is quite light and hence does not put heavy burden on a computer.

The R is an upgradation of another programming language called S that was developed by John Chambers in the year 1976. The programming language R was created by Ross Ihaka and Robert Gentleman in the University of Auckland (New Zealand). The project was started in the year 1992 and its final version was released in the year 2000. The R program is updated regularly by a team of developers termed as "R Core Development Team". The capability of R has been extended to cover many specific fields such as bioinformatics, pharmacology, molecular biology etc. with the development of user-defined R packages. These specific R packages are available in repositories such as Comprehensive R Archive network (CRAN), Bioconductor (packages related to bio-informatics) and GitHub.

The working of the basic R package can be improved by use of some special interface called R-studio. This interface make it easy to upload the data from excel or csv file format. In addition, while typing the commands in the R-studio, the suggestions for the correct command are displayed, hence lowering the possibility of typing incorrect commands.

I am enlisting few R packages specific for fields of biological sciences.

1. Agricole: This is the package specific for agricultural research. It has broad range of applications such as design of agricultural experiments, algorithms related to experiments based on improvement of plants and other applications such as multiple comparison of means that includes methods of Bonferroni, Duncan, Student-Newman-Keuls, Scheffe, Ryan, Einot and Gabriel and Welsch multiple range tests in addition to LSD and Tukey HSD tests. The designs of experiments that it contains are lattice, alpha, cyclic, balanced incomplete block design, Latin, Graeco-latin, augmented block design, split plot and strip plot (Mendiburu, 2020).

2. pcr: This package is used for analysis of data generated from real time PCR (quantitative PCR). It can calculate the amplification efficiencies and curves from q-PCR data. It calculates the relative expression from the q-PCR data using delta-delta CT and standard curve method. It also tests the statistical significance of the test groups and linear regression (Ahmed, 2020).

3. openPrimR: this R-package is used for designing, evaluating and comparing primer sets for multiplex PCR (Doring, 2020).

4. msa: this R-package is used for multiple sequence alignment. It performs the multiple sequence alignment based on algorithms ClustalW, ClustalOmega and Muscle (Bonatesta *et al*., 2020).

5. ggplot2: This is a data visualization R-package. It can be used to make various types of plots of publication standard such as bar plots, line plots, pie plots, histograms, correlograms, strip plots etc. Importantly, this package can be used to overlay two different types of plots such as overlaying of strip plots with bar plot or line plots. This package has the option of adding the graphical parameters in layer by layer mode (Wickham *et al*., 2020).

6. clinPK: this R-package is commonly used in clinical pharmacokinetics and clinical pharmacology. It contains the algorithms for dose individualization, compartmental pharmacokinetics, drug exposure etc. (Keizer, 2017).

7. BiodiversityR: this R-package is used to analyse the data related to biodiversity (Kindt, 2020).

Working example of one way ANOVA in R

Data entry in R:

In this section, data entry and performance of ANOVA is demonstrated. The data is prepared in excel sheet in the following form.

A	B	C	D	E
0.209801	1.467962	0.634078	2.098009	0.557136
0.43667	2.871927	0.219548	4.3667	0.172311
0.180761	2.697349	0.336517	1.807608	0.892684

(Showing only the first three rows).

The first row comprises the names of the variables (“A” to “E”). The table is copied from MS Excel. The R program is opened and the following command is entered in order to enter this data. This data is named as say “bione”.

bione = read.table(“clipboard”,header=TRUE)

(The string “header=TRUE” tells R to consider the first row as header. This command will paste the data into the R). The command gets executed by pressing enter.

The data can be seen in R by simply typing “bione” and pressing enter.

bione

Conversion of the data from wide form to long form

Now the next step is converting the data to long form. This can be done using “stack” command. This new data in long form is named as say “bione2”

bione2=stack(bione)

bione2

	values	ind
1	0.209801	A
2	0.43667	A
3	0.180761	A
4	0.894583	A
5	0.625422	A
6	0.790754	A

7	0.624502	A
8	1.467962	B
9	2.871927	B
10	2.697349	B
11	2.566964	B
12	0.643127	B

(Showing only the first 12 rows)

One Way ANOVA

Now the ANOVA is performed using the aov command. The ANOVA is named as say “bione3”

bione3=aov(values~ind,data=bione2)

summary(bione3)

	Df	**Sum-Sq**	**Mean-Sq**	**F-value**	**Pr(>F)**	
ind	4	126.25	31.562	18.53	9.13e-08	***
Residuals	30	51.09	1.703			
Signif. codes: 0 ‘*’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1**						

As the p value of ANOVA is less than 0.05, hence the multiple comparison can be performed using Tukey’s HSD test using TukeyHSD command. The comparison is named as say “bione4”.

bione4=TukeyHSD(bione3)

http://babrone.edu.in/blog/?p=3119

```
bione4
```

This gives the table of multiple comparison as follows:

	diff	lwr	upr	p.adj
B-A	1.55613839	-0.4671659	3.5794426	0.1961431
C-A	-0.20872828	-2.2320325	1.8145760	0.9981516
D-A	4.83749109	2.8141868	6.8607954	0.0000010

(Showing the first three rows only).

The grouping of the variables is carried out by using R-package “multcompView”. The package can be loaded by library command (the mulcompView package should be installed in R).

```
library(multcompView)
```

```
bione5=bione4$ind[,'p adj']
```

```
multcompLetters(bione5)
```

This puts the letter display against the variables.

B	C	D	E	A
"a"	"a"	"b"	"a"	"a"

From this output, it appears that only the group D is significantly different from other groups. While other groups does not bear statistically significant difference among them as they bear same letter.

Conclusion

The R program is one of the fast developing statistical softwares through development of newer versions. In addition, new R-packages are added to the repository each year of the specific field. Owing to its ease in handling and requirement of small space in the computer, it has become the statistical program of choice of many institutes and companies.

References:

1. Ahmed, M. (2020) Package ‘pcr’. https://cran.r-project.org/web/packages/pcr/pcr.pdf(accessed16 August 2020).

2. Bonatesta E,Horejs-Kainrath C, Bodenhofer U (2020) Package ‘msa’. http://bioconductor.org/packages/release/bioc/manuals/msa/man/msa.pdf. (accessed16 August 2020).

3. Doring M (2020)Package ‘openPrimeR’. https://www.bioconductor.org/packages/devel/bioc/manuals/openPrimeR/man/openPrimeR.pdf. (accessed16 August 2020).

4. Keizer R (2017) Package ‘clinPK’. https://cran.r-project.org/web/packages/clinPK/clinPK.pdf. (accessed16 August 2020).

5. Kindt R (2020) Package ‘BiodiversityR’. https://cran.r-project.org/web/packages/BiodiversityR/BiodiversityR.pdf. (accessed16 August 2020).

6. Mendiburu F (2020)Agricolae tutorial (version 1.3-3) https://cran.r-project.org/web/packages/agricolae/vignettes/tutorial.pdf(accessed16 August 2020).

7. R activity around the world. http://rapporter.net/custom/R-activity/#score_user/5(accessed16 August 2020).

8. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D (2020)Package ‘ggplot2’. https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. (accessed16 August 2020).