USING THE R STATISTICAL SOFTWARE IN INITIAL TERMS OF CONTROL ENGINEERING COURSE

Marcos Antonio Cruz Moreira – macruz@iff.edu.br
IF Fluminense – Campus Macaé – Curso de Engenharia de Automação e Controle
Rodovia Amaral Peixoto km 164
27.973-030 – Macaé - RJ
Natália Souto – nati.souto@yahoo.com.br

Abstract: *This paper describes the experience of using the R package as a tool of Statistics learning in the initial terms of an engineering graduation course. The proceedings adopted during a one semester course – corresponding to the third term – are presented. Some excerpts from the assignments are commented. The work aimed the acquaintance of the students with an environment for statistical computing and graphics, as well as to facilitate the learning of that syllabus' subject through graphical facilities for data analysis. The students' opinions about the proceeding, taken one year after the closing of the term are also presented.*

Palavras-chave: *Statistics, R*

1. INTRODUCTION

The option of R environment use in engineering education in IF Fluminense (RJ) have aroused from coexistence on campus Macaé of the Control Engineering graduation course and a MSc course in Environmental Engineering. Due to the widespread use of R in environmental studies (REIMANN, C. et al., 2008) it seems straightforward to extend its application to the basics matters of graduation course. The syllabus of the Control Engineering course in IFF / Campus Macaé comprises two occasions of Probability and Statistics studies. The first one at the second term of the course, includes sampling, randomization and basic concepts of probability. The second group of classes, at the third term, comprises Discrete and Continuous Probability Distributions, Confidence Intervals and Linear Regression. In this second group of subjects the use of R as a learning tool was tested and this experience is related here. Next section shows the methodology applied during the course development, as well as examples of assignments proposed to be done by the students out of the classes. It follows some excerpts from the assignments, and the assessment of the experience performed by the students. These are now attending the 5th term of their Engineering course, half way to conclude it.

2. METODOLOGY

At all the classes the theoretical concepts were presented and furthermore, the R commands available to perform the required statistical analysis. When similar functions were also available at usual commercial software, these were presented too, though students were strongly recommended to try the R package. The exploration of graphical resources and looping possibilities of the R environment were left as assignments to be solved along with statistical calculations and delivered in a one week period. Some examples are presented ahead. The survey of students' views made one year after the closing of the discipline was a research intentional decision to avoid bias in the response of those who were still studying the subject and concerned about possible implications of their responses.

2.1. Binomial Distribution

Concerning Binomial Distribution, the proposed assignment to be in R environment developed was to calculate and plot binomial distributions from 0 to 20 successful events on a set of 20 Bernoulli experiments. This should be done for given success probabilities associated with the experiment, in the case 5 %, 10 % and 25 %.

2.2. Normal Distribution

The assignment asks the students to identify and plot using the R package the z values such that a gaussian distribution contains a given probability distribution between $\mu \pm z\sigma$.

2.3. Linear Regression

Using data from Statistical Abstracts of the United States (AGRESTI & FINLAY, 2009) students were requested to obtain pair wise plots using *pairs* function as a useful high-level plotting function. It works as a tool for displaying and exploring multivariate data, producing a scatter plot between all possible pairs of variables in a dataset, and using the same scale. Once achieved these dispersion graphs, identify, among them, those that seemed to show closer linear association between two variables. Then, based on any reasonable concept obtained from a reference literature, choose among them one as explanatory and the other as response variable, and find out the linear regression equation that could be used to estimate the relationship between them.

3. RESULTS

Among the responses to the assignments, the clearer ones were selected (from the second author) and are shown in this section. The results obtained using R environment code, show a straightforward association with theoretical concepts and also illustrate the usefulness of R graphical methods to introduce students in statistical literacy, reasoning and thinking.

3.1. Binomial Distribution

R Script and Graph (Figure 1) shown as follows

Script:

```
y1 <- array(0:0, dim=c(1,21))
y2 <- array(0:0, dim=c(1,21))
y3 <- array(0:0, dim=c(1,21))
for (j in 1:21) {y1[j]<-(dbinom(j-1,20,0.05));y2[j] <-(dbinom(j-1,20,0.10));
y3[j]<-(dbinom(j-1,20,0.25))}
i <- array(0:20, dim=c(1,21))
plot(i,y1,pch=16,col="red",type="l",lwd=4,xlab="number of events",ylab="probability")
points(i,y2,pch=16,col="green",type="l",lwd=4)
points(i,y3,pch=16,col="blue",type="l",lwd=4,lty=2)
title(main="Binomial Distribution") legend(locator(1),c("p=0.05","p=0.10","p=0.25"),
col=c("red","green","blue"),pch=16,bty="o")
```

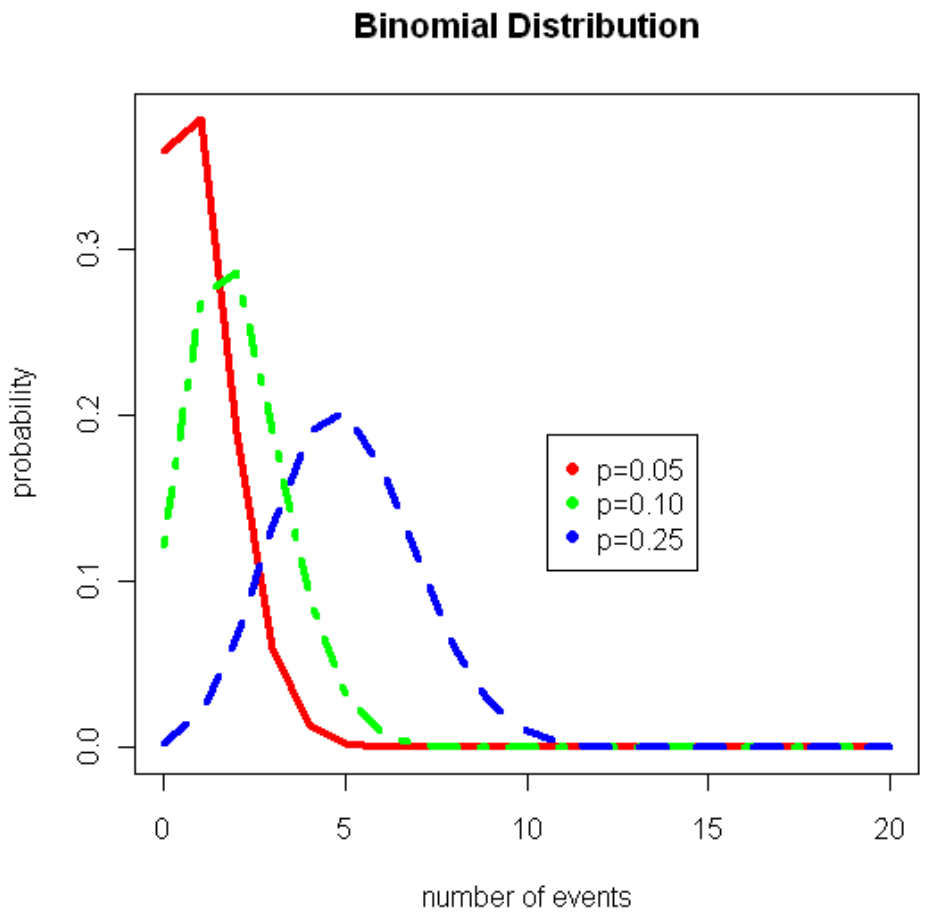

Figure 1 – Binomial Distribution Plot

3.2. Normal Distribution

R Script and Graph (Figure 2) shown below

Script:

```
band2<-function(prob) {x=seq(-6,6,length=200);
y=dnorm(x);
plot(x,y,type="l", lwd=2, col="blue");
tails<-(1-prob)/2;
lim1<-qnorm(tails,0,1)
x=seq(lim1,-lim1,length=200)
y=dnorm(x)
polygon(c(lim1,x,-lim1),c(0,y,0),col="green")}
band2(0.90); p<-0.90;
z = qnorm(0.05,0,1);
mtext(bquote(p == .(p*100)), line= 3)
mtext(bquote(z == .(z)), line= .25)
```

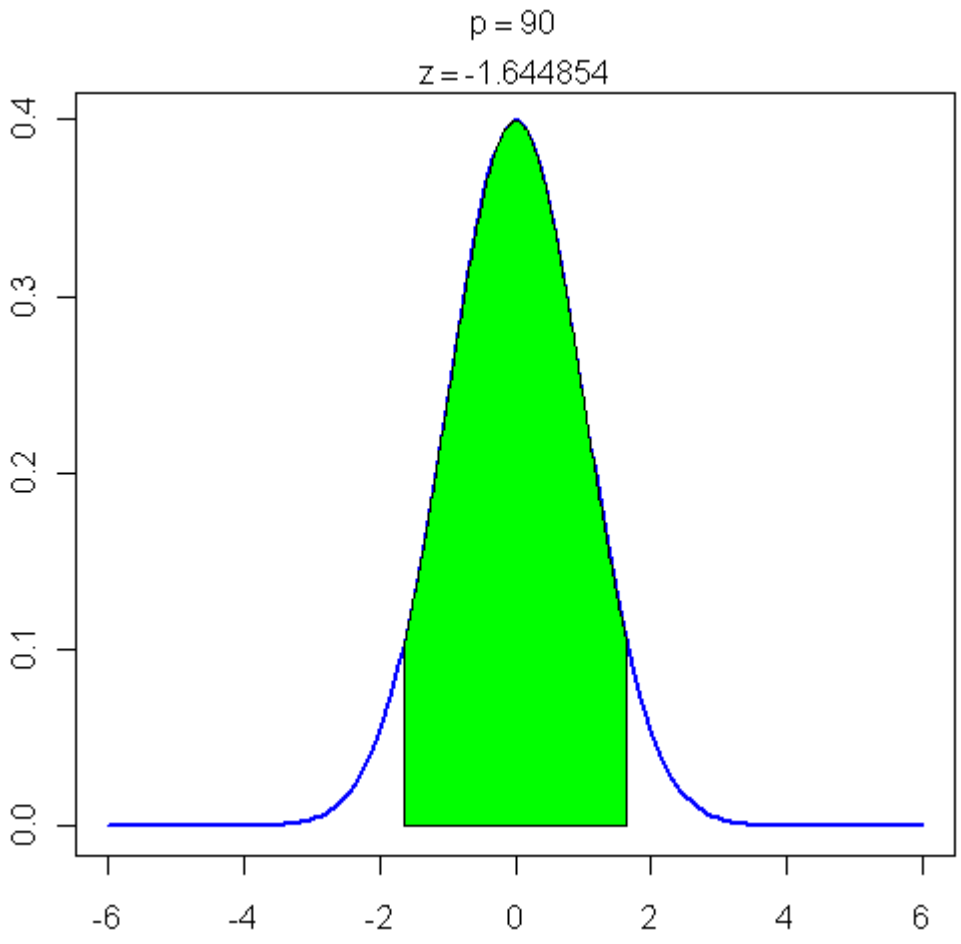

Figure 2 – Normal Distribution Plot

3.3. Linear Regression

Table 1 – Statewide Data Used to Illustrate Regression Analyses

Violent_Crime	Murder_rate	Poverty_rate	Single_Parent
761	9.0	9.1	14.3
780	11.6	17.4	11.5
593	10.2	20.0	10.7
715	8.6	15.4	12.1
1078	13.1	18.2	12.5
567	5.8	9.9	12.1
456	6.3	8.5	10.1
686	5.0	10.2	11.4
1206	8.9	17.8	10.6
723	11.4	13.5	13.0
261	3.8	8.0	9.1
326	2.3	10.3	9.0
282	2.9	13.1	9.5
960	11.4	13.6	11.5
489	7.5	12.2	10.8
496	6.4	13.1	9.9
463	6.6	20.4	10.6
1062	20.3	26.4	14.9
805	3.9	10.7	10.9
998	12.7	9.7	12.0
126	1.6	10.7	10.6
792	9.8	15.4	13.0
327	3.4	11.6	9.9
744	11.3	16.1	10.9
434	13.5	24.7	14.7

R Script and Graphs (Figure 3) shown below
Script:

```
VC <- read.table("G:/RL.TXT", header=TRUE)
# data acquisition from a txt file
cor(VC)
# correlation between variables
```

	Violent_Crime	Murder_rate	Poverty_rate	Single_Parent
Violent_Crime	1.0000000	0.7103507	0.3432887	0.5061586
Murder_rate	0.7103507	1.0000000	0.6979947	0.7574187
Poverty_rate	0.3432887	0.6979947	1.0000000	0.4930857
Single_Parent	0.5061586	0.7574187	0.4930857	1.0000000

```
pairs(VC)
# dispersion plots of  variables' pairs
lm(formula = Murder_rate ~ Single_Parent, data = VC)
# linear regression coefficients taken between the variables that appeared to show more
# association in dispersion plots
Call:
lm(formula = Murder_rate ~ Single_Parent, data = VC)
```

```
Coefficients:
  (Intercept)   Single_Parent
      -15.058          2.044
# results
```

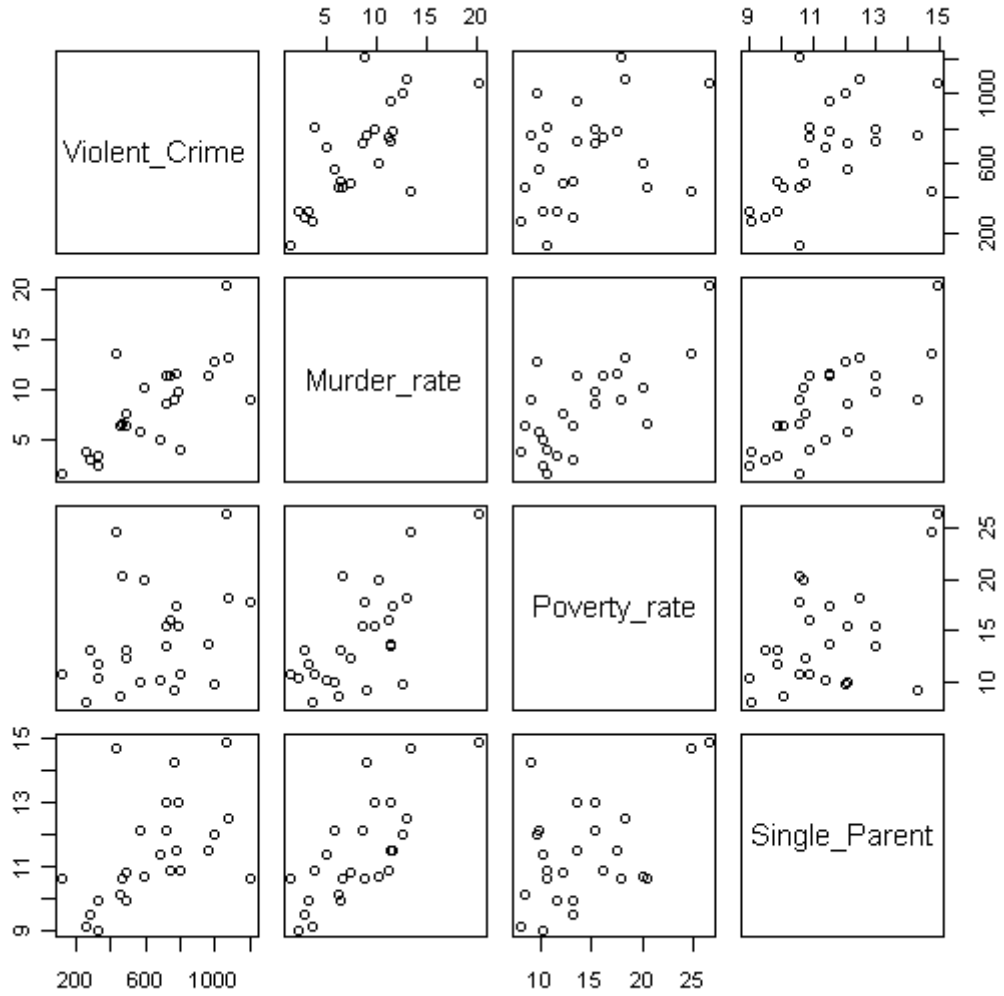

Figure 3 – Paired Observations

4. ASSESSMENT

The evaluation questionnaires were presented to 10 students taking this course, one year after the completion of the period. We sought to identify through six questions using Likert scale, the perception of students in two respects: the R package's usefulness as a tool for statistical calculations and its usefulness as a tool to support the learning process.

Counting the results, it was found predominantly positive feedback, which can be categorized as good or regular as regards the use of R as a tool for calculating and a predominantly negative evaluation categorized as fair or poor with regard to the use in the support learning process. Of course these results are subject to inaccuracies because the

sampling was limited to 1/4 of the original class and has yet a bias due to have been answered by those who made it voluntarily.

5. CONCLUSION

The software has an interface heavily based on the command line and a poor graphical user interface (GUI). Although there is extensive documentation available online in various sites, could hardly be called user friendly. These features make the R environment more familiar and interesting to students who are biased to scholarly works than to students who cares pragmatically in joining the labor market soon. Nevertheless, it constitutes a powerful tool to be used in introductory courses in statistics in engineering as it easily associates graphical representation to basic concepts of probability and statistics. Of course this does not preclude the use and applicability of the software on more advanced topics of the engineering course, given its ability to treat advanced statistics.

Thanks
To students who have completed their tasks in the course and those who responded to the survey assessment.

6. REFERENCES

AGRESTI, Alan; FINLAY, Barbara. Statistical Methods for the Social Sciences. 4. ed. New Jersey: Pearson Prentice Hall, 2009. 256 p.

REIMANN, Clemens; FILZMOSER, Peter; GARRET, Robert G.; DUTTER, Rudolf. Statistical Data Analysis Explained: Applied Environmental Statistics with R. 1. ed. West Sussex: John Wiley & Sons, Ltd., 2008.

KUHNERT, Petra; VENABLES, Bill. **An Introduction to R: Software for Statistical Modelling & Computing**. Available at <http://www.csiro.au/Organisation-Structure/Divisions/Mathematics-Informatics-and-Statistics/Rcoursenotes.aspx> Access on 08 may 2011.

R Documentation. Available at < http://www.r-project.org/> Access on 01 June 2012.