Session-12-Assignment-2

Aditya Mulay

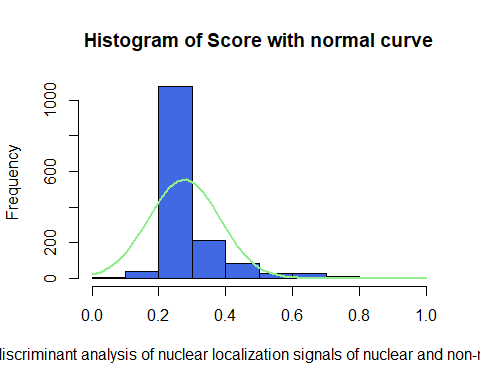
July 18, 2018

# 1. Use the given link Data Set.  
# Answer the below questions:  
Yeast <- read.table("C:/Users/Aditya/Desktop/Yeast.txt", quote="\"", comment.char="")  
View(Yeast)  
yeastdata <- Yeast  
dim(yeastdata)

## [1] 1484 10

View(yeastdata)  
library(reshape)  
yeastdata <- rename (yeastdata, c(V1 = "Sequence Name",  
 V2 = "mcg",  
 V3 = "gvh",  
 V4 = "alm",  
 V5 = "mit",  
 V6 = "erl",  
 V7 = "pox",  
 V8 = "vac",  
 V9 = "nuc",  
 V10 = "Class Distribution"))  
View(yeastdata)

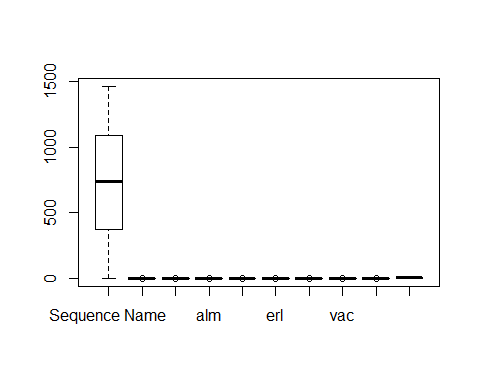
# a. What are the assumptions of ANOVA, test it out?  
# normality check  
x <- yeastdata$nuc  
h <- hist(x, breaks = 10, col = "royalblue", main = "Histogram of Score with normal curve", xlab = "Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins")  
xfit <- seq(min(x), max(x), length= 40)  
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))  
yfit <- yfit\*diff(h$mids[1:2]\* length(x))  
lines(xfit, yfit , col = "lightgreen", lwd = 2)  
#for checking skewness or kurtosis and variances   
library(psych)



describe(yeastdata)

## vars n mean sd median trimmed mad min  
## Sequence Name\* 1 1484 732.77 420.07 736.50 733.08 534.48 1.00  
## mcg 2 1484 0.50 0.14 0.49 0.49 0.12 0.11  
## gvh 3 1484 0.50 0.12 0.49 0.49 0.12 0.13  
## alm 4 1484 0.50 0.09 0.51 0.50 0.07 0.21  
## mit 5 1484 0.26 0.14 0.22 0.24 0.10 0.00  
## erl 6 1484 0.50 0.05 0.50 0.50 0.00 0.50  
## pox 7 1484 0.01 0.08 0.00 0.00 0.00 0.00  
## vac 8 1484 0.50 0.06 0.51 0.50 0.04 0.00  
## nuc 9 1484 0.28 0.11 0.22 0.25 0.00 0.00  
## Class Distribution\* 10 1484 5.13 3.04 6.00 5.21 2.97 1.00  
## max range skew kurtosis se  
## Sequence Name\* 1462.00 1461.00 -0.01 -1.19 10.90  
## mcg 1.00 0.89 0.60 0.45 0.00  
## gvh 1.00 0.87 0.42 0.55 0.00  
## alm 1.00 0.79 -0.22 1.59 0.00  
## mit 1.00 1.00 1.44 2.27 0.00  
## erl 1.00 0.50 10.14 100.87 0.00  
## pox 0.83 0.83 10.26 105.23 0.00  
## vac 0.73 0.73 -1.79 9.45 0.00  
## nuc 1.00 1.00 2.41 7.73 0.00  
## Class Distribution\* 10.00 9.00 -0.40 -1.50 0.08

#for checking outliers  
boxplot(yeastdata)



# b. Why ANOVA test? Is there any other way to answer the above question?  
  
# ANOVA allows researcher to evaluate all the mean differences in a single hypothesis test using a single ??-level and thereby keep the risk of a Type I error under control, no matter how many different means are being compared. A regression analysis will accomplish the same goal as an ANOVA.