# Problem Statement

1. Use the given link [Data Set.](https://archive.ics.uci.edu/ml/datasets/Yeast)

Answer the below questions:

yeastdata <- read.table("D:/BIG DATA/DATA ANALYTICS WITH R, EXCEL & TABLEAU/12 NON-LINEAR MODELS/yeast\_data.txt", quote="\"", comment.char="")

View(yeastdata)

dim(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)

1. What are the assumptions of ANOVA, test it out?

#To use the ANOVA test we made the following assumptions:

|  |
| --- |
| > #1) Each group sample is drawn from a normally distributed population  > #2) All populations have a common variance  > #3) All samples are drawn independently of each other  > #4) Within each sample, the observations are sampled randomly and independently of each other  > #5) Factor effects are additive  > #for checking normality assumption  > x <- yeastdata$nuc  > h <- hist(x, breaks = 10, col = "red", xlab = "Score of discriminant analysis of nuclear localization signals  + of nuclear and non-nuclear proteins", main = "Histogram of Score with normal curve")    > xfit <- seq(min(x), max(x), length= 40)  > yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))  > yfit <- yfit\*diff(h$mids[1:2]\* length(x))  > lines(xfit, yfit , col = "blue", lwd = 2)    > #for checking skewness or kurtosis and variances  > library(psych)  > describe(yeastdata)  vars n mean sd median trimmed mad min max range skew  Sequence Name\* 1 1484 732.77 420.07 736.50 733.08 534.48 1.00 1462.00 1461.00 -0.01  mcg 2 1484 0.50 0.14 0.49 0.49 0.12 0.11 1.00 0.89 0.60  gvh 3 1484 0.50 0.12 0.49 0.49 0.12 0.13 1.00 0.87 0.42  alm 4 1484 0.50 0.09 0.51 0.50 0.07 0.21 1.00 0.79 -0.22  mit 5 1484 0.26 0.14 0.22 0.24 0.10 0.00 1.00 1.00 1.44  erl 6 1484 0.50 0.05 0.50 0.50 0.00 0.50 1.00 0.50 10.14  pox 7 1484 0.01 0.08 0.00 0.00 0.00 0.00 0.83 0.83 10.26  vac 8 1484 0.50 0.06 0.51 0.50 0.04 0.00 0.73 0.73 -1.79  nuc 9 1484 0.28 0.11 0.22 0.25 0.00 0.00 1.00 1.00 2.41  Class Distribution\* 10 1484 5.13 3.04 6.00 5.21 2.97 1.00 10.00 9.00 -0.40  kurtosis se  Sequence Name\* -1.19 10.90  mcg 0.45 0.00  gvh 0.55 0.00  alm 1.59 0.00  mit 2.27 0.00  erl 100.87 0.00  pox 105.23 0.00  vac 9.45 0.00  nuc 7.73 0.00  Class Distribution\* -1.50 0.08  > #for checking outliers  > boxplot(yeastdata) |
|  |
| |  | | --- | | > | |

1. 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.

#The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant

#differences between the means of three or more independent (unrelated) groups

#The one-way ANOVA compares the means between the groups you are interested in and determines whether

#any of those means are statistically significantly different from each other

#In the situation where there multiple response variables you can test them simultaneously using a

#multivariate analysis of variance (MANOVA)

#As in salescity dataset we see that our categorical variables has more than 2 levels hence we are using

#anova t test ,otherwise if our independent variables that is city here in salescity dataset

#has got only 2 levels than we used "independent sample t test

#Thus we can use independent sample t test also if our independent variables has got only 2 levels

#the other ways for testing which we can used are by plotting

#histogram

#scatterplot

#box plot

#qq plot

#by this too

#Levene's test

#Fligner Killeen test

#Bartlett's tes