|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Assignment 12.2**  **Problem Statement**  Use the given link data set   |  |  | | --- | --- | |  | yeast <-read.table("C:/Users/Mymaster/Desktop/Data Analytics/Assignments\_Questions/yeast.txt", quote="\"", comment.char="") | |  | View(yeast) | |  | yeastdata<- yeast | |  | dim(yeastdata) | |  | install.packages("reshape") | |  | 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) | |  |  | |  | #Answer the below questions: | |  | # a. 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) | |  | #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. | |
| install.packages("RcmdrPlugin.IPSUR") |
| library(RcmdrPlugin.IPSUR) |
| head(RcmdrTestDrive) |
|  |
| mean(RcmdrTestDrive$salary) |
|  |
| install.packages("plyr") |
| library(plyr) |
|  |
| library(reshape2) |
| library(plyr) |
| library(ggplot2) |
| install.packages("RcmdrPlugin.IPSUR") |
| library(RcmdrPlugin.IPSUR) |
|  |
| #a Calculate the average salary by gender and smoking status. |
| #of salary |
| tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean) |
|  |
| #of smoking status |
| tapply(RcmdrTestDrive$salary, RcmdrTestDrive$smoking, mean) |
|  |
|  |
| #b. Which gender has the highest mean salary? |
| # genders mean salary respectively |
| #Female Male |
| #698.0911 743.3915 |
| #soits the gender male which is highest |
|  |
|  |
| #cReport the highest mean salary. |
| #if we are considering the mean of salary then |
| mean(RcmdrTestDrive$salary) |
| #724.5164 |
| #its the mean of salary |
|  |
| # if we talk about which has the highest salary of all then |
| which.max(RcmdrTestDrive$salary) |
| #152 |
| #so at 152 its the highest salary present which is 1156.16 |
|  |
|  |
| #dCompare the spreads for the genders by calculating the standard deviation of salary by gender. |
| tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, sd) |
|  |
| #Female Male |
| #130.7053 158.5423 |
|  |
| #for answering the compareness of spreads of genders lets plot boxplot |
| boxplot(salary~gender,data=RcmdrTestDrive,main="salary versus gender",xlab="gender",ylab="salary",col=topo.colors(2)) |
|  |
| #see mean too |
| tapply(RcmdrTestDrive$salary, RcmdrTestDrive$gender, mean) |
| #as from mean only there is sd deviate takes place |
|  |
| #we can plot histogram by genders to compare spreadness |
| hist(which(RcmdrTestDrive$gender=="Male") ,xlab="gender male", ylab="frequency",main="histogram of gender",col="red") |
| hist(which(RcmdrTestDrive$gender=="Female") ,xlab="gender female", ylab="frequency",main="histogram of gender",col="blue") |
|  |
|  |
| #so higher the sd higher the members of a group differ from the mean value for the group |
| #that the data spreadness in gender male is more comparatively to gender female |