PS01_Armstrong_RebeccaKatherine_15320332

2024-09-30

load libraries

set wd

clear global .envir

remove objects

rm(list=ls()) # detach all libraries detachAllPackages <- function() { basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base") package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)] package.list <- setdiff(package.list, basic.packages) if (length(package.list)>0) for (package in package.list) detach(package, character.only=TRUE) } detachAllPackages()

load libraries

pkgTest <- function(pkg){ new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])] if (length(new.pkg)) install.packages(new.pkg, dependencies = TRUE) sapply(pkg, require, character.only = TRUE) }

here is where you load any necessary packages

ex: stringr

lapply(c("stringr"), pkgTest)

lapply(c(), pkgTest)

Problem 1

y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98, 80, 97, 95, 111, 114, 89, 95, 126, 98) #Question 1 - find the CI for the same student IQ scores 'y' given a 90% confidence interval ci<-t.test(y, conf.level = 0.9) ci #The 90% confidence interval for the sample student IQ scores with a confidence level of 90% is (93.95993,102.92007) #Question 2 - considering the national average IQ is 100, determine if the average student IQ score in the school is greater than the national average, using the sample set y, with alpha=0.05 # Null Hypothesis: The average student IQ score in the school is 100

(H0: u=100) # Alternate hypothesis: The average student IQ score in the school is greater than 100 (HA: u>100) #As determining if IQ is greater than 100, use a one-tailed (right, so positive) t-test hyp<- t.test(y, mu=100, alternative="greater", conf.level=0.95) hyp #P-value is 0.7215, the data does not have any statistical importance at the 95% confidence level as p>0.05, we cannot reject the null hypothesis as the dataset of sample student IQ scores is insufficient evidence to determine if the average student IQ score of the school is greater than the national average of 100.

Problem 2

expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T) #Question 1 - examining the variables within expenditure for correlation. library(corrplot) str(expenditure) numExp<-expenditure[sapply(expenditure, is.numeric)] corRel<-cor(numExp) corrplot(corRel) #We can see the strongest variable relationship correlation between X1(per capita personal income in state) and X3(Number of people per thousand residing in urban areas in state), with a correlation of 0.4 or higher. Y(per capita expenditure on shelters/housing assistance in state) has a not as strong but still notable correlation with X1, X2(Number of residents per 100,000 that are 'financially insecure' in state), and X3. We expect to see rough trajectories of linearly increasing relationships when the scatterplots of these variables are created. Region has very weak correlational relationships with all variables except X2, with which it has a midly positive relationship. plot(numExp) #From the scatterplots generated, the relationships are all reflective of what the correlation plot matrix inferred. Region has parallel lines of point densities, which is to be expected as regions are numerical natural numbers - no decimal values were possible. (Parallel in that when region is on the X-axis we see horizontal parallel lines when all other variables are held individually against it on the y-axis, and vertical when it is repeated but with region on the y-axis and all other variables compared to it on the x-axis). The cleasrest trendline can be seen as between X1 and X3, with the densest areas of points following along an increasning linear slope (only one notable outlier can be seen on the top left corner when X1 is on the y-axis and X3 on the x-axis). Linearly increasing relationships are visible when Y is held against X1, X2 and X3, but not to the same degree of clarity as that between X1 and X3.

#Question 2 - Plotting the relationship between Y and Region: plot(expenditure$Y, expenditure$Region)

#Which region has the highest per capita expenditure on housing assistance? avRegion<-tapply(expenditure$Y, expenditure$Re mean) avRegion #from this, we can see that region 4 has the highest average per capita expenditure on housing assistance.

#Question 3 - Plot the graph between Y and X1,and describe the relationship plot(expenditure$Y, expenditure$X1) #There is an increasing linear relationship between X1 and Y. We can see that in general that as Y increases, so too does X1. This is clearly not a hard and fast rule, as there are several outliers, one in particular appearing on the left hand side of the graph with one value appearing slightly above 2000 on the X1 axis with it's corresponding Y value being only slightly beyond 40 (if this point were to follow the general trend of the other points on the graph, we would expect to see it surpassing 80 on Y, roughly double it's actual value). The points while increasing also do not have the level of density the X1-X3 graph depicted around the central increasing slope.

#Reproduce graph but include variable region, with it colour-coded and ustilising symbols to differentiate between each of the four regions colour<-c("pink","purple","orange","green")[expenditure$Region] $symbol <- -c(4, 8, 12, 16)[expenditure$Region]$ plot(expenditure$X1, expenditure$Y, col=colour, pch=symbol)