**SEMINAR 1-2**

Let’s use the dataset **attitude** embedded in R. To find information about the data attitude, use:

> help(attitude)

# list the structure of mydata

> str(mydata)

Let’s investigate the variable **rating**

> fivenum(attitude$rating)

>var(attitude$rating)

>sd(attitude$rating)

>mean(attitude$rating)

>coefvar <- 100\*sd(attitude$rating)/mean(attitude$rating)

Let’s use the dataset **mtcars** embedded in R.

**Description**

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

**Format**

A data frame with 32 observations on 11 variables.

|  |  |  |
| --- | --- | --- |
| [, 1] | mpg | Miles/(US) gallon |
| [, 2] | cyl | Number of cylinders |
| [, 3] | disp | Displacement (cu.in.) |
| [, 4] | hp | Gross horsepower |
| [, 5] | drat | Rear axle ratio |
| [, 6] | wt | Weight (1000 lbs) |
| [, 7] | qsec | 1/4 mile time |
| [, 8] | vs | V/S |
| [, 9] | am | Transmission (0 = automatic, 1 = manual) |
| [,10] | gear | Number of forward gears |
| [,11] | carb | Number of carburetors |

> boxplot(mtcars$mpg)

> boxplot(mtcars$mpg~mtcars$cyl, ylab="Miles/(US) gallon", # plot and label y-axis

 names=c("6 cylinders","8 cylinders","10 cylinders"), # group names on x-axis

 main="Miles per gallon by number of cylinders")



>dotchart(mtcars$mpg,labels=row.names(mtcars),cex=.7, main="Gas Milage for Car Models", xlab="Miles Per Gallon")



# Dotplot: Grouped Sorted and Colored

# Sort by mpg, group and color by cylinder

x <- mtcars[order(mtcars$mpg),] # sort by mpg

x$cyl <- factor(x$cyl) # it must be a factor

x$color[x$cyl==4] <- "red"

x$color[x$cyl==6] <- "blue"

x$color[x$cyl==8] <- "darkgreen"

dotchart(x$mpg,labels=row.names(x),cex=.7,groups= x$cyl,

 main="Gas Milage for Car Models\ngrouped by cylinder",

 xlab="Miles Per Gallon", gcolor="black", color=x$color)



### Create data frame with mean and std dev

x <- data.frame(mean=tapply(mtcars$mpg, list(mtcars$cyl), mean), sd=tapply(mtcars$mpg, list(mtcars$cyl), sd) )

### Add lower and upper levels of confidence intervals

x$LL <- x$mean-2\*x$sd

x$UL <- x$mean+2\*x$sd

### plot dotchart with confidence intervals

title <- "MPG by Num. of Cylinders with 95% Confidence Intervals"

dotchart(x$mean, col="blue", xlim=c(floor(min(x$LL)/10)\*10, ceiling(max(x$UL)/10)\*10), main=title )

for (i in 1:nrow(x)){

 lines(x=c(x$LL[i],x$UL[i]), y=c(i,i))

}

grid()



**Stripchart()** produces one dimensional scatter plots (or dot plots) of the given data. These plots are a good alternative to boxplots when sample sizes are small.

>stripchart(mtcars$mpg~mtcars$cyl)



> cyl<- factor(mtcars$cyl, labels=c("6 cylinders","8 cylinders","10 cylinders"))

>stripchart(mtcars$mpg~cyl, method="stack",pch=19, xlab="miles per gallon")



> stripchart(mtcars$mpg~cyl,method="stack", pch=19, col=c("red","blue","green"), xlab="miles per gallon")

>title("Miles per gallon by cylinders")



BARPLOT

data<-table(mtcars$gear,mtcars$cyl)

data

par(mfrow=c(2,2))

barplot(data)

barplot(data, beside=TRUE)

barplot(t(data))

barplot(t(data), beside=TRUE)

par(mfrow=c(1,1))

> data

 4 6 8

 3 1 2 12

 4 8 4 0

 5 2 1 2



ex <- factor(mtcars$gear,levels=c("3", "4","5"))

barplot(table(ex,mtcars$cyl),beside=TRUE, legend=TRUE,ylab="Frequency", col=c("green","blue","red"),main="Gear by number of cylinders")



**PIECHARTS**

table(mtcars$cyl)

mtcars$cyl

pie(table(mtcars$cyl))



cyl.tip<-table(mtcars$cyl)

names(cyl.tip) <-c("4 cylinders", "6 cylinders", "8 cylinders")

cyl\_procent<-round(cyl.tip/length(mtcars$cyl)\*100,1)

cyl\_procent<-paste(cyl\_procent, "%", sep="")

pie(cyl.tip, main="Pie Chart - Cars by number of cylinders", col=heat.colors(3), labels=cyl\_procent)

legend("topright", 0.5, c("4 cylinders", "6 cylinders", "8 cylinders"), cex=0.6, fill=heat.colors(3))



#piechart 3D

cyl.tip<-table(mtcars$cyl)

cyl\_procent<-round(cyl.tip/length(mtcars$cyl)\*100,1)

cyl\_procent<-paste(cyl\_procent, "%", sep="")

names(cyl.tip) <-c("4 cylinders", "6 cylinders", "8 cylinders")

lbls<-paste(names(cyl.tip), cyl\_procent, sep=" ", cex=0.5)

library(plotrix)

pie3D(cyl.tip, main="Pie Chart cars by number of cylinders", col=heat.colors(3), labels=cyl\_procent)

legend("topright", 0.5, c("4 cylinders", "6 cylinders", "8 cylinders"), cex=0.6, fill=heat.colors(3))



#Stem and leaf plot

stem(mtcars$mpg,scale=2)

 The decimal point is at the |

 10 | 44

 12 | 3

 14 | 3702258

 16 | 438

 18 | 17227

 20 | 00445

 22 | 88

 24 | 4

 26 | 03

 28 |

 30 | 44

 32 | 49

HISTOGRAm

library(MASS)

par(mfrow=c(1,2)) # Make device region 1 by 2

hist(mtcars$mpg,prob=TRUE)

lines(density(mtcars$mpg)) # Add density to Histogram

plot(density(mtcars$mpg)) # Create density by itself

par(mfrow=c(1,1))



par(mfrow=c(1,2)) # Make device region 1 by 2

hist(mtcars$mpg,prob=TRUE,col="cyan", main="Histogram cars")

lines(density(mtcars$mpg,),lwd=3,col="red")

plot(density(mtcars$mpg,),lwd=3,col="red",main="")

title(main="Density Plot")

par(mfrow=c(1,1))



#Check normality

qqnorm(mtcars$mpg)

qqline(mtcars$mpg)



hist(mtcars$mpg,prob=TRUE,col="cyan", main="Histogram cars")

degM = mean(mtcars$mpg)

degSD = sd(mtcars$mpg)

curve(dnorm(x,degM, degSD), add = T, col = "red") #add the normal curve



#Check normality

plot(density(mtcars$mpg))

curve(dnorm(x, degM, degSD), add = T, col = "red")



> mean(mtcars$mpg)

[1] 20.09062

> median(mtcars$mpg)

[1] 19.2

> quantile(mtcars$mpg)

 0% 25% 50% 75% 100%

10.400 15.425 19.200 22.800 33.900

> sort(mtcars$mpg)

 [1] 10.4 10.4 13.3 14.3 14.7 15.0 15.2 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.2 19.2 19.7 21.0

[20] 21.0 21.4 21.4 21.5 22.8 22.8 24.4 26.0 27.3 30.4 30.4 32.4 33.9

> fivenum(mtcars$mpg)

[1] 10.40 15.35 19.20 22.80 33.90

> var(mtcars$mpg)

[1] 36.3241

> sd(mtcars$mpg)

[1] 6.026948

> range(mtcars$mpg)

[1] 10.4 33.9

> IQR(mtcars$mpg)

[1] 7.375

> summary(mtcars$mpg)

 Min. 1st Qu. Median Mean 3rd Qu. Max.

 10.40 15.42 19.20 20.09 22.80 33.90

> prop.table(table(mtcars$cyl,mtcars$mpg),1)



#Compare distributions

m1<-c(0,2,2,4,5,14,14,14,13,17,17,15)

m2<-c(0,6,7,9,11,13,16,16,16,17,18,20,21)

par(mfrow=c(1,2))

boxplot(m1,m2,names=c("Method 1", "Method 2"))

plot(density(m1), ylim=c(0, 0.07), main = "Density plots of m1 and m2")

lines(density(m2), lty=2)

par(mfrow=c(1,1))

