Chapter 3 Graphical Displays

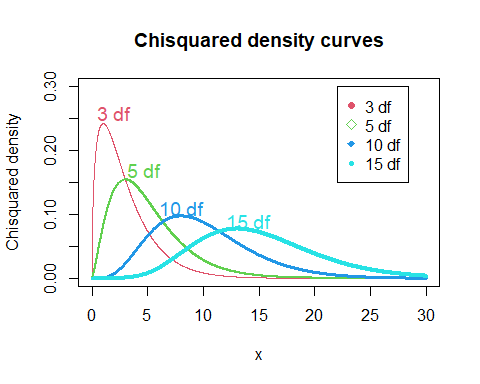
Tanweer Shapla

2023-10-10

## Drawing chi-squared density curves

If you want to draw series of graphs in one display, then first draw a single curve using plot() function and then use lines() function to add more curves.

x=seq(0,30,0.01)#generating 31 values spaced by 0.01  
chx=dchisq(x,3) # density of chisquared distribution with 3 degrees of freedom  
plot(x, chx, type="l", col=2, lwd=1,   
 xlab="x", ylab="Chisquared density", main="Chisquared density curves",   
 xlim=c(0,30), ylim=c(0,0.3))  
lines(x, dchisq(x,5), type="l", col=3, lwd=2)  
lines(x, dchisq(x,10), type="l", col=4, lwd=3)  
lines(x, dchisq(x,15), type="l", col=5, lwd=4)  
  
text( c(2, 4.7, 8, 14), c(.258, .17, .11, 0.09),  
 labels = c("3 df", "5 df", "10 df", "15 df"), col=2:5, cex=1.2)  
  
legend(22,0.3, c("3 df", "5 df", "10 df", "15 df"), col=2:5, pch = c(16,5,18,20))



type=“l” (it is letter ‘l’, not number 1) connects the dot using a line

col is for color, color can be indicated by text or number. There are eight colors that can be addressed by numbers 1 through 8.You may also spell out the color name in the col parameter. For example, col=c(“green”, “red”).

lwd for line width indicating thickness of the plotted line.

cex is for character expansion factor to control font for text or plotting symbol.

In legend(), first two numbers are used to specify the position of the legend (x,y coordinates), then the legend text, color, legend symbol.

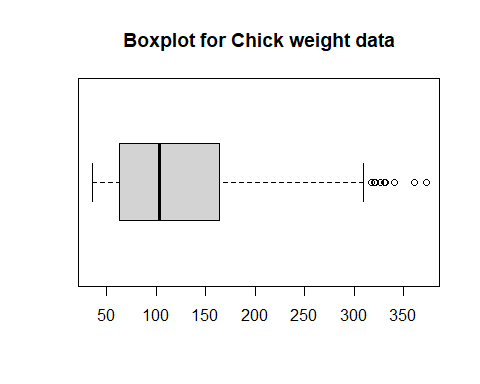
pch= option is for selecting legend symbol, pch=16 is for closed dots, pch=1 for open dot, pch=5 for diamond, etc.

## Drawing a univariate boxplot

Boxplot is a graphical display of five number summary. It also displays possible outliers in a dataset.

An observation that is more than away from the nearest quartile is considered an outlier, and is displayed by an open circle in a boxplot produced by R.

chick.wt<-ChickWeight$weight  
boxplot(chick.wt, horizontal = T, main="Boxplot for Chick weight data")

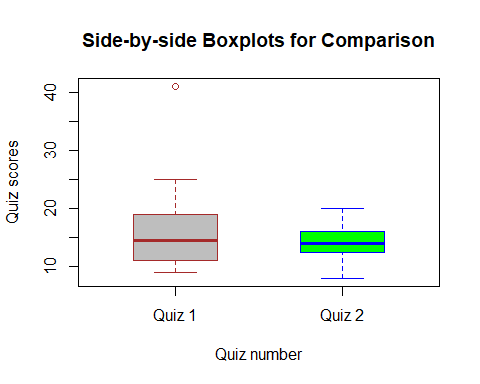


By default we get a vertical boxplot, horizontal = T produces a horizontal boxplot.

## Drawing side-by-side boxplots for comparison

We have scores from Quiz 1 and Quiz 2 for 8 students. We draw two boxplots for comparison.

Quiz1<-c(16, 13, 11, 9, 18, 12, 11, 19, 25,41)  
Quiz2<-c(20, 14, 8, 13, 12, 17, 15, 14)  
boxplot(Quiz1, Quiz2, main="Side-by-side Boxplots for Comparison",   
 col=c("grey", "green"), border=c("brown","blue"), horizontal = F,   
 ylab="Quiz scores", xlab="Quiz number", names=c("Quiz 1", "Quiz 2"), boxwex=0.5)

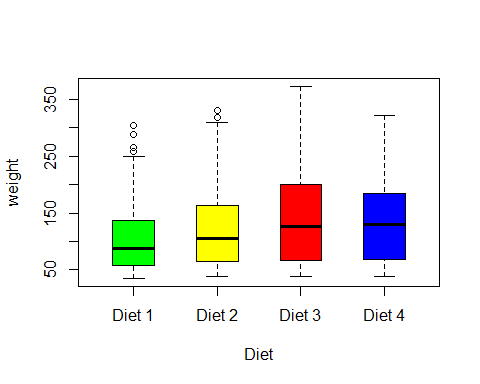


boxwex is a scale factor, the appearance of the plot can be improved by making boxes narrower.

## Drawing side-by-side boxplots for comparison: Another example

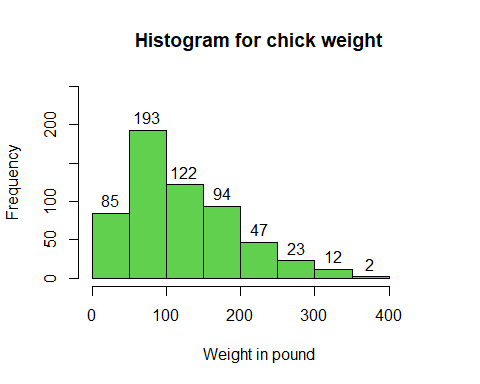
In this example, we wish to draw multiple boxplots for weight variable for each diet level for the ChickWeight data. Multiple boxplots help us compare the distribution of a response variable across different levels of a grouping variable. The function takes the form boxplot(x~y), where x is the response variable and y is the grouping variable.

boxplot(weight~Diet, data=ChickWeight, horizontal = F, col=c("green", "yellow", "red", "blue"), names=c("Diet 1", "Diet 2", "Diet 3", "Diet 4"), boxwex=0.5)



## Drawing a histogram for a univariate data

chick.wt<-ChickWeight$weight  
hist(chick.wt, main="Histogram for chick weight", col=3, xlab="Weight in pound", ylab="Frequency", labels = T, xlim=c(0,450), ylim=c(0,250))



## Drawing a bi-variate boxplot

A bi-variate boxplot is useful in identifying possible outliers in a scatterplot. It consists of a pair of concentric ellipses, the inner one containing 50% of the data, and the outer one about 95%. Any point outside the outer ellipse is considered an outlier.

In addition, regression lines of both y on x and x on y are shown, with their intersection showing the bi-variate location estimator.

The acute angle between the regression lines will be small for a large absolute value of correlations and large for a small one.

Below we will draw bi-variate boxplot for Agriculture and Examination variables from swiss data. Click on the Help tab from upper right pane, type swiss and get detail information about the data.

head(swiss)

Fertility Agriculture Examination Education Catholic  
Courtelary 80.2 17.0 15 12 9.96  
Delemont 83.1 45.1 6 9 84.84  
Franches-Mnt 92.5 39.7 5 5 93.40  
Moutier 85.8 36.5 12 7 33.77  
Neuveville 76.9 43.5 17 15 5.16  
Porrentruy 76.1 35.3 9 7 90.57  
 Infant.Mortality  
Courtelary 22.2  
Delemont 22.2  
Franches-Mnt 20.2  
Moutier 20.3  
Neuveville 20.6  
Porrentruy 26.6

names(swiss) #listing variables in swiss data

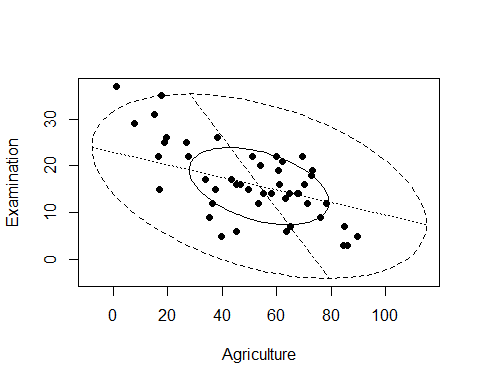
[1] "Fertility" "Agriculture" "Examination" "Education"   
[5] "Catholic" "Infant.Mortality"

attach(swiss)  
#attach() means the dataset is searched by R and objects (i.e, variables)   
#in the dataset can be accessed simply typing their names in the subsequent codes  
n=nrow(swiss) #gives number of observations in swiss data  
n

[1] 47

First we draw the bi-variate boxplot using bvbox() function from MVA package. It appears that there are two outlying values in the y direction.

#install.packages("MVA")  
library(tools)  
library(HSAUR2)  
library(MVA)  
  
bvbox(cbind(Agriculture,Examination), xlab="Agriculture", ylab="Examination",  
 pch = 16, cex = 1, col = 1)



In order to identify two outlying observations, we use

out=match(sort(Examination)[(n-1):n],Examination)  
out

[1] 42 45

swiss[out,2:3]

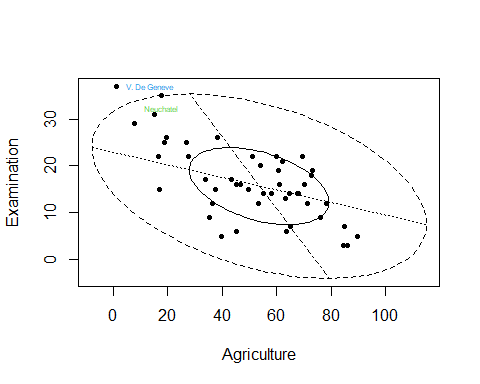
Agriculture Examination  
Neuchatel 17.6 35  
V. De Geneve 1.2 37

sort(Examination)[(n-1):n] gives the second highest and the highest values in Examination variable. Then, match() gives the index or position of argument 1 in the second, that is, the index of second highest and the highest values in Examination variable.

The last command above gives us the name of the provinces that are outlying observations displayed by the bivariate boxplot.

Now we draw the bi-variate boxplot with outlying values labeled in the graph.

bvbox(cbind(Agriculture,Examination), xlab="Agriculture", ylab="Examination",  
 pch = 16, cex = 0.7, col = 1)  
text(Agriculture[out], Examination[out], labels = c("Neuchatel", "V. De Geneve"), cex = 0.5, col=c(3,4), pos = c(1,4))



pos parameter is a position specifier for the text. Values of 1, 2, 3 and 4 respectively indicate positions below, to the left of, above and to the right of the specified (x,y) coordinates.

## Calculating correlation coefficient with and without V.De Geneve province

cor(Agriculture,Examination) #correlation with all data points included

[1] -0.6865422

cor(Agriculture[-45],Examination[-45]) #corr with V. De Geneve removed

[1] -0.6434664

## Calculating mean of Agriculture and Examination variable

mean(Agriculture)

[1] 50.65957

mean(Examination)

[1] 16.48936

#Or  
colMeans(cbind(Agriculture,Examination))

Agriculture Examination   
 50.65957 16.48936

## Calculating some summary statistics

var(Agriculture) # giving sample variance of agriculture

[1] 515.7994

cov(cbind(Agriculture,Examination)) #producing covariance matrix

Agriculture Examination  
Agriculture 515.7994 -124.39283  
Examination -124.3928 63.64662

cov(Agriculture,Examination) #giving covariance of agriculture and examination

[1] -124.3928

cor(cbind(Agriculture,Examination))# giving sample correlation matrix

Agriculture Examination  
Agriculture 1.0000000 -0.6865422  
Examination -0.6865422 1.0000000

Finally we detach the data.

detach(swiss)