**###補充教材--- Unit 1: 資料分析與趨勢判別**

data(iris)

attributes(iris)

str(iris)

**summary(iris)**

**head(iris, 5)**

**attach(iris)**

var=c(1:4)

colMeans(iris[,-5])

cor**(Sepal.Length, Sepal.Width)**

corr=cor(iris[var], use="pairwise")

corr

cov**(Petal.Length, Petal.Width)**

covv=cov(iris[var], use="pairwise")

covv

install.packages("timeDate")

library(timeDate)

skewness(iris[,1:4])

kurtosis(iris[,1:4])

偏態值 > 0，為正偏態，分配集中在平均數以下，低分群的個體較多。  
偏態值 = 0，為平均數左右對稱的分配，平均數上下的個體數相等。  
偏態值 < 0，為負偏態，分配集中在平均數以上，高分群的個體較多。

         峰度值 > 0，為高狹峰，較常態分配來得高瘦。  
峰度值 = 0，為常態峰。  
峰度值 < 0，為低闊峰，較常態分配來得低寬。

**常態性檢定**

**hist(Sepal.Length, breaks=seq(4.0, 8.0, 0.25))**

**hist(Sepal.Length, breaks=seq(4.0, 8.0, 0.25), prob=TRUE)**

**qqnorm(Sepal.Length, xlab=** "**Z-score**"**, ylab=**"**Sepal.Length**"**)  
qqline(Sepal.Length, col="red")**

**curve(dnorm(x, mean(Sepal.Length), sd(Sepal.Length)), 4.0, 8.0, col="red")**

**shapiro.test(Sepal.Length)**

**qqplot(Sepal.Length, Sepal.Width)**

install.packages("nortest")

**library(nortest)  
ad.test(Sepal.Width)**

**sf.test(Sepal.Width)**

**cvm.test(Sepal.Width)**

##先以var.test函數進行變異數相同與否的F檢定：若變異數相同，則執行t.test時設定var.equal=TRUE，若變異數不相同，則設定var.equal= FALSE或省略。

**雙群樣本平均數檢定 (F test/T test)**

**xtabs(~Species)  
setosa=subset(iris, Species=="setosa")  
versicolor=subset(iris, Species=="versicolor")  
var.test(setosa$Petal.Width, versicolor$Petal.Width)  
t.test(setosa$Petal.Width, versicolor$Petal.Width, var.equal=FALSE)**

#想檢定三個品種的Petal.Width平均數是否有顯著差異？

**變異數分析 (ANOVA)**

bartlett.test(Petal.Width~Species, data = iris)

Wid<- aov(Petal.Width~Species, data = iris)

summary(Wid)

**卡方檢定/比例檢定**

male= c(Bush=315, Perot=152, Clinton=337)  
female= c(Bush=346, Perot=126, Clinton=571)

rbind(male, female)  
chisq.test(rbind(male, female))

citizen= c(sum(male), sum(female))

bush= c(male[1], female[1])

prop.test(bush, citizen, alternative="greater")

perot= c(male[2], female[2])

prop.test(perot, citizen, alternative="two.sided")

clinton= c(male[3], female[3])

prop.test(clinton, citizen, alternative="less")

**\*\*\*\*\*以性別資料集簡單練習**

gender\_size <-read.csv("[gender\_size.csv](file:///\\gender_size.csv)")

attach(gender\_size)

gender\_size[,4]<- as.factor(gender\_size[,4])

Male=subset(gender\_size, Gender==**"male"**)

Female=subset(gender\_size, Gender==**"female"**)var.test(Male$Height, Female$Height)

t.test(Male$Height, Female$Height, var.equal=FALSE)

var.test(Male$Weight, Female$Weight)

t.test(Male$Weight, Female$Weight, var.equal=FALSE)

var.test(Male$Waist, Female$Waist)

t.test(Male$Waist, Female$Waist, var.equal=FALSE)

###**離群值檢定**

gender\_outlier <-read.csv("[gender\_outlier.csv](file:///\\gender_outlier.csv)")

attach(gender\_outlier)

Male=subset(gender\_outlier, Gender==**"male"**)

Female=subset(gender\_outlier, Gender==**"female"**)

Ma.mean<-colMeans(Male[, -4])

Fe.mean<-colMeans(Female[, -4])

Ma.var<-cov(Male[c(1:3)], use="pairwise")

Fe.var<-cov(Female[c(1:3)], use="pairwise")

Male$mdis<- mahalanobis(Male[, -4], Ma.mean, Ma.var)

Female$mdis<-mahalanobis(Female[, -4], Fe.mean, Fe.var)

Male$maout<-(Male$mdis< qchisq(df=3, p=0.95))

Female$feout<-(Female$mdis< qchisq(df=3, p=0.95))

**####線性迴歸模型的離群值測試**

MR<- lm(Waist~Weight+Height, gender\_outlier[50:146, -4])

summary(MR)

install.packages("car")

library(car)

cooks.distance(MR)

which(as.vector(cooks.distance(MR))>1) #判定outlier

outlierTest(MR)

###**機率分配**

curve(dunif(x, min=160, max=190), from=150, to=200, main="probability density")

curve(punif(x, min=160, max=190), from=150, to=200, main="cumulative density")

curve(dnorm(x, mean=50, sd=10), from=20, to=80, main="probability density")

curve(pnorm(x, mean=50, sd=10), from=20, to=80, main="probability density")

male\_height= runif(100, min=160, max=190)

x=ecdf(male\_height)

plot(x, main="empirical uniform probability")

female\_weight=rnorm(100, mean=50, sd=10)

y=ecdf(female\_weight)

plot(y, main="empirical normal probability")

arrival=rpois(60, 5)

hist(arrival)

lambda=5; k=seq(0,15)

plot(k, dpois(k, lambda), type='h', main='dpois(lambda=5)', xlab='arrival rate')

dice=rbinom(80, 10, 0.4)

hist(dice)

n=20; p=0.4; k=seq(0, n)

plot(k, dbinom(k, n ,p), type='h', main='dbinom(k=0:20, n=20, p=0.4)', xlab='k')

**\*\*\*\*\*畫出散佈圖、長條圖、盒狀圖、圓餅圖**  
**plot(iris) #散佈圖矩陣**

**plot(Sepal.Length, Sepal.Width) #散佈圖**

**plot(Species, Sepal.Length,** main="Distribution of Sepal.Length"**)**

x=boxplot(iris[, 1:4], main="Three species") #**盒狀圖**x$stats

y=which(iris[, 5]=="setosa")

boxplot(iris[y, 1:4], main="setosa")

num\_species=with(iris, c(sum(iris[,5]=="setosa"), sum(iris[,5]=="versicolor"), sum(iris[,5]=="virginica")))

barplot(num\_species, names.arg=c("setosa","versicolor","virginica"), xlab="species", ylab="number") **#長條圖**

pie(num\_species, labels= c("setosa","versicolor","virginica")) #圓餅圖

pie(num\_species, labels=c(sum(iris[,5]=="setosa"), sum(iris[,5]=="versicolor"), sum(iris[,5]=="virginica")))

percent=round(num\_species/sum(num\_species)\*100)

label=paste(paste(iris$Species,":"), percent,"%")

pie(num\_species, label)

plot(Sepal.Length[Species=="setosa"], Petal.Length[Species=="setosa"], pch=1, col="blue", xlim=c(3,8), ylim= c(0,9), main="scatter plot", xlab="SepalLen", ylab="PetalLen")

points(Sepal.Length[Species=="virginica"], Petal.Length[Species=="virginica"], pch=3, col="green")

points(Sepal.Length[Species=="versicolor"], Petal.Length[Species=="versicolor"], pch=2, col="red")

legend(3, 9, c("setosa","versicolor","virginica"), col=c(1,2,3), pch=c(1,2,3))

hist(iris$Sepal.Length, breaks=20, labels=TRUE, col="blue", border="red", main="Histogram of frequency")

hist(iris$Sepal.Length, freq=FALSE, main="Histogram of density")

lines(density(iris$Sepal.Length))

X=ecdf(iris$Sepal.Width)

plot(X, xlab="Sepal.Width", main="Cumulative frequency")

**\*\*\*\*\*自訂函數 (標準化資料集)**

scale1=function (x)

{

ncol=dim(x) [2]-1 #Y變數不進行標準化

nrow=dim(x) [1]

new=matrix(0, nrow, ncol)

for (i in 1:ncol)

{

max=max(x[, i])

min=min(x[, i])

for (j in 1:nrow)

{

new[j,i]=(x[j, i]-min)/(max-min)

}

}

new

}

**\*\*\*\*\*multiple regression (線性複迴歸)/性別資料集**

gender\_size<- read.csv("[gender\_size.csv](file:///\\gender_size.csv)")

attach(gender\_size)

#data\_gender <-gender\_size

#data\_gender[,1:3]<- scale1(gender\_size)

#data\_gender[,4]<- as.factor(data\_gender[, 4])

MR1<- lm(Waist~Weight+Height, gender\_size[,-4])

anova(MR1)

M1=summary(MR1)

names(M1)

M1$coefficients

confint(MR1, level=0.95)

predict(MR1, data.frame(Weight=68, Height=175), se=T)

attach(gender\_size)

BMI=Weight/((Height/100)^2) #加入新變數BMI

data\_BMI<-data.frame(Height, Weight, Waist, BMI, Gender)

attach(data\_BMI)

MR2<- lm(Waist~Height+Weight+BMI, data\_BMI[, -5], na.action=na.exclude)

summary(step(MR2), k=2, method="forward")

summary(step(MR2), k=log(nrow(data\_BMI)), method="backward")

confint(MR2, level=0.95)

install.packages("car")

library(car)

vif(MR2) #共線性檢定

**#One sample/ Two sample T-test**

attach(gender\_size)

sexwt=split(Weight, Gender)

sexws=split(Waist, Gender)

t.test(sexwt$female, mu=50, alternative="two.sided")

var.test(sexws$male, sexws$female)

t.test(sexws$male, sexws$female, alternative="greater", var.equal=FALSE)

**#####趨勢判斷(銀行資料集)**

bank=read.csv("bank.csv", header=TRUE, sep=";")

summary(bank)

attach(bank)

hist(age)

char=0 ###離散化變數

for (i in 1:nrow(bank))

{

if (bank[i,1]<35) char[i]="young"

else if(bank[i,1]<50) char[i]="middle"

else char[i]="old"

}

bank$agg=factor(char)

summary(bank$agg)

**###年紀與借貸是否相關**

young\_y=sum((bank$agg=="young") & (bank$loan=="yes"))

young\_n=sum((bank$agg=="young") & (bank$loan=="no"))

middle\_y=sum((bank$agg=="middle") & (bank$loan=="yes"))

middle\_n=sum((bank$agg=="middle") & (bank$loan=="no"))

old\_y=sum((bank$agg=="old") & (bank$loan=="yes"))

old\_n=sum((bank$agg=="old") & (bank$loan=="no"))

age\_y<-c(young=young\_y, middle=middle\_y, old=old\_y)

age\_n<-c(young=young\_n, middle=middle\_n, old=old\_n)

rbind(age\_y, age\_n)

chisq.test(rbind(age\_y, age\_n))

**###婚姻與借貸是否相關**

mar\_y<-c(single=nrow(subset(bank, marital=="single" & loan=="yes")), married=nrow(subset(bank, marital=="married" & loan=="yes")), divorced=nrow(subset(bank, marital=="divorced" & loan=="yes")))

mar\_n<-c(single=nrow(subset(bank, marital=="single" & loan=="no")), married=nrow(subset(bank, marital=="married" & loan=="no")), divorced=nrow(subset(bank, marital=="divorced" & loan=="no")))

rbind(mar\_y, mar\_n)

chisq.test(rbind(mar\_y, mar\_n))

**###已婚是否與借貸比例**

married\_y=sum((bank$marital=="married") & (bank$loan=="yes"))

unmarried\_y=sum((bank$marital!="married") & (bank$loan=="yes"))

x=c(married\_y, unmarried\_y)

y=c(sum((bank$marital=="married")), sum((bank$marital!="married")))

prop.test(x, y)

**###單身是否與帳戶餘額**

single=subset(bank, marital=="single")

nonsingle=subset(bank, marital!="single")

var.test(single$balance, nonsingle$balance)

t.test(single$balance, nonsingle$balance, var.equal=FALSE)

**\*\*\*one-way anova/two-way anova**

fit1<- lm(balance~marital)

summary(fit1)

ano1<- aov(balance~marital, data= bank)

summary(ano1)

fit2<-lm(balance~marital+age)

summary(fit2)

ano2 <- aov(balance~marital+age, data= bank)

summary(ano2)

fit3<-lm(balance~marital+age+marital\*age)

summary(fit3)

ano3 <- aov(balance~marital+age+marital\*age, data= bank)

summary(ano3)

interaction.plot(age, marital, balance, col=1:3)

#single\_blan=bank[marital=="single", "balance"]

#married\_blan=bank[marital=="married", "balance"]

#divorced\_blan=bank[marital=="divorced", "balance"]

#blan.survey <- data.frame( blan= c(single\_blan, married\_blan, divorced\_blan), mar= factor(c(rep("1", length(single\_blan)), rep("2", length(married\_blan)), rep("3", length(divorced\_blan)))) )

#lm(blan~mar, data= blan.survey)

#aov(blan~mar, data= blan.survey)