Chapter 4

Descriptive measures of association for bivariate distributions

Exercise 14 *Answer* :

```
> duration = faithful$eruptions  # the eruption durations
> waiting = faithful$waiting     # the waiting period
> cov(duration, waiting)         # apply the cov function
[1] 13.978
```

The covariance of the eruption duration and waiting time is 13.978. It indicates a positive linear relationship between the two variables.

Exercise 15 *Answer* :

```
> CA<-c(70,65,90,95,110,115,120,140,155,150)
> YA<-c(80,100,120,140,160,180,200,220,240,260)
> CB<-c(75,50,95,120,90,115,110,140,175,140)
> YB<-c(80,100,120,140,160,180,200,220,240,260)
> data<-data.frame(CA,YA,CB,YB)
```

1.
   - ```
     > meanCA<-mean(CA)
     > meanYA<-mean(YA)
     > cat("meanCA=",meanCA," meanYA=",meanYA," sdCA=",sd(CA)," sdYA=",sd(YA),"\n")
     meanCA= 111 meanYA= 170 sdCA= 31.42893 sdYA= 60.55301
     ```
   - ```
     > answer1<-data.frame(CA,YA,col1=(CA-meanCA)^2,col2=(YA-meanYA)^2,
     + col3=(CA-meanCA)*(YA-meanYA))
     > answer1
         CA  YA col1 col2 col3
     1   70  80 1681 8100 3690
     2   65 100 2116 4900 3220
     3   90 120  441 2500 1050
     4   95 140  256  900  480
     5  110 160    1  100   10
     6  115 180   16  100   40
     7  120 200   81  900  270
     8  140 220  841 2500 1450
     9  155 240 1936 4900 3080
     10 150 260 1521 8100 3510
     ```

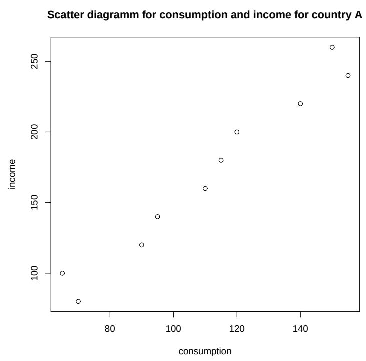
     If we use the previous datas

```
> n<-dim(answer1)
> covA=sum(answer1$col3)/(n-1)
> sdCA=sqrt(sum(answer1$col1)/(n-1))
> sdYA=sqrt(sum(answer1$col2)/(n-1))
> corA=covA/(sdCA*sdYA)
> cat("covA=",covA," corA=",corA,"\n")
```

If we use the `R` commands

```
> cov(CA,YA)
[1] 1866.667
> cor(CA,YA)
[1] 0.9808474
```

2.
```
> plot(CA,YA,main="Scatter diagramm for consumption and income for country A",
+ xlab="consumption",ylab="income")
```



3.
   - ```
     > meanCB<-mean(CB)
     > meanYB<-mean(YB)
     > cat("meanCB=",meanCB," meanYB=",meanYB," sdCB=",sd(CB)," sdYB=",sd(YB),"\n")
     meanCB= 111 meanYB= 170 sdCB= 35.88562 sdYB= 60.55301
     ```
   - ```
     > answer2<-data.frame(CB,YB,col1=(CB-meanCB)∧2,col2=(YB-meanYB)∧2,
     + col3=(CB-meanCB)∗(YB-meanYB))
     > answer2
         CB  YB col1 col2 col3
       1  75  80 1296 8100 3240
       2  50 100 3721 4900 4270
       3  95 120  256 2500  800
       4 120 140   81  900 -270
       5  90 160  441  100  210
       6 115 180   16  100   40
       7 110 200    1  900  -30
       8 140 220  841 2500 1450
       9 175 240 4096 4900 4480
      10 140 260  841 8100 2610
     ```

     By using the previous datas,

     ```
     > covB=sum(answer2$col3)/(n-1)
     > sdCB=sqrt(sum(answer2$col1)/(n-1))
     > sdYB=sqrt(sum(answer2$col2)/(n-1))
     ```

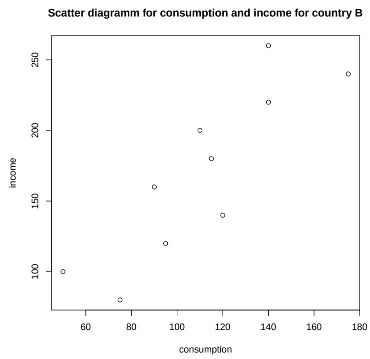
```
> corB=covB/(sdCB*sdYB)
> cat("covB=",covB," corB=",corB,"\n")
covB= 1866.667 corB= 0.8590345
```

If we use the `R` commands

```
> cov(CB,YB)
[1] 1866.667
> cor(CB,YB)
[1] 0.8590345
```

•

```
> plot(CB,YB,main="Scatter diagramm for consumption and income for country B",
+ xlab="consumption",ylab="income")
```



- Since `YA` and `YB` identical, we have `meanYA=meanYB` and `sdYA=sdYB`. We remark next that `meanCA=meanCB` (even if `CA`$\neq$`CB`) and `sdCA<sdCB` which means that the incomes of country A are less dispersed around their mean than the ones of country B.

4. We note that `covA=covB`. Since the covariance measures the degree to which two variables are linearly associated, the consumptions and incomes of country A are as much as (linearly) linked than the ones of country B.

5. (a)
```
> CAa<-CA*100
> YAa<-YA*100
> cov(CAa,YAa)
[1] 18666667
```

The covariance is multiplied by 100^2.

(b)
```
> CAb<-CAa
> YAb<-YA
> cov(CAb,YAb)
[1] 186666.67
```

The covariance is multiplied by 100.

(c)
```
> CAc<-CA
> YAc<-YA*7.85
> cov(CAc,YAc)
[1] 14653.33
```

The covariance is multiplied by 7.85.

(d)
```
> CAd<-CA+10
> YAd<-YA+10
> cov(CAd,YAd)
[1] 1866.667
```

The covariance remains unchanged.

(e)
```
> CAe<-CA
> YAe<-YA+10
> cov(CAd,YAd)
[1] 1866.667
```

The covariance remains unchanged.

All these results are well known. Indeed, let a_1, a_2, b_1 and b_2 be scalars, and let X and Y be random variables. Then the covariance of $a_1X + b_1$ and $a_2Y + b_2$ can be given by

$$\mathrm{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\mathrm{Cov}(X, Y).$$

Exercise 16

```
> xi<-c(-20,-10,0,10,20)
> yi<-c(60,40,35,20,20)
> cor(xi,yi)
[1]  -0.9534626
```

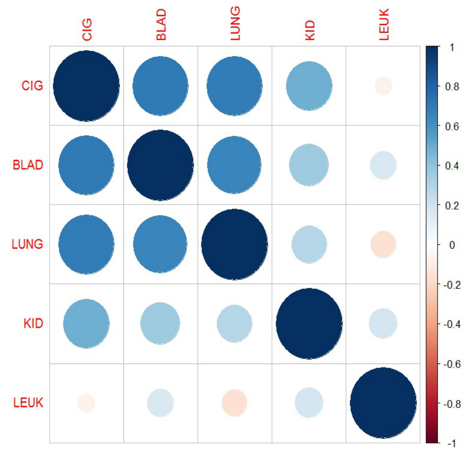
From the rule of thumb, we may conclude that X and Y are very strongly related.

Exercise 17 We first download the datas and save them under csv format ("exo17.csv" in this answer). Next, when we have selected the correct directory, we may use the commands

```
> dat<-read.csv("exo17.csv",header=T,row.names=1)
> cor(dat)
             CIG       BLAD       LUNG       KID        LEUK
  CIG   1.00000000  0.7036219  0.6974025  0.4873896  -0.06848123
 BLAD   0.70362186  1.0000000  0.6585011  0.3588140   0.16215663
 LUNG   0.69740250  0.6585011  1.0000000  0.2827431  -0.15158448
  KID   0.48738962  0.3588140  0.2827431  1.0000000   0.18871294
 LEUK  -0.06848123  0.1621566 -0.1515845  0.1887129   1.00000000
```

The higher the coefficient is in absolute values, the stronger is the relation between two variables. We may note from the previous correlation matrix that bladder and lung cancers are strongly associated with smoking. Help to `R`, we may provide "graphical" correlations.

```
> library('corrplot') # package corrplot
> corrplot(cor(dat), method = "circle") # plot matrix
```



Exercise 18

```
> cor.test(judge1,judge2,method="pearson")

Pearson's product-moment correlation

data:  judge1 and judge2
t = 2.9448, df = 8, p-value = 0.01857
alternative hypothesis:  true correlation is not equal to 0
95 percent confidence interval:
0.1677685 0.9289897
sample estimates:
cor
0.7212121
```

Consequently, there exists a strong rank correlation between judge 1 and judge 2.

Exercise 19

Those tied in the x rankings are given a value of $\frac{2+3}{2} = 2.5$ and those tied in y are allocated $\frac{6+7+8}{3} = 7$. (In general, each tie is given the mean of the places that would have been occupied if a strict order had been produced.) The table, therefore, becomes

Ranks x	1	$\boxed{2.5}$	$\boxed{2.5}$	5	4	6	7	8
Ranks y	1	3	4	2	5	$\boxed{7}$	$\boxed{7}$	$\boxed{7}$

We may apply next the usual commands in R.

```
> ranks_x<-c(1,2.5,2.5,5,4,6,7,8)
> ranks_y<-c(1,3,4,2,5,7,7,7)
> cor.test(ranks_x,ranks_y,method="pearson")

Pearson's product-moment correlation

data:  ranks_x and ranks_y
t = 3.5386, df = 6, p-value = 0.01224
alternative hypothesis:  true correlation is not equal to 0
95 percent confidence interval:
0.2794869 0.9667584
```

```
sample estimates:
cor
0.8222252
```

We may conclude that there is positive correlation between rankings.

Exercise 20 We use the commands:

```
> mat=matrix(c(91,104,235,39,73,48,18,31,161),nrow=3)
> chisq.test(mat,correct=FALSE)

Pearson's Chi-squared test

data:  mat
X-squared = 86.023, df = 4, p-value < 2.2e-16
```

Since the p-value is almost zero we reject the null hypothesis at $\alpha = 0.05$ that financial condition and education level are independent. In another words, financial condition and education level are not independent attributes and we may not reject the evidence of a relationship between the two variables.

Exercise 21 We use the commands:

```
> mat=matrix(c(18,8,7,17),nrow=2)
> chisq.test(mat,correct=FALSE)

Pearson's Chi-squared test

data:  mat
X-squared = 8.0128, df = 1, p-value = 0.004645
```

Since the p-value is 0.004645 we reject the null hypothesis at $\alpha = 0.05$ level. Thus, age and cholesterol are not independent attributes.