Frankfurt University of Applied Sciences
Prof. Dr. Christina Andersson

Solutions to R Exercises

1 Introduction

2 Manipulation of Vectors and Numbers

1. (a) `> x=c(8,9,7,5)`

   (b) `> y=c(1,2,3,4)`

   (c) `> s=x+y`

   (d)
   ```
   > s[2]
   [1] 11
   ```

2. (a) `> p=matrix(c(1,2,3,3,5,1,4,2), nrow=4, ncol=2,byrow=T)`

   (b)
   ```
   > p
        [,1] [,2]
   [1,]    1    2
   [2,]    3    3
   [3,]    5    1
   [4,]    4    2
   ```

   (c)
   ```
   > p[,2]
   [1] 2 3 1 2
   ```

   (d)
   ```
   > p[3,2]
   [1] 1
   ```

3. (a) `> x=c(5,6,7)`

   (b)
   ```
   > max(x)
   [1] 7
   ```

   (c)
   ```
   > rev(x)
   [1] 7 6 5
   ```

4. (a)
   ```
   > s=seq(6,76, by=1)
   > s
    [1]  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
   [26] 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
   [51] 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76
   ```

   (b)
   ```
   > sample(s, 5, replace = F )
   [1] 66 17 51 59 32
   ```

   (c)
   ```
   > sample(s, 5, replace = T )
   [1] 63 47 15 15 45
   ```

5. (a) `> d=seq(0,99,by=1)`

   (b) `> e=rnorm(100,3,4)`

   (c) `> f = d + e`

6.
   ```
   > x = c(2,3,4)
   > y = c(5,6,7)
   > test = data.frame(x,y)
   > test
   > test
     x y
   ```

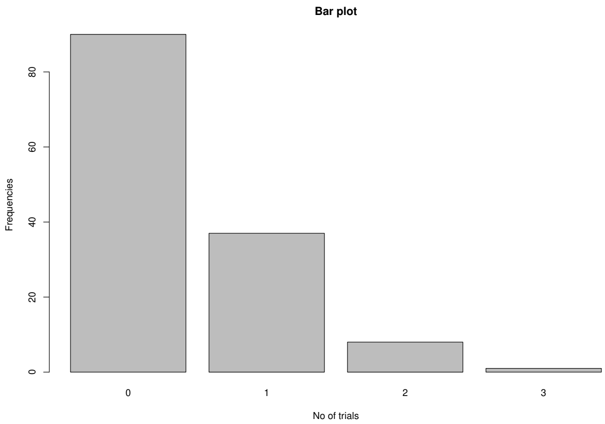
```
1 2 5
2 3 6
3 4 7
```

7. (a)
```
> library(datasets)
> nrow(airquality)
[1] 153
> ncol(airquality)
[1] 6
```
   (b)
```
>  head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

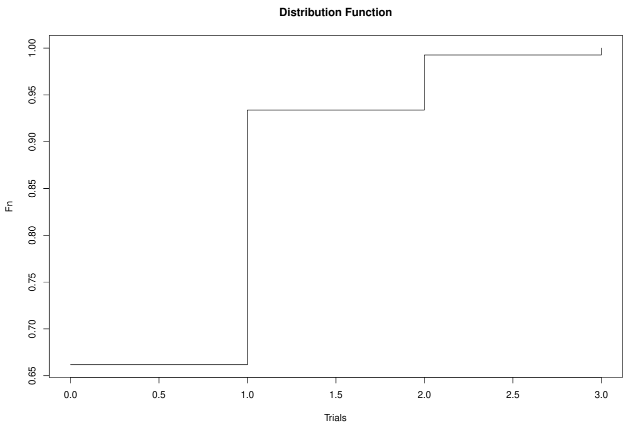
3 Tables and Graphs

1. (a)
```
exam = c(0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,
1,0,0,2,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,3,1,1,2,0,0,1,
0,0,0,0,2,0,1,0,1,0,1,0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,0,1,0,0,0,0,0,1,
0,1,1,1,2,0,0,1,0,2,1,1,0,0,0,0,0,1,0,1,1,0,0,0,2,1,1,1,0,2,0,1,1,2,
1,1,0,1,0)

abs_freq = table(exam)
> abs_freq
exam
 0  1  2  3
90 37  8  1
```
   (b)
```
> rel_freq = abs_freq/length(exam)
> rel_freq
exam
          0           1           2           3
0.661764706 0.272058824 0.058823529 0.007352941
```
   (c)
```
> barplot(abs_freq,names.arg=c("0","1","2","3"),main="Bar plot",
xlab="No of trials",ylab="Frequencies")
```



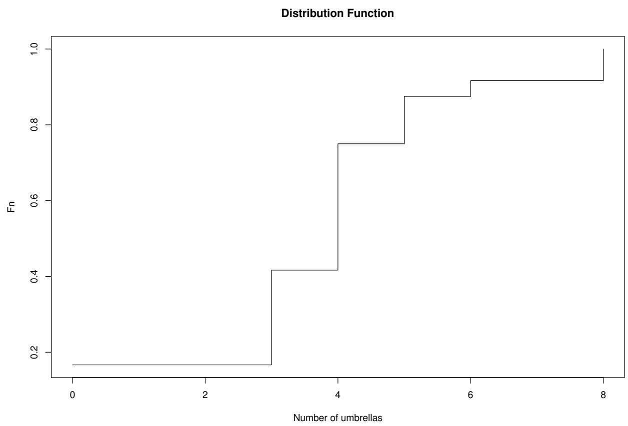
(d)
```
fn = cumsum(rel_freq)
plot(sort(unique(exam)),fn,type="n",xlab="Trials",
ylab="Fn",main="Distribution Function")
lines(sort(unique(exam)),fn,type="s")
```



2. (a)
```
> umbrella = data.frame(no_of_umbrellas=c(0,3,4,5,6,8),days=c(4,6,8,3,1,2))
> abs_freq = umbrella$days
> abs_freq
[1] 4 6 8 3 1 2
> cum_abs = cumsum(abs_freq)
> cum_abs
[1]  4 10 18 21 22 24
```

(b)
```
> rel_freq = umbrella$days/sum(abs_freq)
> rel_freq
[1] 0.16666667 0.25000000 0.33333333 0.12500000 0.04166667 0.08333333
> cum_rel <- cumsum(rel_freq)
> cum_rel
[1] 0.1666667 0.4166667 0.7500000 0.8750000 0.9166667 1.0000000
```

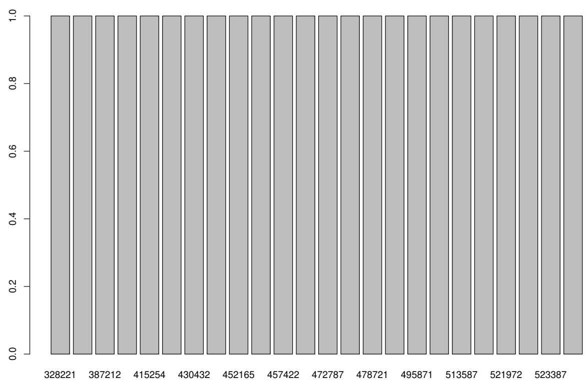
(c)
```
> fn = cumsum(rel_freq)
> plot(umbrella$no_of_umbrellas,fn,type="n",xlab="Number of umbrellas",
+ ylab="Fn",main="Distribution Function")
> lines(umbrella$no_of_umbrellas,fn,type="s")
```



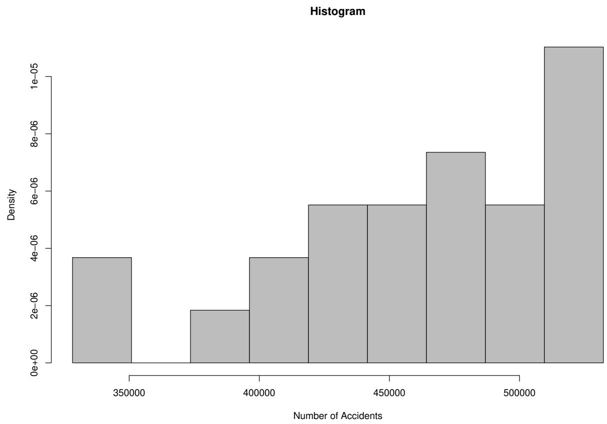
(d) > fn[5]
[1] 0.9166667
During 91.67 % of the days 6 umbrellas or less are sold.

3. (a) boots = data.frame(Year=1985:2008,No_of_Accidents=c(472787,
495871,532220,523387,499666,513587,487654,478721,521972,476544,
430432,452165,432589,456436,457422,466064,519482,343091,328221,
522169,415077,387212,415254,423731))

(b) > barplot(table(boots$No_of_Accidents))

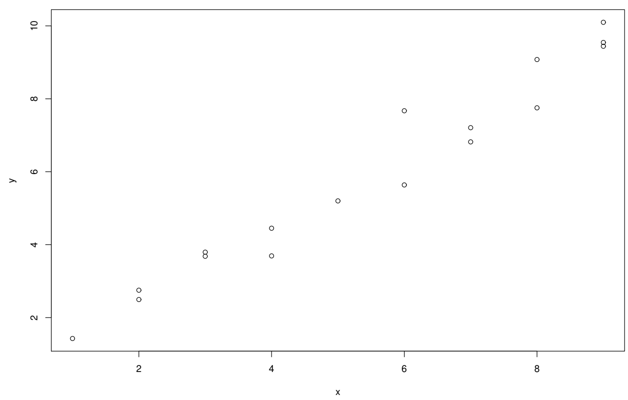


(c) hist(boots$No_of_Accidents,breaks=seq(min(boots$No_of_Accidents),
max(boots$No_of_Accidents),(max(boots$No_of_Accidents)-
min(boots$No_of_Accidents))/9),freq=F,main="Histogram",
xlab="Number of Accidents",col="grey")

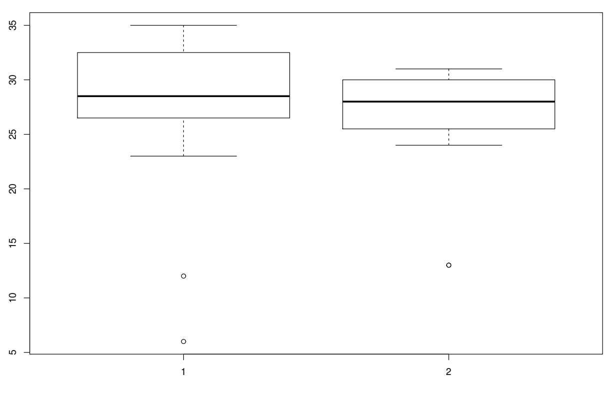


In the last class.

4.
```
> x=c(1,2,2,3,3,4,4,5,6,6,7,7,8,9,8,9,9)
> y =c(1.426865, 2.495512, 2.751945, 3.794935, 3.682121, 3.692246
4.451148, 5.200307, 5.638318, 7.672076, 6.819001, 7.208195 9.076866,
9.441328, 7.752522, 9.545205, 10.097847)
> plot(x,y)
```

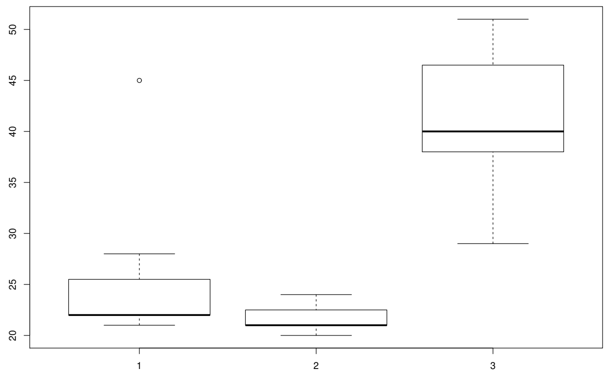


5.
```
> group1 = c(12, 23, 34, 33, 35, 33, 32, 31, 30, 29, 28, 28, 27, 27, 6, 26)
> group2 = c(13, 13, 24, 30, 31, 31, 30, 30, 31, 28, 28, 29, 26, 25, 26, 26)
> boxplot(group1, group2)
```



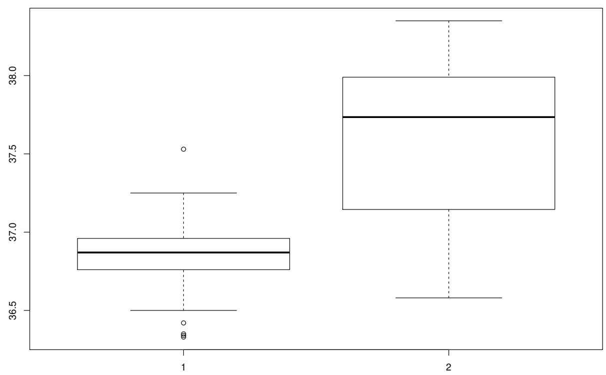
6. (a) `Gr1=c(22,22,28,23,45,21,22)`
   `Gr2=c(21,23,21,24,22,20,21)`
   `Gr3=c(48,45,51,29,38,40,38)`

   (b) `boxplot(Gr1, Gr2, Gr3)`
   Interpretation: -



7. (a) `> library(datasets)`

   (b) `> boxplot(beaver1$temp, beaver2$temp)`

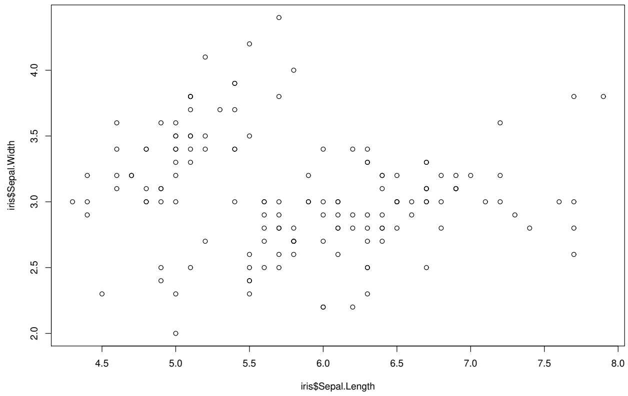


   Interpretation?

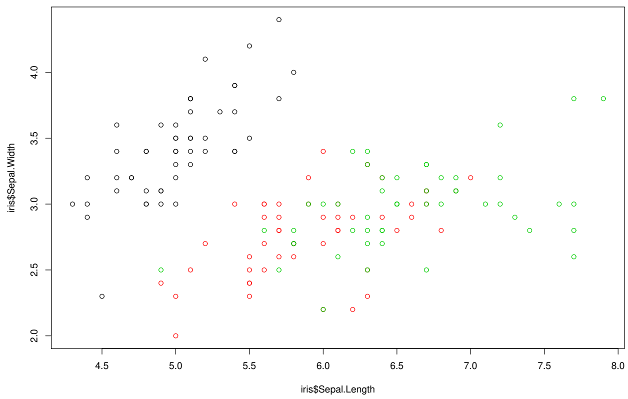
8. (a) `> library(MASS)`

   (b) `> max(iris$Sepal.Width)`
   `[1] 4.4`

   (c) `> plot(iris$Sepal.Length, iris$Sepal.Width)`



(d) > plot(iris$Sepal.Length,iris$Sepal.Width,col=iris$Species)



4 Measures of Central Tendency

1.
```
> strike= data.frame(Country=c("Aland", "Bland", "Cland", "Dland"),
Days=c(77, 45, 76, 83))
> strike
  Country Days
1   Aland   77
2   Bland   45
3   Cland   76
4   Dland   83
```

(a)
```
> median(strike$Days)
[1] 76.5
```

(b)
```
> mean(strike$Days)
[1] 70.25
```

(c) -

2.
```
> students=data.frame(name=c("Anton", "Kim", "Harald", "Inga", "Mona", "Sigrid"),
+ height=c(170,167,169,172,171,170), weight=c(70,75,120,87,88,87))
> students
    name height weight
```

```
1  Anton    170     70
2    Kim    167     75
3 Harald    169    120
4   Inga    172     87
5   Mona    171     88
6 Sigrid    170     87
```

(a)
```
> mean(students$weight)
[1] [1] 87.83333
> median(students$weight)
[1] 87
```

(b)
```
> students2=subset(students,name!="Harald")
> students2
    name height weight
1  Anton    170     70
2    Kim    167     75
4   Inga    172     87
5   Mona    171     88
6 Sigrid    170     87
> mean(students2$weight)
[1] 81.4
> median(students2$weight)
[1] 87
```

(c) -

3. `grades<-c(2,3,3,3,4,1,5,2,4,2,2,2,3,4,4,3,2,1,3,3,3,2,2)`

(a)
```
> table(grades)
grades
1 2 3 4 5
2 8 8 4 1
```

(b)
```
> mean(grades)
[1] 2.73913
> median(grades)
[1] 3
```

From the table, we see that the modes are 2 and 3.

4.
```
> pumpkins=data.frame(no_of_pumpkins=c(0,1,2,3,4,5,6,7),
days=c(10,2,3,5,4,2,4,5))
> sum(pumpkins$no_of_pumpkins*pumpkins$days)/sum(pumpkins$days)
[1] 3.085714
```

5. (a) `> library(datasets)`

(b)
```
> mean(airquality$Temp)
[1] 77.88235
```

(c)
```
> median(airquality$Solar.R)
[1] NA
```

(d)
```
> median(airquality$Solar.R, na.rm=T)
[1] 205
```

6.
```
> (90*0+4*1+6*2)/100
[1] 0.16
```
The average number of server problems per day is 0.16.

7.
```
> flats=data.frame(rooms=1:7,no_of_flats=c(56,55,35,22,12,6,2))
> sum(flats$rooms*flats$no_of_flats)/sum(flats$no_of_flats)
[1] 2.494681
```

8.
```
> mydata=data.frame(obs_values=c(50,45,25,20),frequency=c(2,4,2,2))
> prod(mydata$obs_values[1]^mydata$frequency[1],
mydata$obs_values[2]^mydata$frequency[2],
mydata$obs_values[3]^mydata$frequency[3],
mydata$obs_values[4]^mydata$frequency[4])^(1/sum(mydata$frequency))
[1] 34.74346
```

9.
```
> mydata=c(5,7,8,9,9)
> prod(mydata)^(1/length(mydata))
[1] 7.432392
```

10.
```
> mydata=c(5,7,8,9,9)
> length(mydata)/sum(1/mydata)
> [1] 7.245543
```

11.
```
> mydata = c(5,8,9,9,9,4,5,6,6,76,43,56,65,65,3,34,45)
> quantile(mydata)
  0%  25%  50%  75% 100%
   3    6    9   45   76
```

12.
```
> scores = c(43,12,11,22,23,34,34,33,34,23,33,32,11,9,45,
56,48,23,23,43,23,21,21,45,23,22,32,32,21,43,11,47)
> quantile(scores, 0.96)
  96%
47.76
```
Those students that had scores 56 and 48.

5 Measures of Spread

1. (a)
```
> scores = c(56, 87, 88, 91, 66)
> sd(scores)
[1] 15.6301
```
   (b)
```
> var(scores)
[1] 244.3
```
   (c)
```
> range(scores)
[1] 56 91
> 91-56
[1] 35
```

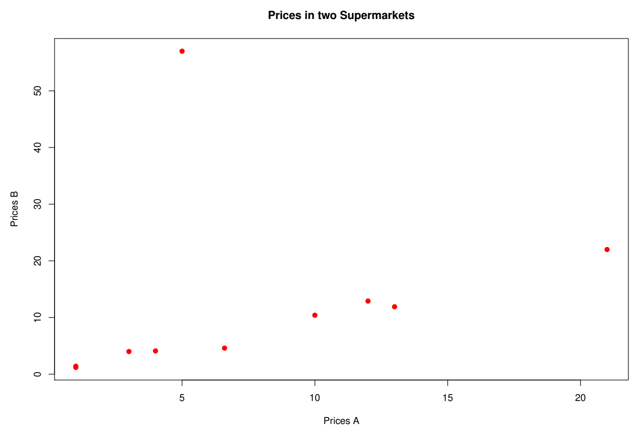
2. (a)
```
> library(datasets)
```
   (b)
```
> var(beaver1$temp)
[1] 0.03741196
> var(beaver2$temp)
[1] 0.1996203
```

(c) -

3.
```
> datawith=c(17, 23, 33, 24, 78)
> datawithout=c(17, 23, 33, 24)
> iqr_with=quantile(datawith, 0.75)-quantile(datawith, 0.25)
> iqr_with
75%
 10
> iqr_without=quantile(datawithout, 0.75)-quantile(datawithout, 0.25)
> iqr_without
 75%
4.75
> range_with=max(datawith)-min(datawith)
> range_with
[1] 61
> range_without=max(datawithout)-min(datawithout)
> range_without
[1] 16
```

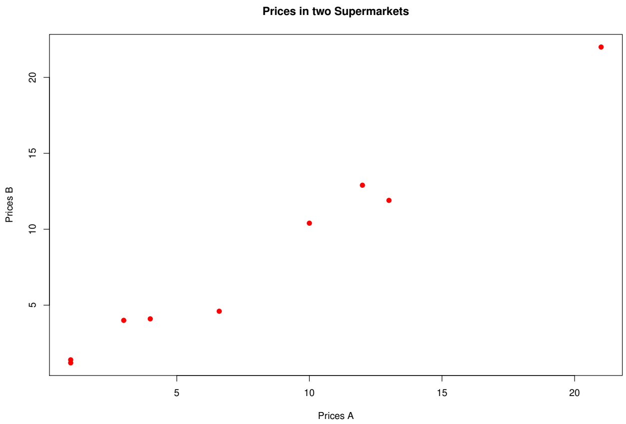
6 Correlation

1. (a)
```
> A=c(1, 5, 6.6, 4, 10, 12, 13, 21, 1, 3)
> B=c(1.2, 57, 4.6, 4.1, 10.4, 12.9, 11.9, 22, 1.4, 4)
> cor(A,B)
[1] 0.2399742
```
   (b) Scatter plot:



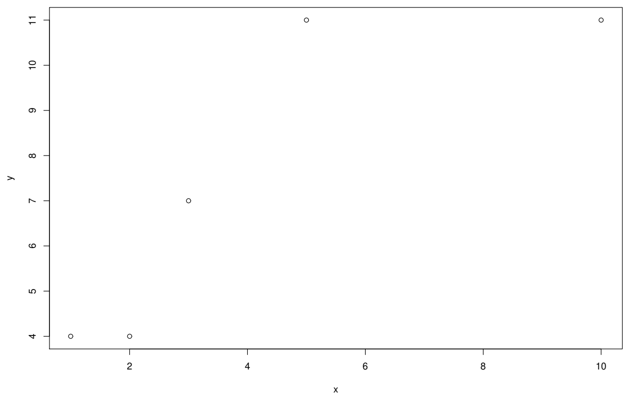

```
> plot(A,B,pch=16,cex=1.2,col="red",main="Prices in two Supermarkets",
xlab="Prices A",ylab="Prices B")
```
   (c) -

   (d)
```
> Anew=c(1, 6.6, 4, 10, 12, 13, 21, 1, 3)
> Bnew=c(1.2, 4.6, 4.1, 10.4, 12.9, 11.9, 22, 1.4, 4)
> cor(Anew,Bnew)
[1] 0.9889279
```
   Scatter plot:




```
> plot(Anew,Bnew,pch=16,cex=1.2,col="red",main="Prices in two Supermarkets",
xlab="Prices A",ylab="Prices B")
```

2. (a)
```
> y=c(4,4,7,11,11)
> x=c(1,2,3,5,10)
> plot(x,y,xlab="x",ylab="y")
```
Scatter plot:



(b)
```
> cor(x,y)
[1] 0.8521091
> cor(x,y, method="spearman")
[1] 0.9486833
```

(c)
```
> rank(x)
[1] 1 2 3 4 5
> rank(y)
[1] 1.5 1.5 3.0 4.5 4.5
```
Double values exist.

3.
```
> W=c(1,2,3,3,6,1,3,8)
> V=c(4,4,5,3,8,1,5,6)
> cor(W,V, method="spearman")
[1] 0.8013814
```

7 Regression

1.
```
> d=c(1,1,1,2,2,2,3,3,5,6,7,8)
> e=c(2,3,4,4,5,6,6,7,8,8,8,9)
```

   (a)
```
> lm(e ~ d)

Call:
lm(formula = e ~ d)

Coefficients:
(Intercept)            d
     3.0336       0.8194
```
   i.e. $e = 3.0336 + 0.8194 * d$

   (b) ...

   (c)
```
> cor(d,e)
> 0.898421
```

   (d) ...

   (e)
```
> cov(d,e)
> 4.984848
```

   (f) ...

2. (a)
```
> cor(d,e, method='spearman')
[1] -0.02683367
```

   (b)
```
> cor(d,e)
[1] 0.1627058
```

   (c) ...

   (d) ...

3. (a)
```
e=c(5,2,3,4,2,1,5,4,7,5,8,9,8,8)
j=c(5,6,3,4,1,1,1,6,7,8,8,7,8,9)

> lm(e~j)

Call:
lm(formula = e ~ j)

Coefficients:
(Intercept)            j
     1.5068       0.6744

> summary(lm(e~j))

Call:
lm(formula = e ~ j)

Residuals:
   Min     1Q Median     3Q    Max
-3.553 -1.018 -0.030  1.017  2.819
```

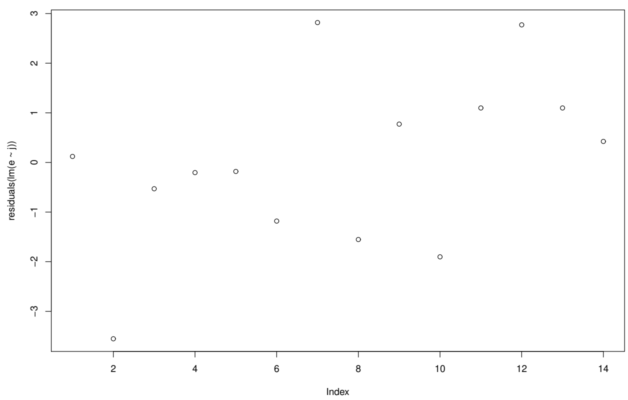
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5068     1.0511   1.434  0.17724
j             0.6744     0.1766   3.819  0.00244 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.808 on 12 degrees of freedom
Multiple R-squared:  0.5486,    Adjusted R-squared:  0.511
F-statistic: 14.58 on 1 and 12 DF,  p-value: 0.002444

> plot(residuals(lm(e~j)))
```

The estimated regression line is $e = 1.5068 + 0.6744 * j$. The p-value 0.00244 tells that the variable j is significant in the model. The value of the multiple R-squared tells that the model explains 54.86% of the variation in e.

The residual plot looks rather random.



(b) -

(c) Take the square root of 0.5486 to get the correlation.

8 Probability Distributions

1. (a)
```
> pbinom(35, 100, 0.4)
[1] 0.1794694
```
   (b) $P(X > 39) = 1 - P(X \leq 39)$
```
> 1-pbinom(39, 100, 0.4)
[1] 0.5379247
```
   (c)
```
> sum( dbinom(36:38, size = 100, prob = 0.4) )
[1] 0.2027183
```

2. (a) $P(X \leq 11)$
```
> pnorm(11, mean = 10, sd = 4)
[1] 0.5987063
```
   (b) $P(X > 13)$

```
> 1 - pnorm(13, mean = 10, sd = 4)
[1] 0.2266274
```

(c) $P(10 \leq X \leq 12)$

```
> pnorm(12, mean = 10, sd = 4) - pnorm(10, mean = 10, sd = 4)
[1] 0.1914625
```

3.
```
> qnorm(0.50, mean=0, sd=10)
[1] 0
```

4.
```
> rexp(10, 5)
 [1] 0.091088054 0.153399162 0.444140176 0.075040071 0.136310121 0.057523302
 [7] 0.008389491 0.089599280 0.046229349 0.183981225
```

5.
```
> qbinom(0.4, 100, 0.3)
[1] 29
```

6.
```
> 1 - pnorm(36, mean=35.42, sd=sqrt(16))
[1] 0.4423554
```

9 Hypothesis Tests

1. We summarize the values needed for the test statistic:

```
> xbar = 14.6
> mu0 = 15.4
> sigma = 2.5
> n = 35
> z =(xbar - mu0)/(sigma/sqrt(n))
> z
[1] -1.893146
```

Then, we compute the critical values using *qnorm*:

```
> qnorm(0.05/2)
[1] -1.959964
```

this means that the critical values are aprox. -1.96 and 1.96.

Since the value of the test statistic is $-1.96 < -1.89 < 1.96$, this means that we cannot reject the null hypothesis at significance level 5%. We cannot reject that the mean weight is 15.4 this year.

2.
```
> 2* pnorm(-1.893146)
[1] 0.05833846
```

where we used the value for $z = -1.893146$ from previous exercise. We cannot reject the null hypothesis, since the p-value is larger than 5%.

3. Hypotheses:

$$\begin{aligned} H_0\colon &\quad \mu \geq 10,000 \\ H_1\colon &\quad \mu < 10,000 \end{aligned}$$

We summarize the values needed for the test statistic:

```
> xbar = 9900
> mu0 = 10000
> sigma = 125
> n = 30
> z =(xbar - mu0)/(sigma/sqrt(n))
> z
[1] -4.38178
```

Then, we compute the critical values using *qnorm*:

```
> qnorm(0.05)
[1] -1.644854
```

This means that the critical value is aprox. -1.64. Since $-4.38178 < -1.64$, we can reject the null hypothesis at level 5%.

4. Using the value from previous exercise, a lower tail p-value is:

```
> pnorm(-4.38178)
[1] 5.885682e-06
```

which means that we can reject the null hypothesis.

5. (a) `> library(datasets)`

   (b)
   ```
   > mean(beaver1$temp)
   [1] 36.86219
   ```

   (c)
   ```
   > t.test(beaver1$temp, mu=37)
            One Sample t-test

   data:  beaver1$temp
   t = -7.6071, df = 113, p-value = 9.038e-12
   alternative hypothesis: true mean is not equal to 37
   95 percent confidence interval:
    36.82630 36.89808
   sample estimates:
   mean of x
    36.86219
   ```
   The p-value tells that we can reject the null hypothesis.

6.
```
> x = matrix(c(208, 230, 282, 241), ncol = 2)
> x
     [,1] [,2]
[1,]  208  282
[2,]  230  241
> chisq.test(x)

        Pearson's Chi-squared test with Yates' continuity correction

data:  x
```

```
X-squared = 3.6919, df = 1, p-value = 0.05468
```

Since the p-value $0.05468 > 0.05$, we cannot reject the null hypothesis. We cannot reject that the two variables can be independent.

7.
```
> library(MASS)
> freq_data<-table(Aids2$state, Aids2$status)
> freq_data

          A    D
  NSW    664 1116
  Other  107  142
  QLD     78  148
  VIC    233  355
> chisq.test(freq_data)

        Pearson's Chi-squared test

data:  freq_data
X-squared = 4.7982, df = 3, p-value = 0.1872
```

Since the p-value $0.1872 > 0.05$, we cannot reject the null hypothesis. We cannot reject that the two variables can be independent.

10 Confidence Intervals

1. (a)
```
> library(datasets)
```
(b)
```
> mean(beaver1$temp)
[1] 36.86219
```
(c)
```
> t.test(beaver1$temp, conf.level=0.99)
   One Sample t-test

data:  beaver1$temp
t = 2034.8, df = 113, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 36.81473 36.90966
sample estimates:
mean of x
 36.86219
```
The confidence interval is $[36.81473, 36.90966]$