Midterm

Sijia Liang

3/15/2018

Exercise 2.2

Write a program in R that is a function of a positive integer j and produces a vector of the form

$$1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5, ..., 1, 2, ..., j.$$

```
series <- function(j) {
  result <- c()
  for (i in 1:j) {
    result <- c(result, seq(i))
  }
  return(result)
}
series(10)
```

```
##  [1]  1  1  2  1  2  3  1  2  3  4  1  2  3  4  5  1  2  3  4  5  6  1  2
## [24]  3  4  5  6  7  1  2  3  4  5  6  7  8  1  2  3  4  5  6  7  8  9  1
## [47]  2  3  4  5  6  7  8  9 10
```

Exercise 2.5

Look at the academic score data in Table 1.2. The various scores that went into this table have been standardized across countries to adjust for cultural differences.

(a) Are the standard deviation of the different scores comparable? Consider using the `bartlett.test()` test for equality of variances. Other tests of equality of variances available in R are `var.test()`, `flinger.test()`, `ansari.test()`, and `mood.test()`. Using the `help()` file to read about these tests.

```
OECD <- read.table("/Users/sijialiang/Desktop/Supp_2/OECD PISA.txt", header=TRUE,row.names=1)
sapply(OECD, sd) # compute sd for every column, yes, they are comparable
```

```
##       Read     Access  Integrate    Reflect Continuous    Non.con
##   51.57607   54.58450   50.31968   54.43443   50.26054   56.36515
##       Math    Science
##   59.80011   56.08678
```

```
bartlett.test(OECD)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  OECD
## Bartlett's K-squared = 3.3366, df = 7, p-value = 0.8522
```

```
help("var.test")
help("ansari.test")
help("mood.test")
```

(b) Examine the correlation matrix of this data using the `cor` function. Describe what you see. See also Exercise 8.7.

```
cor(OECD)
```

```
##                 Read    Access Integrate   Reflect Continuous   Non.con
## Read       1.0000000 0.9920751 0.9960944 0.9864269  0.9985971 0.9938858
## Access     0.9920751 1.0000000 0.9883184 0.9684007  0.9895695 0.9882465
## Integrate  0.9960944 0.9883184 1.0000000 0.9707004  0.9953654 0.9863709
## Reflect    0.9864269 0.9684007 0.9707004 1.0000000  0.9849382 0.9849555
## Continuous 0.9985971 0.9895695 0.9953654 0.9849382  1.0000000 0.9877283
## Non.con    0.9938858 0.9882465 0.9863709 0.9849555  0.9877283 1.0000000
## Math       0.9478621 0.9509377 0.9585066 0.9029881  0.9432874 0.9410056
## Science    0.9818384 0.9791585 0.9854737 0.9542091  0.9793031 0.9753858
##                 Math   Science
## Read       0.9478621 0.9818384
## Access     0.9509377 0.9791585
## Integrate  0.9585066 0.9854737
## Reflect    0.9029881 0.9542091
## Continuous 0.9432874 0.9793031
## Non.con    0.9410056 0.9753858
## Math       1.0000000 0.9708641
## Science    0.9708641 1.0000000
```

```
# All of the test scores are very highly correlated with each other.
# This suggest any one test score is representative of the whole data from each country.
```

Exercise 2.8

Consider the function

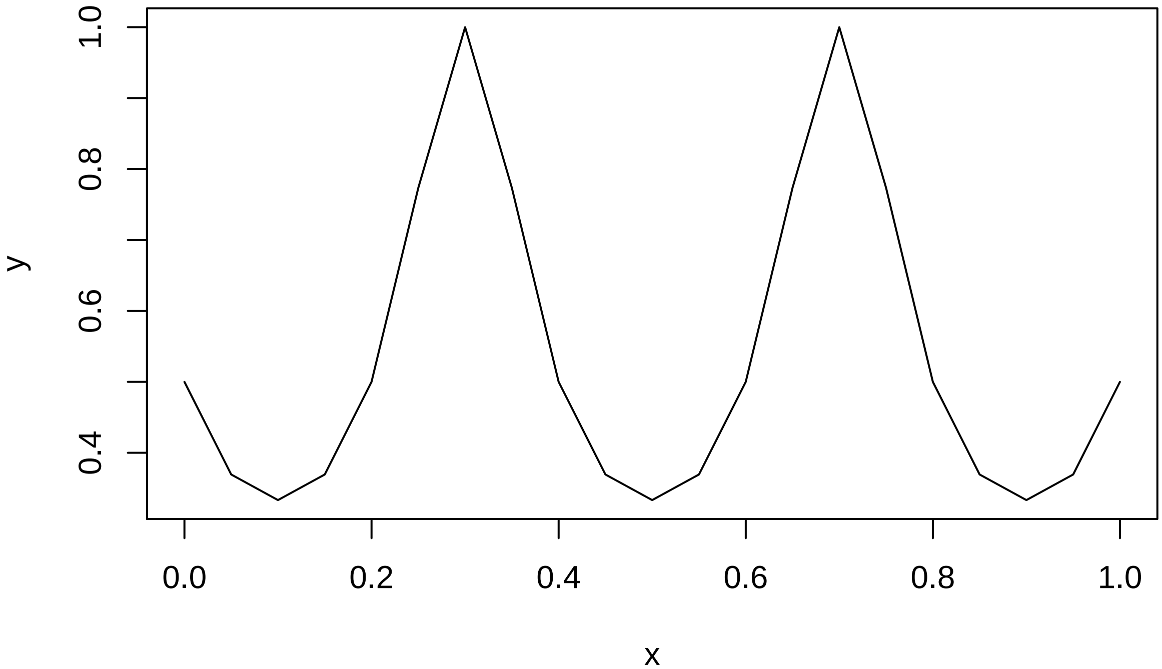
$$f(x) = \frac{1}{2 + \sin(5\pi x)}$$

for values of

$$0 \leqslant x \leqslant 1.$$

(a) Plot the function f in R. (See Sect.3.1 for an example of how to do this.)

```
x <- seq(0, 1, by=0.05)
f <- function(x) {
  return(1/(2+sin(5*pi*x)))
}

y <- f(x)
plot(x, y, type="l")
```



(b) Find the area under this curve using numerical quadrature with `integrate()`.

```
integrate(f, 0, 1)
```

```
## 0.5388603 with absolute error < 6.5e-07
```

(c) Identify maximums and minimums of this function in R using `nlm()`. Show how the results depend on the starting values.

```
#  minimum value
nlm(f, 0.9)$minimum
```

```
## [1] 0.3333333
```

```
g <- function(x) {
  return(0-f(x))
}
#maximum value
-nlm(g, 0.2)$minimum
```

```
## [1] 1
```

```
nlm(g, 0.1)
```

```
## $minimum
## [1] -1
##
## $estimate
## [1] 0.2999995
##
## $gradient
## [1] 5.329071e-09
##
## $code
## [1] 1
##
## $iterations
```

```
## [1] 8
```

```r
nlm(g, 0.2)
```

```
## $minimum
## [1] -1
##
## $estimate
## [1] 6.7
##
## $gradient
## [1] 1.847204e-08
##
## $code
## [1] 1
##
## $iterations
## [1] 6
```

```r
nlm(g, 0.9)
```

```
## $minimum
## [1] -1
##
## $estimate
## [1] 1.099999
##
## $gradient
## [1] 5.046471e-09
##
## $code
## [1] 1
##
## $iterations
## [1] 8
```

```r
# indenpent on starting value since the local max are all the same
```

Exercise 2.12

What does R do when we try to access subscripts that are out of range? Suppose we start with `> x <- 1:3`
(a) A negative subscript such as `x[-2]` will omit the second element of the vector. What does `x[-5]` yield in this example?

```r
# it does not show elements that are out of range

x <- 1:3
# omit the second element of the vector
x[-2]
```

```
## [1] 1 3
```

```r
# it remains the same on apperance, but it actually omit the 5th element of the vector
x[-5]
```

```
## [1] 1 2 3
```

(b) What does `x[0]` give us? Can you explain this outcome?

```
x[0]
```

```
## integer(0)
```

```
# gives zero, because R starts with 1, unlike Java, C++ start with zero
```

(c) Suppose we try to assign a value to an invalid element of the vector, such as in `x[7] <- 9`. What does this produce?

```
x[7] <- 9
x
```

```
## [1]  1  2  3 NA NA NA  9
```

```
# it assigned the value 9 to the 7th element of the vector
# but at the same time, returns NotAvailable(NA) for 4,5,6th element
```

(d) Is an empty subscript `x[]` different from `x` with no subscript at all? As an example, how is this `x[] <- 3` different from `x <- 3`?

```
x[]
```

```
## [1]  1  2  3 NA NA NA  9
```

```
x # its the same
```

```
## [1]  1  2  3 NA NA NA  9
```

```
x[] <- 3
x # assign 3 to the vector
```

```
## [1] 3 3 3 3 3 3 3
```

```
x <- 3
x # assign the value 3 to x
```

```
## [1] 3
```

Exercise 5.1

Consider some data such as the apartment rent values or the CD4 counts. Transform these values to more normal-looking distributions using the Box-Cox transformation (5.5) for different values of λ. Find the value of λ that provides the best fit according to the Jarque-bera test, the Kolmogorov-Smirnov test, or the Shapiro-Wilk test. Are these different λs close in value? Explain why this may (or not) be the case.

```
# Different lambdas are similar
# because all tests are testing whether data are normally distributed
# we try out all different tests and lambda values in the following
# to find the best fit

# Loading packages
#install.packages("fBasics")
library(fBasics)
```

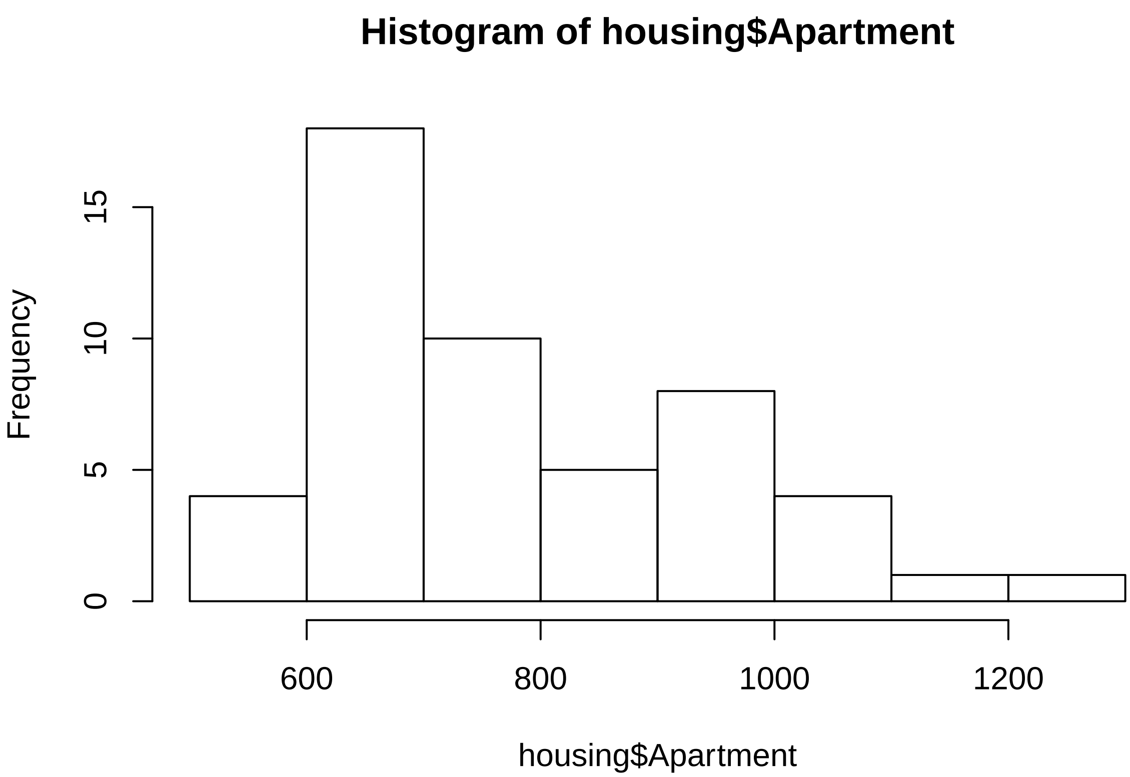
```
## Warning: package 'fBasics' was built under R version 3.4.2

## Loading required package: timeDate

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/America/New_York'
```

```
## Loading required package: timeSeries

## Warning: package 'timeSeries' was built under R version 3.4.2

##
## Attaching package: 'timeSeries'

## The following object is masked _by_ '.GlobalEnv':
##
##     series
```

```r
#install.packages("akima")
library(akima)


# Loading housing dataset and plot
housing <- read.table("/Users/sijialiang/Desktop/Supp_2/housing.txt", header=TRUE,row.names=1)
housing$Apartment
```

```
##  [1]  949  631  606  866 1135  848  970 1011  917  947  787 1298  607  690
## [15]  811  670  654  578  698  991 1074  702  706  734  657  638  631  694
## [29]  534  626  914 1068  668 1011  953  667  614  780  726  850  675  569
## [43]  660  768  784  934  797  874  704  528  636
```

```r
hist(housing$Apartment)
```

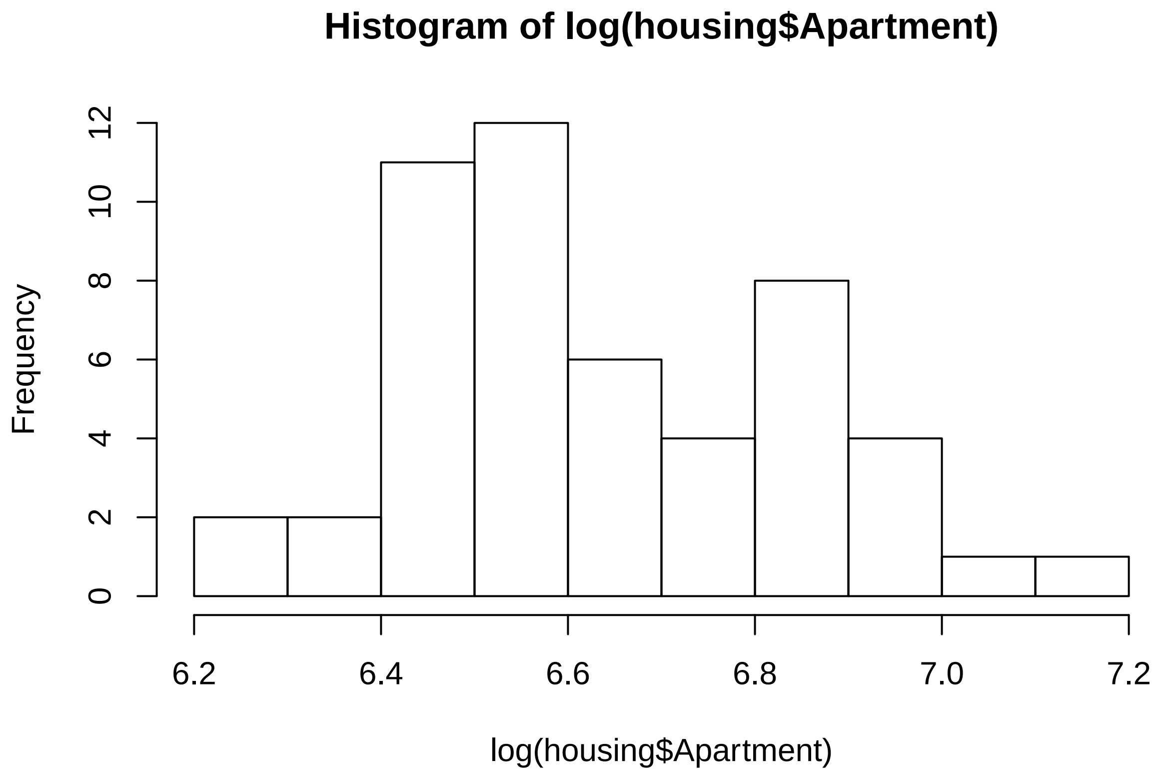



```r
log(housing$Apartment)
```

```
##  [1] 6.855409 6.447306 6.406880 6.763885 7.034388 6.742881 6.877296
##  [8] 6.918695 6.821107 6.853299 6.668228 7.168580 6.408529 6.536692
## [15] 6.698268 6.507278 6.483107 6.359574 6.548219 6.898715 6.979145
## [22] 6.553933 6.559615 6.598509 6.487684 6.458338 6.447306 6.542472
## [29] 6.280396 6.439350 6.817831 6.973543 6.504288 6.918695 6.859615
```

```
## [36] 6.502790 6.419995 6.659294 6.587550 6.745236 6.514713 6.343880
## [43] 6.492240 6.643790 6.664409 6.839476 6.680855 6.773080 6.556778
## [50] 6.269096 6.455199
```

```r
hist(log(housing$Apartment))
```




```r
# Loading CD4 dataset
require(boot)
```

```
## Loading required package: boot
```

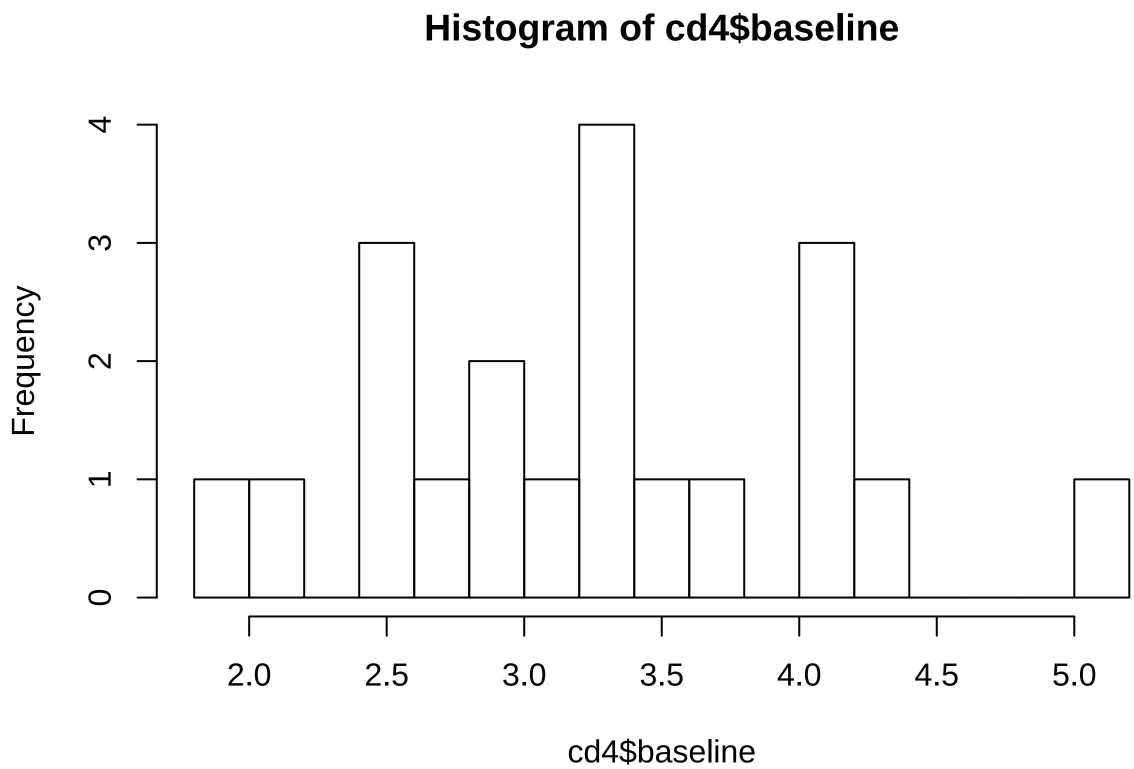
```
## Warning: package 'boot' was built under R version 3.4.1
```

```r
cd4
```

```
##    baseline oneyear
## 1      2.12    2.47
## 2      4.35    4.61
## 3      3.39    5.26
## 4      2.51    3.02
## 5      4.04    6.36
## 6      5.10    5.93
## 7      3.77    3.93
## 8      3.35    4.09
## 9      4.10    4.88
## 10     3.35    3.81
## 11     4.15    4.74
## 12     3.56    3.29
## 13     3.39    5.55
## 14     1.88    2.82
## 15     2.56    4.23
## 16     2.96    3.23
```

```
## 17      2.49     2.56
## 18      3.03     4.31
## 19      2.66     4.37
## 20      3.00     2.40
```

```r
# stem and leaf plot for small dataset
hist(cd4$baseline, breaks=20)
```




```r
stem(cd4$baseline)
```

```
##
##   The decimal point is at the |
##
##   1 | 9
##   2 | 15567
##   3 | 000444468
##   4 | 0124
##   5 | 1
```

```r
# assuming statistical signifiance = 5%, 0.05

# jb Test for housing$Apartment
jbTest(housing$Apartment,
       title="Original apartment rents")
```

```
## Warning in interpp.old(x, y, z, xo, yo, ncp = 0, extrap = FALSE, duplicate
## = duplicate, : interpp.old() is deprecated, future versions will only
## provide interpp()

## Warning in interpp.old(x, y, z, xo, yo, ncp = 0, extrap = FALSE, duplicate
## = duplicate, : interpp.old() is deprecated, future versions will only
## provide interpp()
```

```
##
## Title:
##  Original apartment rents
##
## Test Results:
##   PARAMETER:
##     Sample Size: 51
##   STATISTIC:
##     LM: 5.499
##     ALM: 6.3
##   P VALUE:
##     LM p-value: 0.043
##     ALM p-value: 0.053
##     Asymptotic: 0.064
##
## Description:
##  Sun Mar 18 17:11:20 2018 by user:
```

```
# p value = 0.043 reject null hypotheis, housing$Apartment is not normal distributed,
# therefore, perform box-cox transformation

# jb Test for cd4
jbTest(cd4$baseline)
```

```
## Warning in interpp.old(x, y, z, xo, yo, ncp = 0, extrap = FALSE, duplicate
## = duplicate, : interpp.old() is deprecated, future versions will only
## provide interpp()

## Warning in interpp.old(x, y, z, xo, yo, ncp = 0, extrap = FALSE, duplicate
## = duplicate, : interpp.old() is deprecated, future versions will only
## provide interpp()
```

```
##
## Title:
##  Jarque - Bera Normality Test
##
## Test Results:
##   PARAMETER:
##     Sample Size: 20
##   STATISTIC:
##     LM: 0.389
##     ALM: 0.377
##   P VALUE:
##     LM p-value: 0.792
##     ALM p-value: 0.814
##     Asymptotic: 0.823
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```
# p value 0.79 can't reject null hypothesis


# One-sample Kolmgorov-Smirnov test for housing$Apartment
z1 <- housing$Apartment
z1 <- (z1-mean(z1)) / sd(z1)
```

```r
ksnormTest(z1)# two-sided = 0.1484 =p-value > 0.05 = can't reject null hypothesis
```

```
## Warning in ks.test(x, "pnorm", alternative = "two.sided"): ties should not
## be present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "less"): ties should not be
## present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "greater"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.1597
##   P VALUE:
##     Alternative Two-Sided: 0.1484
##     Alternative      Less: 0.5633
##     Alternative   Greater: 0.07421
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```r
#### this is the best since it has the largest p value among all other tests

z2 <- -(housing$Apartment)^(-1/2)
z2 <- (z2-mean(z2)) / sd(z2)
ksnormTest(z2) # two-sided = 0.4315 > 0.05 = can't reject
```

```
## Warning in ks.test(x, "pnorm", alternative = "two.sided"): ties should not
## be present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "less"): ties should not be
## present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "greater"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.1222
##   P VALUE:
##     Alternative Two-Sided: 0.4315
##     Alternative      Less: 0.4789
##     Alternative   Greater: 0.218
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```r
z3 <- log(housing$Apartment)
z3 <- (z3-mean(z3)) / sd(z3)
```

```r
ksnormTest(z3)# two-sided =  0.2976> 0.05 = can't reject
```

```
## Warning in ks.test(x, "pnorm", alternative = "two.sided"): ties should not
## be present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "less"): ties should not be
## present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "greater"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.1366
##   P VALUE:
##     Alternative Two-Sided: 0.2976
##     Alternative      Less: 0.5417
##     Alternative   Greater: 0.1493
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```r
z4 <- (housing$Apartment)^(1/2)
z4 <- (z4-mean(z4)) / sd(z4)
ksnormTest(z4)# two-sided = 0.2068 > 0.05 = can't reject
```

```
## Warning in ks.test(x, "pnorm", alternative = "two.sided"): ties should not
## be present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "less"): ties should not be
## present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "greater"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.1491
##   P VALUE:
##     Alternative Two-Sided: 0.2068
##     Alternative      Less: 0.6269
##     Alternative   Greater: 0.1035
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```r
# One-sample Kolmgorov-Smirnov test for cd4
cdks <- cd4$baseline
cdks <- (cdks-mean(cdks)) / sd(cdks)
ksnormTest(cdks)# two-sided=0.9884> 0.05 = not reject null hypothesis
```

```
## Warning in ks.test(x, "pnorm", alternative = "two.sided"): ties should not
## be present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "less"): ties should not be
## present for the Kolmogorov-Smirnov test

## Warning in ks.test(x, "pnorm", alternative = "greater"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## Title:
##  One-sample Kolmogorov-Smirnov test
##
## Test Results:
##   STATISTIC:
##     D: 0.0999
##   P VALUE:
##     Alternative Two-Sided: 0.9884
##     Alternative      Less: 0.7717
##     Alternative   Greater: 0.6707
##
## Description:
##  Sun Mar 18 17:11:52 2018 by user:
```

```
##### cd4 data, 0.9884 best lambda among others




# Shapiro test for housing$Apartment
shapiro.test(housing$Apartment) #0.007472 < 0.05 = reject, not normal, cox-box transformation
```

```
##
##  Shapiro-Wilk normality test
##
## data:  housing$Apartment
## W = 0.93455, p-value = 0.007472
```

```
shapiro.test(-housing$Apartment^(-1/2)) # lambda=-1/2, p=0.3136>0.05, (-1/2) is better than take log
```

```
##
##  Shapiro-Wilk normality test
##
## data:  -housing$Apartment^(-1/2)
## W = 0.97372, p-value = 0.3136
```

```
shapiro.test(log(housing$Apartment)) # when lambda=0, take log, p= 0.1481 >0.05, obey normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(housing$Apartment)
## W = 0.96586, p-value = 0.1481
```

```
shapiro.test((housing$Apartment)^(1/2)) # when lambda=1/2 >0, p=0.0414 < 0.05, failed
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  (housing$Apartment)^(1/2)
## W = 0.95283, p-value = 0.0414
```

```
# Shapiro test for cd4$baseline
shapiro.test(cd4$baseline) # p=0.9434 >0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cd4$baseline
## W = 0.98075, p-value = 0.9434
```

Question 6

Let X have covariance matrix

$$\sum = \begin{pmatrix} 25 & -2 & 4 \\ -2 & 4 & 1 \\ 4 & 1 & 9 \end{pmatrix}$$

(a) Determine R (correlation matrix) and (diagonal matrix with standard deviations on the diagonals)

```
V_half <- diag(c(5,2,3),3,3)
R <- matrix(c(1,-1/5.0,4/15.,-1/5.,1,1/6.,4/15., 1/6.,1 ), nrow=3)
R
```

```
##            [,1]       [,2]      [,3]
## [1,]  1.0000000 -0.2000000 0.2666667
## [2,] -0.2000000  1.0000000 0.1666667
## [3,]  0.2666667  0.1666667 1.0000000
```

(b) Multiply your matrices to check the relation

```
V_half %*% R %*% V_half
```

```
##      [,1] [,2] [,3]
## [1,]   25   -2    4
## [2,]   -2    4    1
## [3,]    4    1    9
```

Question 7

Using the vectors and ,verify the extended Cauchy-Schwarz inequality if

$$B = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix}$$

```
b <- c(-4,3)
b
```

```
## [1] -4  3
```

```
d <- c(1,1)
d
```

```
## [1] 1 1
```

```r
B <- matrix(c(2,-2,-2,5), nrow=2)
B
```

```
##      [,1] [,2]
## [1,]    2   -2
## [2,]   -2    5
```

```r
b%*%d
```

```
##      [,1]
## [1,]   -1
```

```r
b %*% B %*% b
```

```
##      [,1]
## [1,]  125
```

```r
d %*% solve(B) %*% d
```

```
##          [,1]
## [1,] 1.833333
```

```r
(b%*%d)^2 <= (b %*% B %*% b)*(d %*% solve(B) %*% d)
```

```
##      [,1]
## [1,] TRUE
```

Question 8

The Table data show the age (x1, in years) and selling prices (x2, in thousands of dollars) of 10 used cars.

(a) Calculate the squared statistical distances.

```r
X <- matrix(c(1,2,3,3,4,5,6,8,9,10,18.95,19.00,17.95,15.54,14.00,12.95,8.94,7.49,6.00,3.99), nrow=10)
Sigma <- cov(X)
Sigma
```

```
##            [,1]      [,2]
## [1,]   9.433333 -16.76678
## [2,] -16.766778  30.85437
```

```r
X_bar <- colMeans(X)
X_bar
```

```
## [1]  5.100 12.481
```

```r
X_demean <- X - matrix(rep(X_bar,each=10),nrow=10)
X_demean
```

```
##       [,1]   [,2]
##  [1,] -4.1  6.469
##  [2,] -3.1  6.519
##  [3,] -2.1  5.469
##  [4,] -2.1  3.059
##  [5,] -1.1  1.519
##  [6,] -0.1  0.469
##  [7,]  0.9 -3.541
##  [8,]  2.9 -4.991
##  [9,]  3.9 -6.481
```

```
## [10,]  4.9 -8.491
```

```r
diag(X_demean %*% solve(Sigma) %*% t(X_demean))
```

```
##  [1] 2.41784482 1.98556553 3.33064161 0.89824572 0.30888467 0.08161164
##  [7] 3.66448769 0.91688749 1.80537393 2.59045691
```

```r
qchisq(.5, df=2)
```

```
## [1] 1.386294
```

(b) Using the distances in part a, determine the proportion of the observations falling within the estimated 50% probability contour of a bivariate normal distribution.

```r
# TRUEs are the data falling in the 50% prob contour
diag(X_demean %*% solve(Sigma) %*% t(X_demean)) <= qchisq(.5, df=2)
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
```