MAT 5030
Chapter 4:
Descriptive Statistics and Graphics

Kazuhiko Shinki

Wayne State University

Tips

A list of datasets included by default in R package:

http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/00Index.html

Summary Statistics for a Single Group

The 'summary' function summarize a data frame.

```
> Loan
   Inst Rate YTM Amount
1    BA  5.6  15 100000
2  Citi  4.7  15 150000
3  Citi  7.5  30 310000
4    BA  5.1  30 510000
5 Chase  4.2   7  90000
6    BA  2.9   5 190000
7 Chase  6.6  30 450000
> summary(Loan)
    Inst         Rate            YTM            Amount
 BA   :3   Min.   :2.900   Min.   : 5.00   Min.   : 90000
 Chase:2   1st Qu.:4.450   1st Qu.:11.00   1st Qu.:125000
 Citi :2   Median :5.100   Median :15.00   Median :190000
           Mean   :5.229   Mean   :18.86   Mean   :257143
           3rd Qu.:6.100   3rd Qu.:30.00   3rd Qu.:380000
           Max.   :7.500   Max.   :30.00   Max.   :510000
```

Graphical Display of Distributions

Histogram

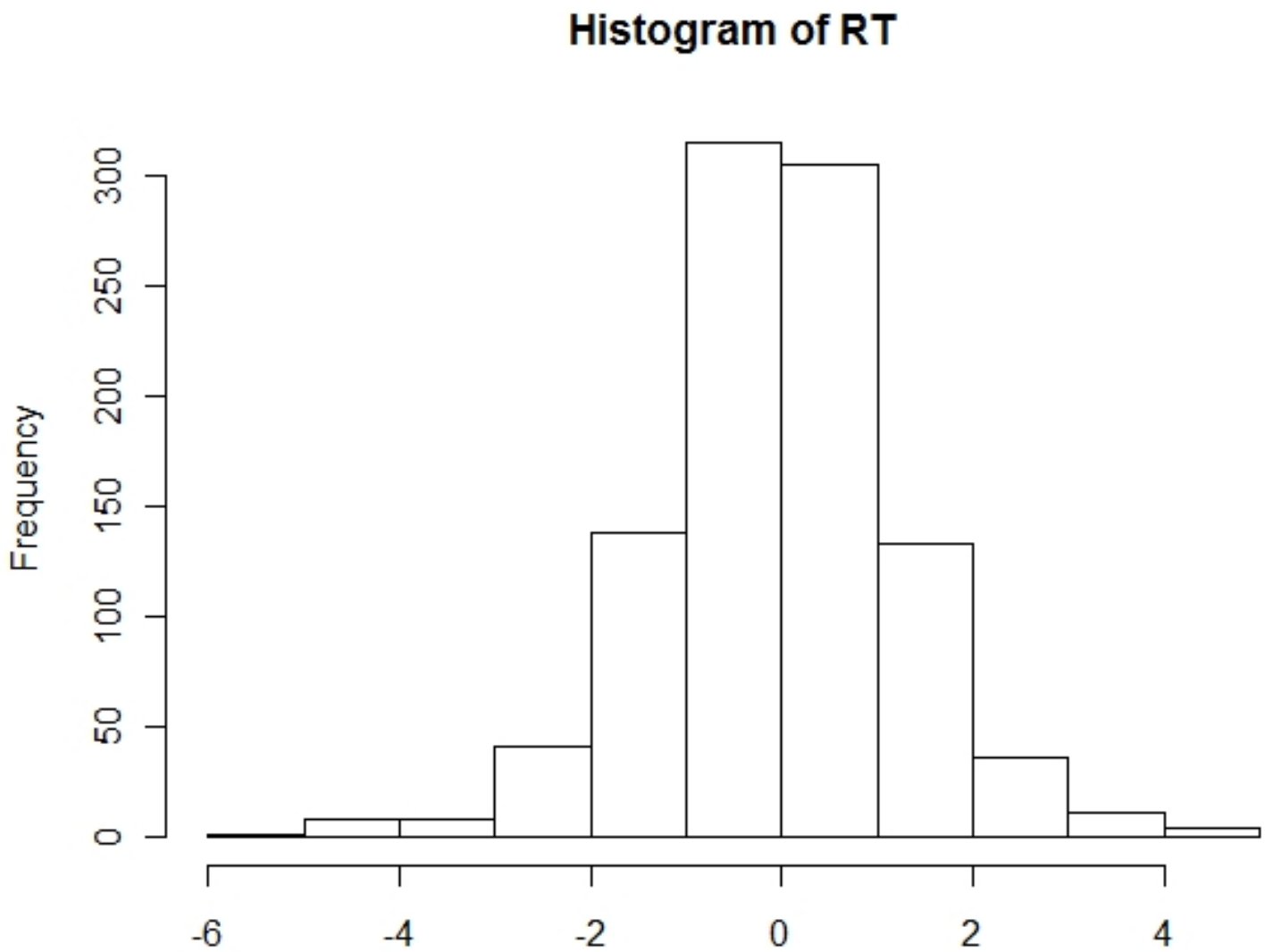
- Use the 'breaks' option to customize break points of a histogram.
- Use the 'freq=F' option to make a density histogram. With this, a histogram can be overlaid with density curves.

Graphical Display of Distributions

```
RT <- rt(1000,df=5)

## Figure 1
hist(RT)

## Figure 2
B <- seq(floor(min(RT)), ceiling(max(RT)), by =0.5)
  # break points
hist(RT, breaks=B)

## Figure 3
hist(RT, breaks=B,
ylim=c(0,0.5), freq=F, xlab="",ylab="")
par(new=T)
X <- 0.02*(-100:100) # x-coordinates of density curve
plot(X, dt(X,df=5),
xlim = range(B), ylim=c(0,0.5), type="l",
xlab="",ylab="Relative Frequency")
par(new=F)
```

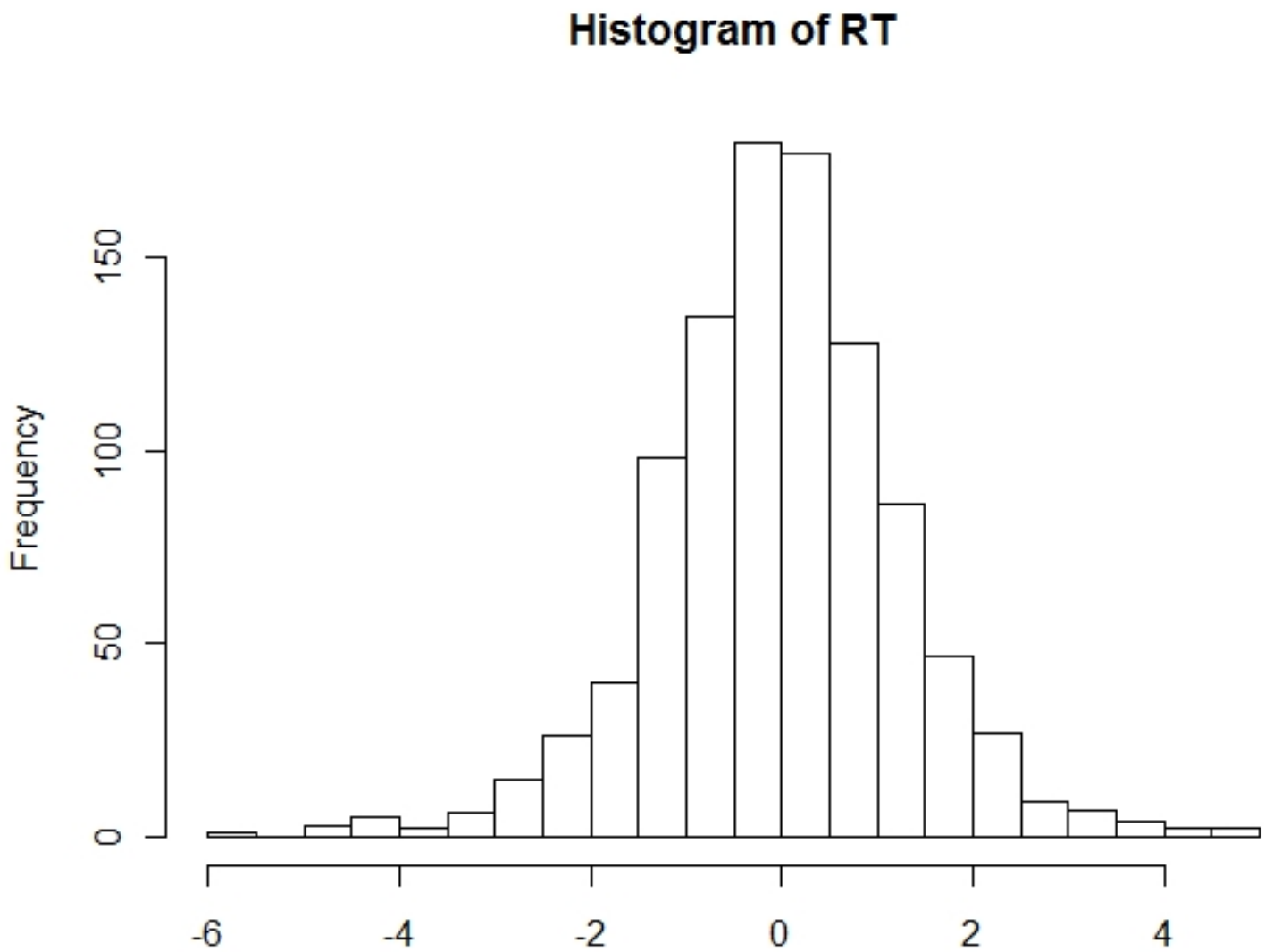
Graphical Display of Distributions

Figure 1



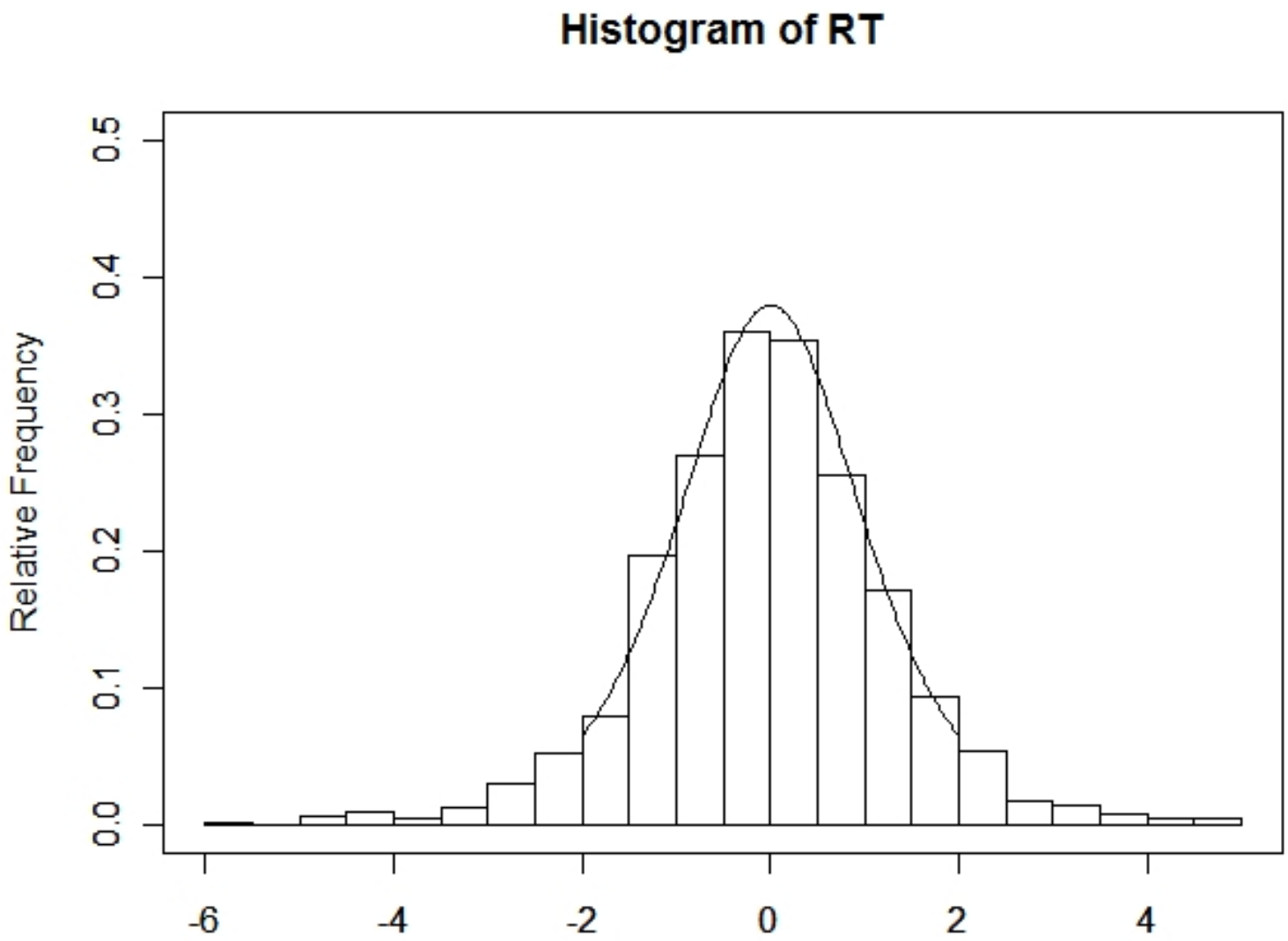
Graphical Display of Distributions

Figure 2



Graphical Display of Distributions

Figure 3



Graphical Display of Distributions

Empirical Cumulative Distribution function

Let X be a random variable. The cumulative distribution function (CDF) F is:

$$F(x) = P(X \leq x).$$

When $x_1, \cdots, x_n$ are observed, the ***empirical CDF*** $\hat{F}$ is defined by:

$$\hat{F}(x) = \frac{\text{Number of observations with } (x_i \leq x)}{n}.$$

The empirical CDF is a right-continuous function which is flat except for the points $x_1, \cdots, x_n$.

Graphical Display of Distributions

Example 1:

Suppose we have 4 observations $\{-1, 0.1, 0.5.1.3\}$,

$$\begin{aligned}
\hat{F}(X) &= 0 \quad (if \ \ x < -1) \\
\hat{F}(X) &= 1/4 \quad (if \ \ -1 \leq x < 0.1) \\
\hat{F}(X) &= 2/4 \quad (if \ \ 0.1 \leq x < 0.5) \\
\hat{F}(X) &= 3/4 \quad (if \ \ 0.5 \leq x < 1.3) \\
\hat{F}(X) &= 1 \quad (if \ \ 1.3 \leq x)
\end{aligned}$$

Graphical Display of Distributions

Example 2:

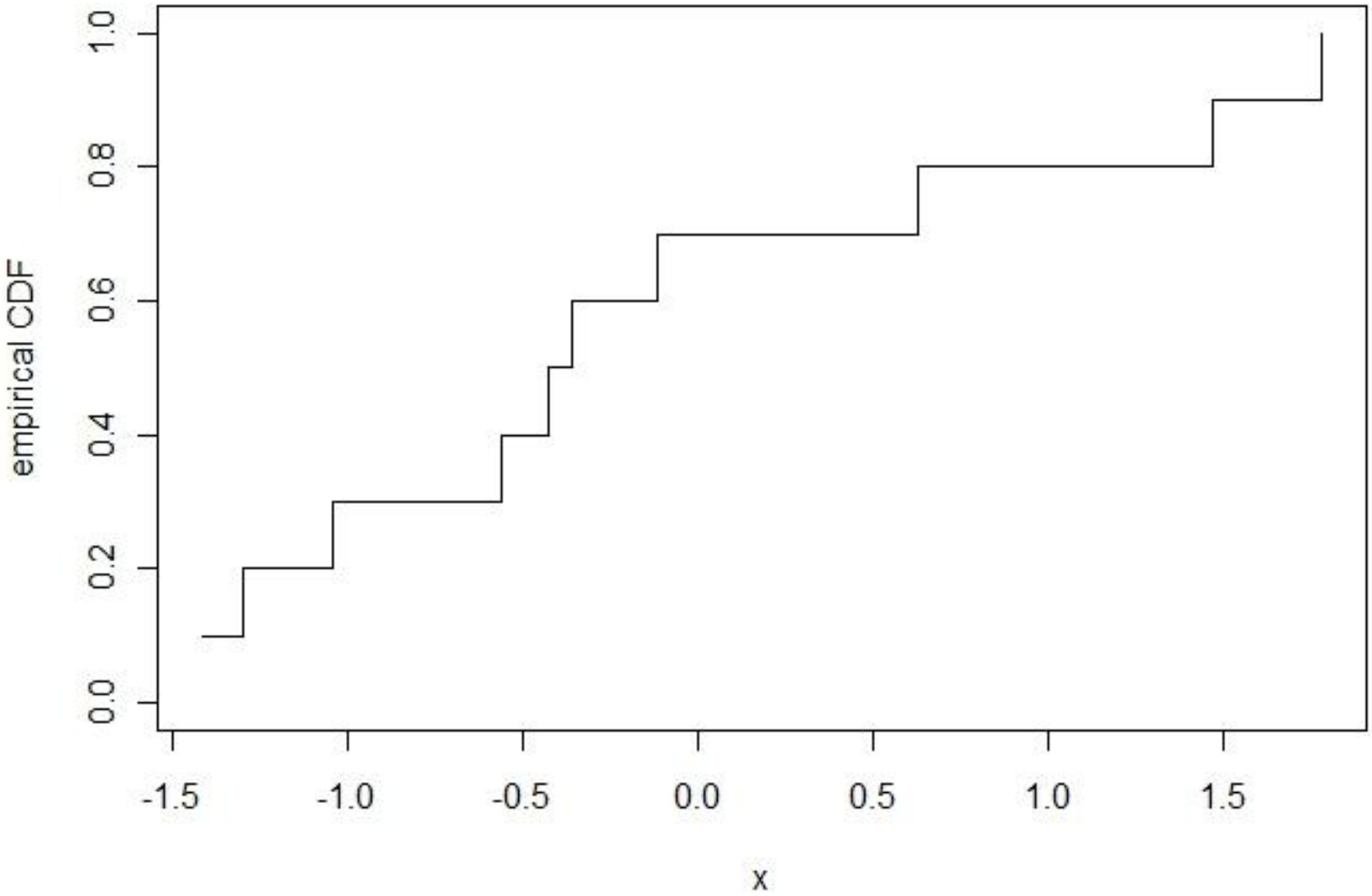
Generate 10 random numbers which follows standard normal, and draw an empirical CDF.

```
X <- sort(rnorm(10)) # 10 r.v.s, increasing order
n <- length(X)

# Figure 1
plot(X, (1:n)/n, type="s", ylim=c(0,1),xlab="x", ylab= "empirical CDF")
```

Note: type=”s” generates a step function.

Graphical Display of Distributions

Figure 1



Graphical Display of Distributions

$\hat{F}(x)$ must be 0 when x is very small, 1 when x is very large.

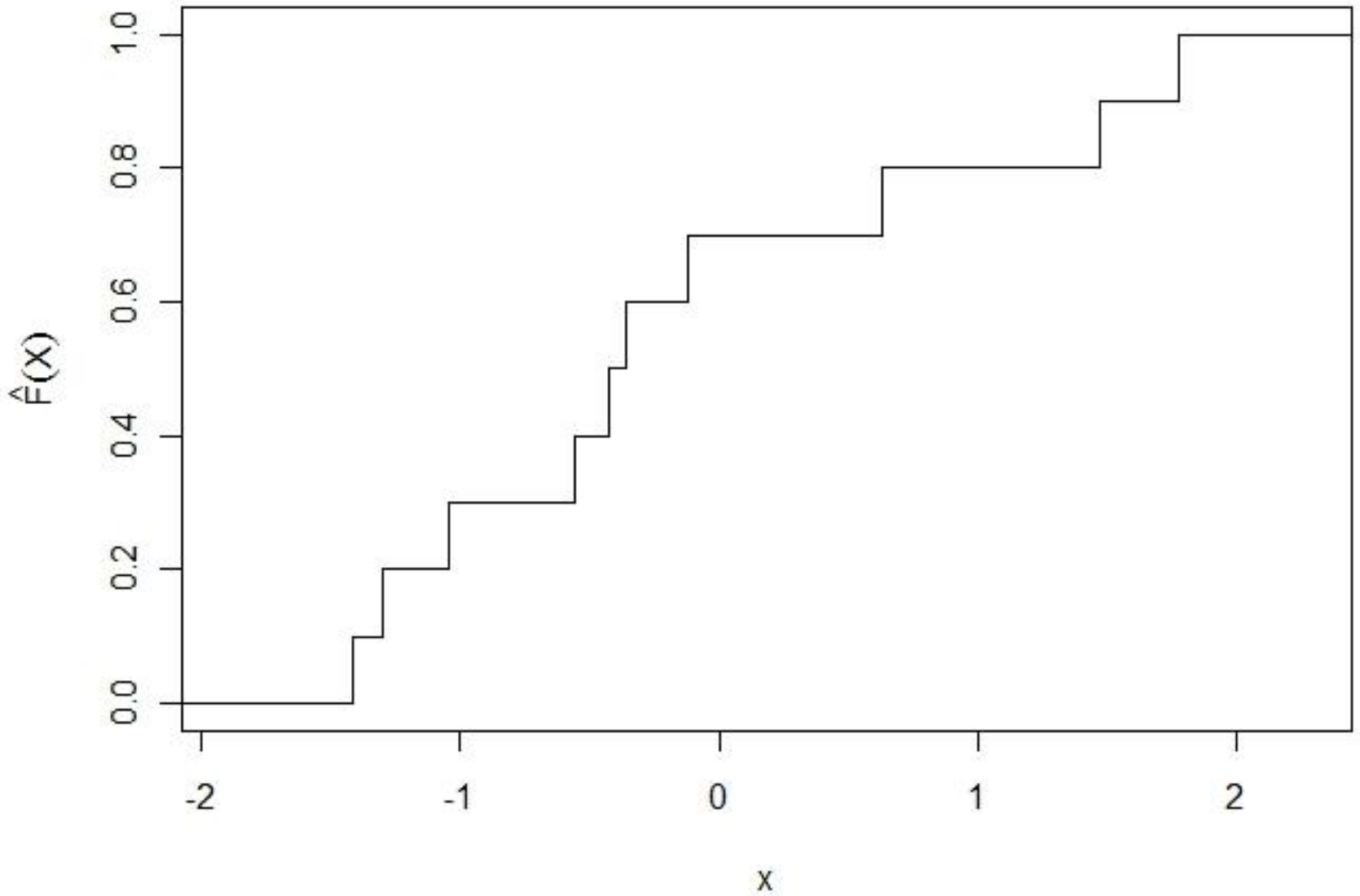
```
# Figure 2
X1 <- c(min(X)-1, sort(X),max(X)+1)
Oldpar <- par()  # store current graphic parameters as 'Oldpar'

par(mai= c(1.02, 1, 0.82, 0.42))
   # set margin size (bottom, left, top, right)
plot(X1, 0:(n+1)/n,
type="s", ylim=c(0,1),xlim=c(min(X)-0.5,max(X)+0.5),
xlab="x", ylab=expression(paste(hat(F), (X)))
)

par(Oldpar) # return to the old parameters
```

Note:

- 'par()' represents all graphic parameters. See 'help(par)' for details.
- 'expression' and 'paste' functions are used for mathematical annotation. see 'help(plotmath)' for details.

Graphical Display of Distributions

Figure 2



Graphical Display of Distributions

Q-Q Plot

Let $x_1, \cdots, x_n$ be ordered observations, and F be the hypothetical CDF for the observations. The plot $(F^{-1}(\frac{i}{n+1}), x_i)$ $(i = 1, \cdots, n)$ is called a ***Q-Q plot*** . (If F is a normal CDF, we say a normal Q-Q plot.)

If the observations come from the hypothetical distribution, approximately $F(x_i) = \frac{i}{n+1}$ (or equivalently $x_i = F^{-1}(\frac{i}{n+1})$). Hence, the Q-Q plot is roughly on the straight line $y = x$. So the Q-Q plot is used to see if the observations fit well to a hypothetical (e.g., normal) distribution.

Graphical Display of Distributions

Example 1:
To check if $\{-1, 0.1, 0.5.1.3\}$ come from standard normal, we have to check if

$$\left(F^{-1}\left(\frac{1}{5}\right), -1\right), \quad \left(F^{-1}\left(\frac{2}{5}\right), 0.1\right), \quad \left(F^{-1}\left(\frac{3}{5}\right), 0.5\right), \quad \left(F^{-1}\left(\frac{4}{5}\right), 1.3\right) \tag{1}$$

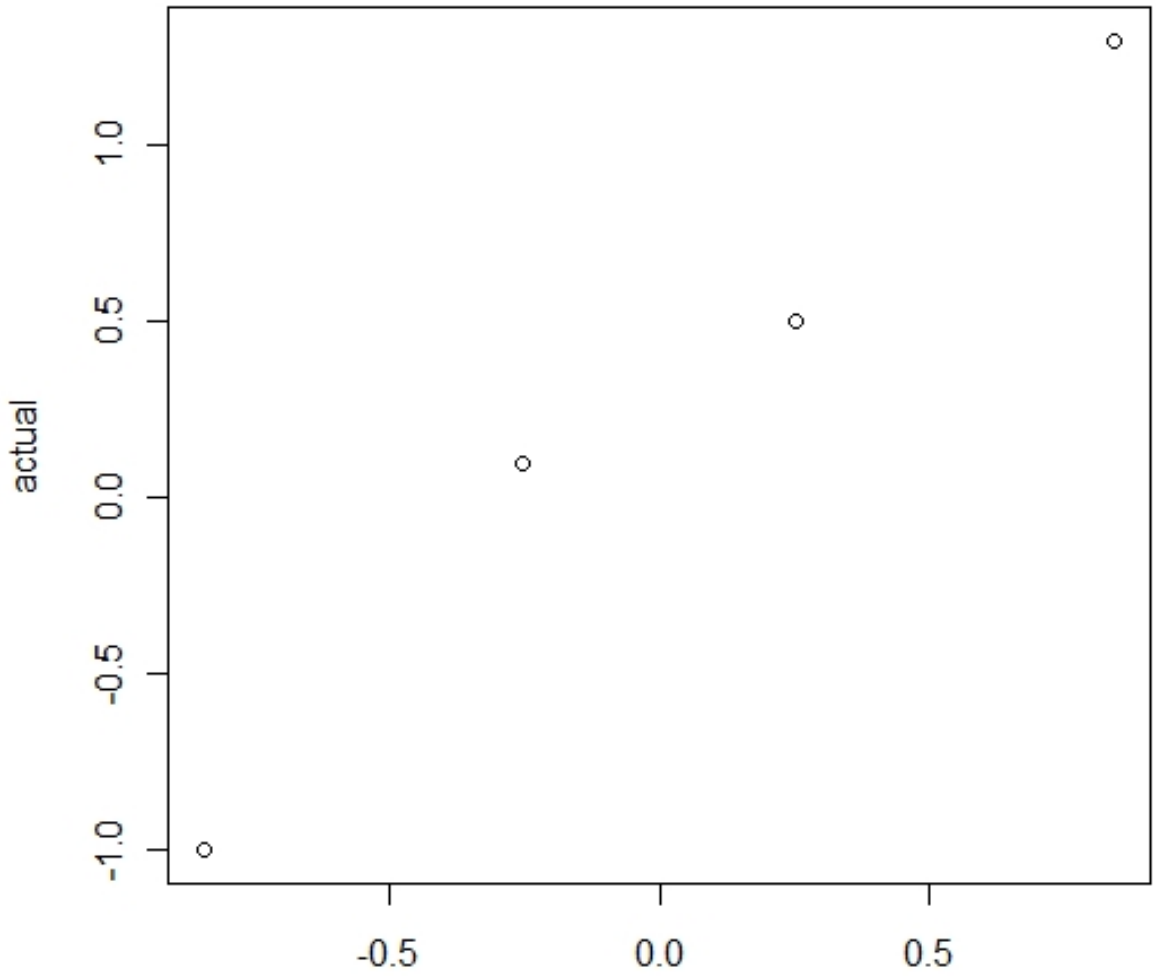
are close to $y = x$.

Code for Example 1:

```
Y <- c(-1, 0.1, 0.5, 1.3)
X <- qnorm((1:4)/5)
plot(X,Y,xlab="theoretical",ylab="actual")
```

Graphical Display of Distributions

Normal Q-Q plot for 4 observations



Graphical Display of Distributions

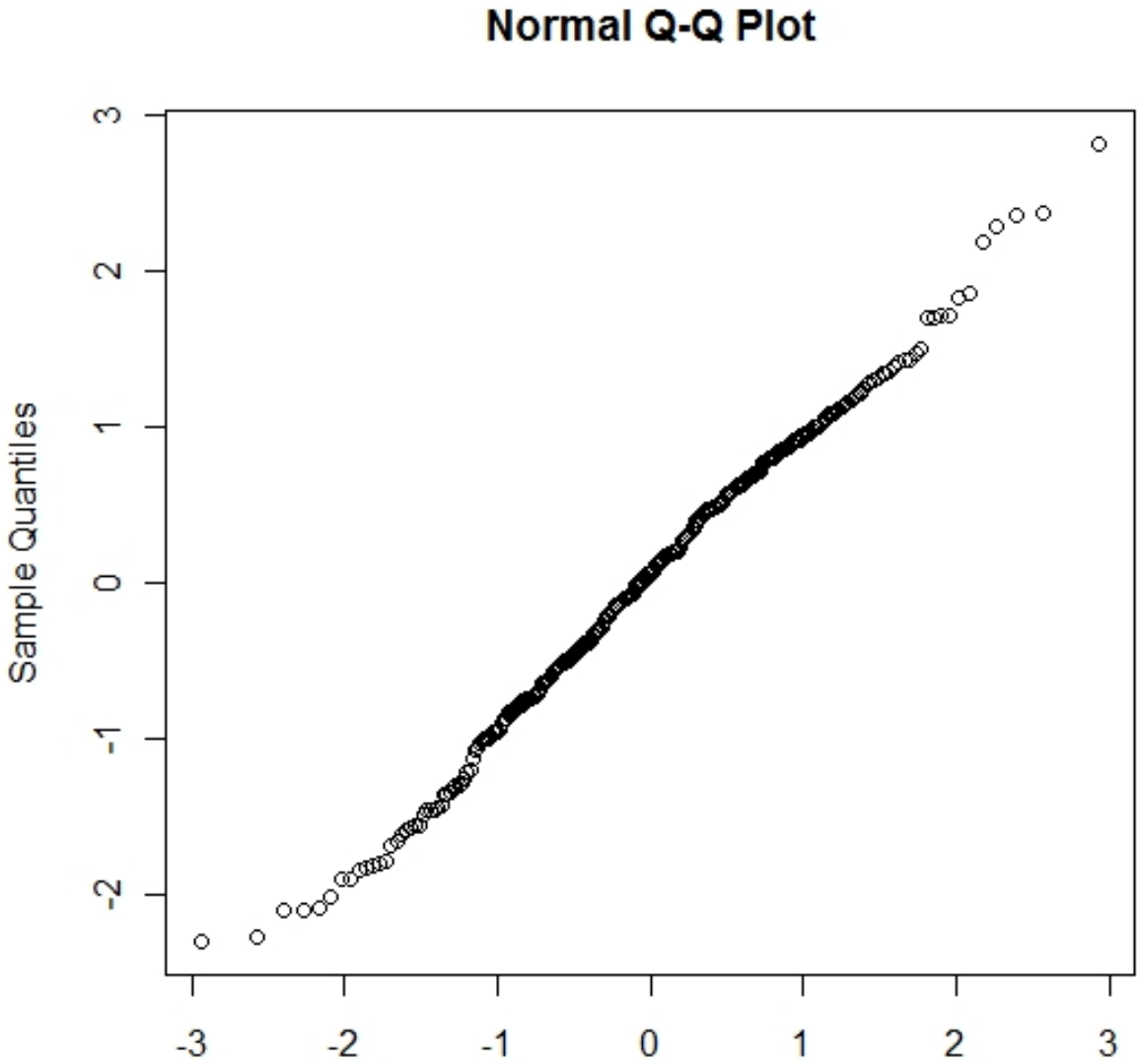
Example 2:
Generate 300 random numbers from a standard normal distribution, then create a normal Q-Q plot.

```
Y <- rnorm(300)
> qqnorm(Y)
```

Since the observations are actually from a standard normal distribution, points should be approximately on the line $y = x$.

Graphical Display of Distributions

normal Q-Q plot for normal observations



Graphical Display of Distributions

Example 2:
Generate 300 random numbers from a normal distribution $N(10, 5^2)$, then create a normal Q-Q plot.

```
Y <- rnorm(300, mean=10, sd=5)
qqnorm(Y)
```

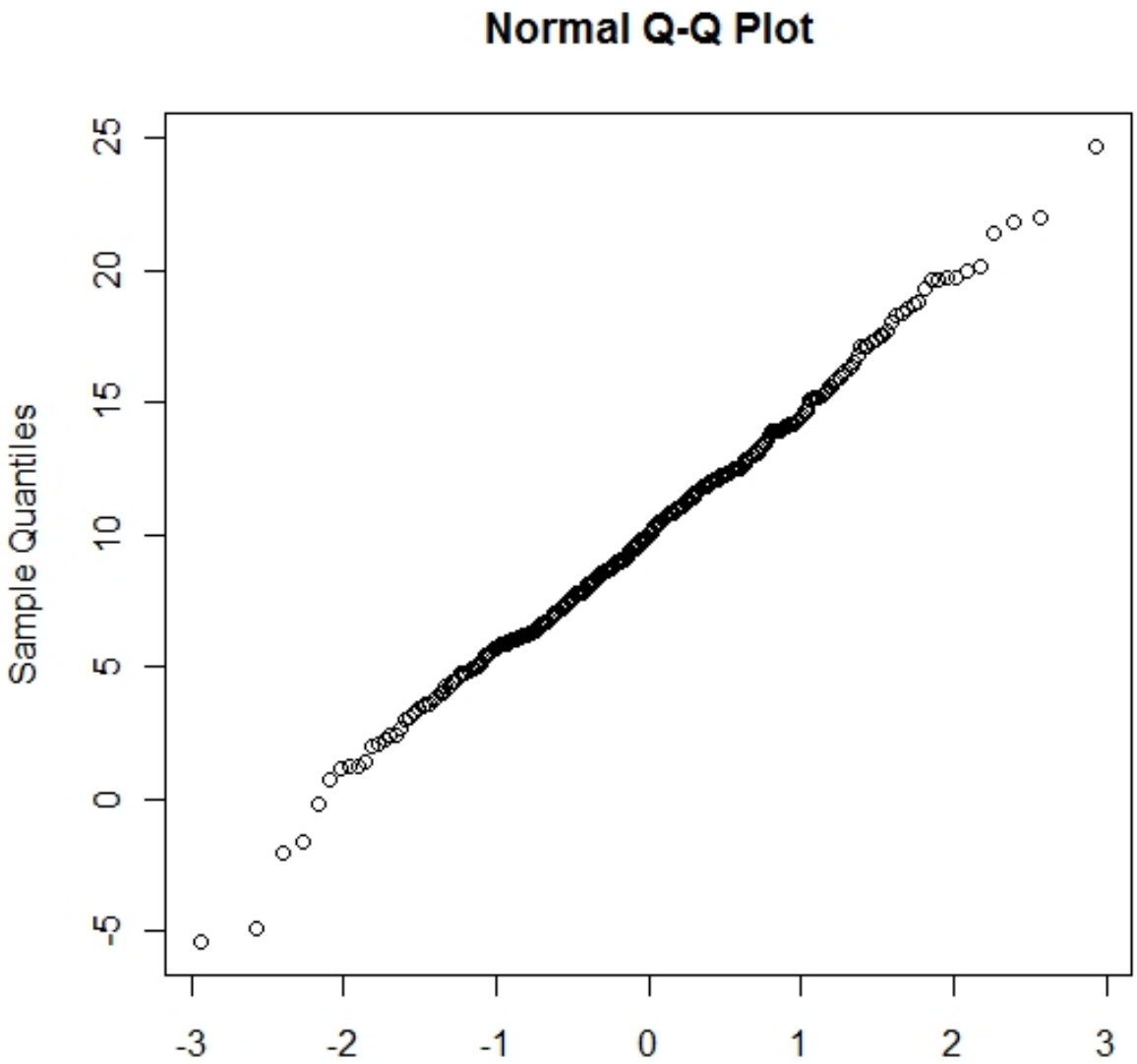
The observations are almost on the straight line: $y = 5x + 10$.

Fact:

Observations follow normal $\Leftrightarrow$ all points are roughly on a straight line.

Graphical Display of Distributions

Normal Q-Q plot for normal observations



Graphical Display of Distributions

Exercise 1:
The t-distribution is known as a heavy-tailed distribution, which have more extreme values than normal distributions.
Generate 300 random numbers from t_3 distribution, and create normal Q-Q plot.

Graphical Display of Distributions

Exercise 2:
A (continuous) uniform distribution ($f(x) = 1/(b-a)$ on $[a, b]$) have upper/lower limits, so there are no extreme values.
Generate 300 random numbers from a uniform distribution on $(0, 1)$, and create normal Q-Q plot.

Summary Statistics by groups

In addition to the 'apply' family in Chapter 1 slides, 'aggregate' and 'by' functions are useful to summarize grouped data.

Summary Statistics by groups

'aggregate'

```
> CO2
  Plant   Type  Treatment conc uptake
1   Qn1 Quebec nonchilled   95   16.0
2   Qn1 Quebec nonchilled  175   30.4
3   Qn1 Quebec nonchilled  250   34.8
... (truncated) ...
83   Mc3 Mississippi    chilled  675   18.9
84   Mc3 Mississippi    chilled 1000   19.9

> aggregate(CO2, CO2["Type"], mean)
         Type Plant Type Treatment conc   uptake
1      Quebec    NA   NA        NA  435 33.54286
2 Mississippi    NA   NA        NA  435 20.88333
Warning messages:
1: In mean.default(X[[1L]], ...) :
  argument is not numeric or logical: returning NA
```

Summary Statistics by groups

'by'

```
> by(CO2, CO2["Type"], mean)
Type: Quebec
    Plant      Type Treatment      conc    uptake
       NA        NA        NA 435.00000  33.54286
------------------------------------------------------------------------------
Type: Mississippi
    Plant      Type Treatment      conc    uptake
       NA        NA        NA 435.00000  20.88333

> by(CO2, CO2["Type"], summary)
Type: Quebec
     Plant             Type         Treatment       conc          uptake
 Qn1    :7   Quebec      :42   nonchilled:21   Min.   :  95   Min.   : 9.30
 Qn2    :7   Mississippi: 0   chilled   :21   1st Qu.: 175   1st Qu.:30.32
 Qn3    :7                                    Median : 350   Median :37.15
 Qc1    :7                                    Mean   : 435   Mean   :33.54
 Qc3    :7                                    3rd Qu.: 675   3rd Qu.:40.15
 Qc2    :7                                    Max.   :1000   Max.   :45.50
 (Other):0
<Truncated>
```

Graphics for grouped data

Box plot

Suppose we want to describe the sepal length in 'iris' data by species.

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
... (truncated) ...
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
... (truncated) ...
149          6.2         3.4          5.4         2.3  virginica
150          5.9         3.0          5.1         1.8  virginica
```

Graphics for grouped data

Box plot (Continued)

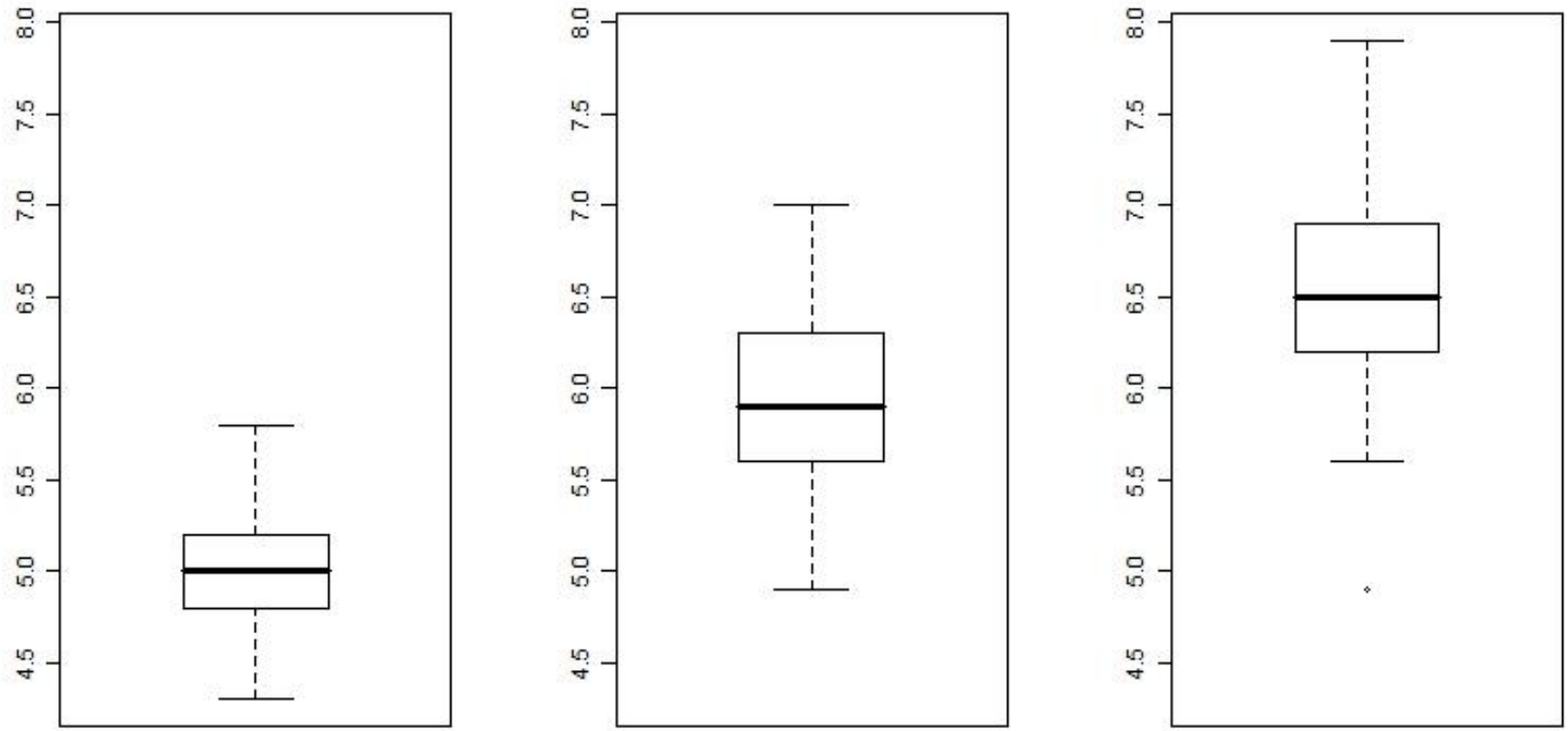
A box plot is a graphical way to describe the 1st quartile, median, 3rd quartile and outliers etc. We may want to lay out 3 box plots to compare these.

Sample R code 1:

```
D1 <- subset(iris, Species=="setosa")
D2 <- subset(iris, Species=="versicolor")
D3 <- subset(iris, Species=="virginica")

par(mfrow = c(1,3))  # align 3 figures as 1 x 3 by row
R <- range(iris$Sepal.Length)
boxplot(D1$Sepal.Length, ylim = R)
boxplot(D2$Sepal.Length, ylim = R)
boxplot(D3$Sepal.Length, ylim = R)
par(mfrow = c(1,1))  # return to single figure
```

Graphics for grouped data

Sample R code 1:



Graphics for grouped data

Sample R code 2:

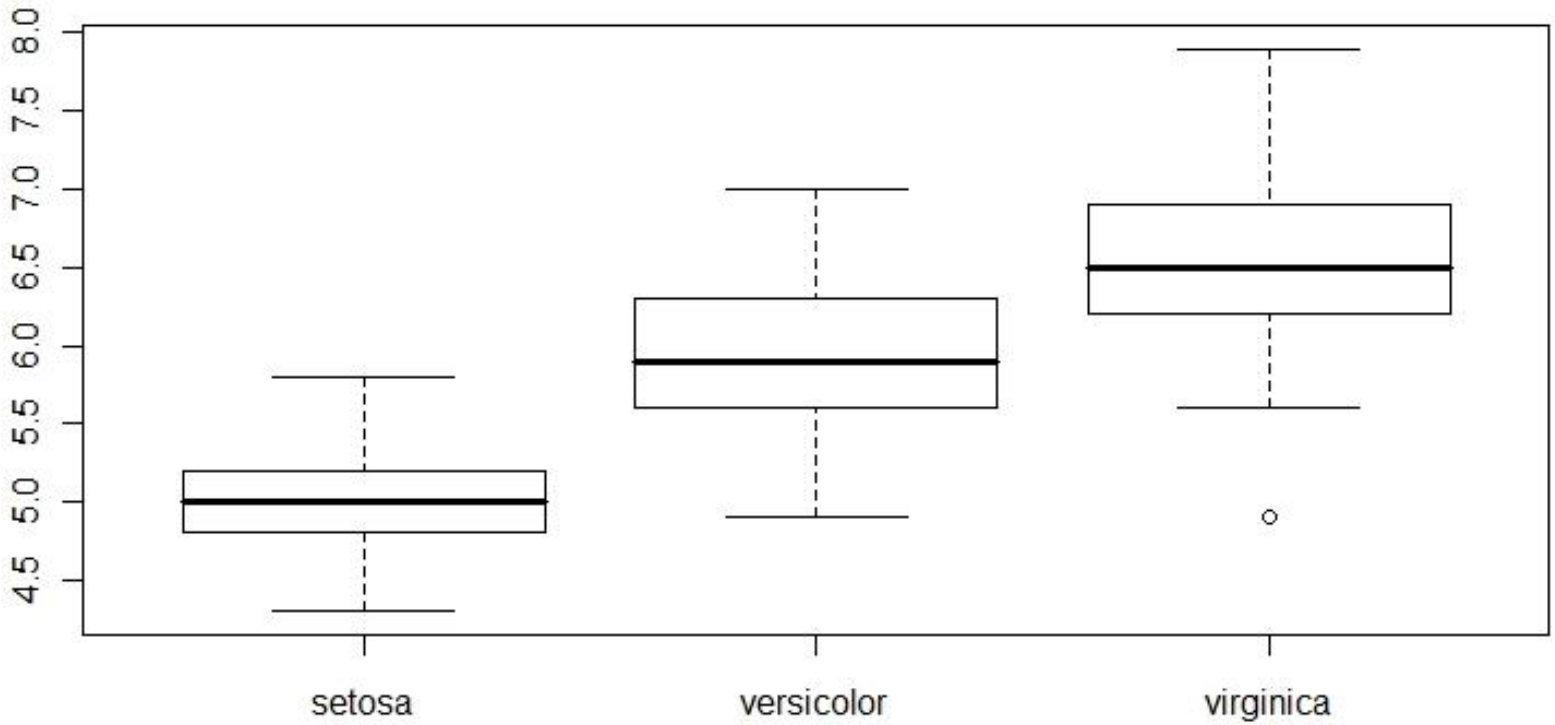
```
boxplot(iris$Sepal.Length ~ iris$Species)
```

Sample R code 3:

```
boxplot(D1$Sepal.Length, D2$Sepal.Length,D3$Sepal.Length)
```

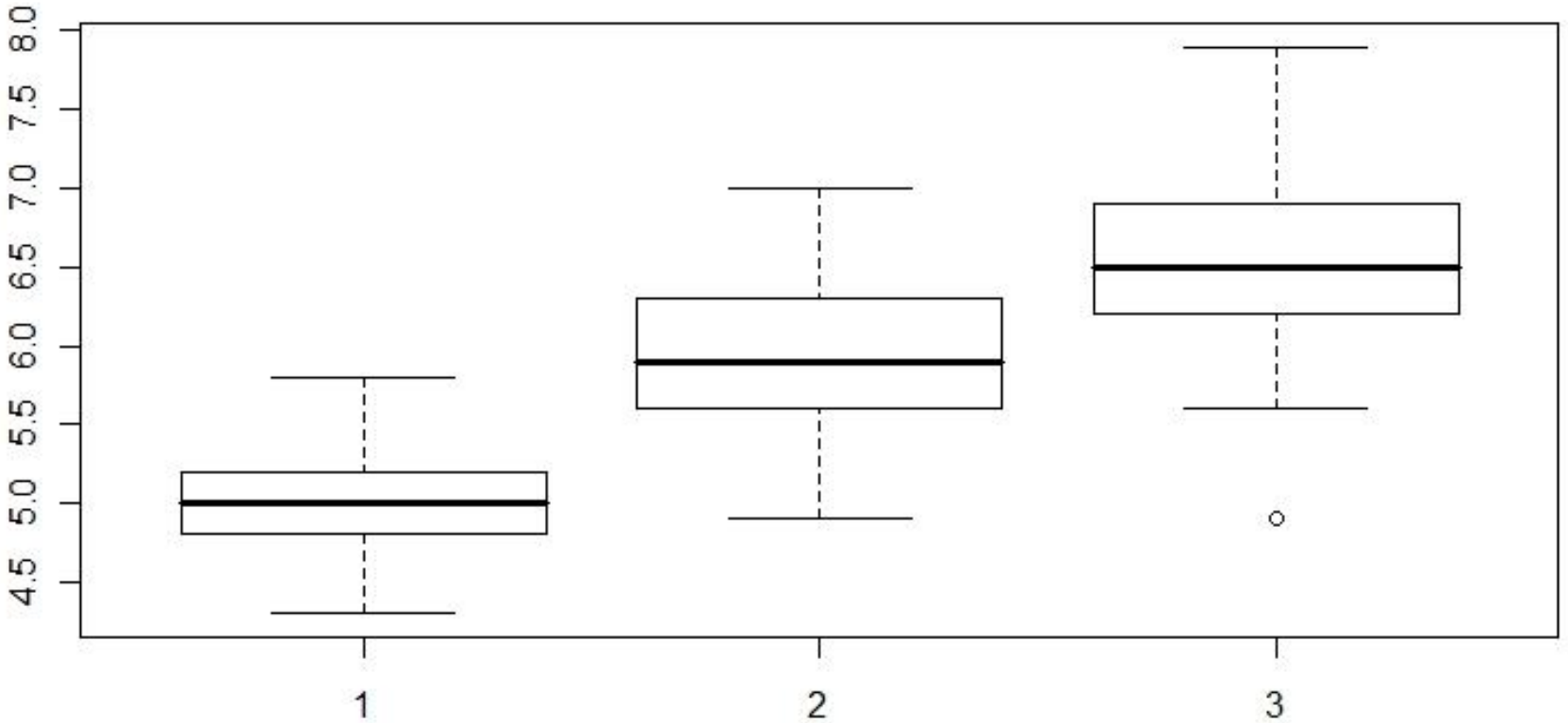
Graphics for grouped data

Sample R code 2:



Graphics for grouped data

Sample R code 3:



Tables

A contingency table in 'table' format can be transformed to a list of frequencies in 'data frame' format.
A 'table' object in R is basically the same as 'matrix'. When you transform it to a data frame, it has to be a 'table'.

Tables

How to make a table?

```
> matrix(c(1,2,3,4),2,2)
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> M <- matrix(c(1,2,3,4),2,2)
> colnames(M) <- c("A","B") # assign column names
> rownames(M) <- c("C","D") # assign row names
> names(dimnames(M)) <- c("Y","X")
  # assign category names for rows and columns
> M
   X
Y   A B
  C 1 3
  D 2 4
```

Tables

How to transform a table to a data frame?

```
> as.data.frame(M) # Nothing happens
  A B
C 1 3
D 2 4

> DF.M <- as.data.frame(as.table(M))
  # transform the table to a data frame
> DF.M
  Y X Freq
1 C A    1
2 D A    2
3 C B    3
4 D B    4
```

Tables

How to summarize a data frame?

Motor Cars Dataset in R package:

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Tables

How to summarize a data frame? – 1

```
> attach(mtcars)

> table(gear)  # summarize by gear
gear
 3  4  5
15 12  5

> table(gear, carb) # summarize by gear and carb
    carb
gear 1 2 3 4 6 8
   3 3 4 3 5 0 0
   4 4 4 0 4 0 0
   5 0 2 0 1 1 1
```

Tables

How to summarize a data frame? – 2

```
> xtabs(~ gear + carb, data=mtcars)
   # same as table(gear, carb)
    carb
gear 1 2 3 4 6 8
   3 3 4 3 5 0 0
   4 4 4 0 4 0 0
   5 0 2 0 1 1 1
```

Tables

How to summarize a data frame? – 3

Function	Description
ftable	Similar to 'xtabs'. It makes a flat (2-dimensional) table even when there are more than 2 variables.
margin.table	Sum up counts by rows (or columns).
prop.table	Transform a table of counts into a table of proportions.

Graphical Display of Tables

Barplot

Dataset:

```
> D <- HairEyeColor[,,1]
   # Hair/Eye color of statistics student (male)
> D
       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

> D.E <- margin.table(D,2) # Students by Eye
> D.E
Eye
Brown  Blue Hazel Green
   98   101    47    33
```

Graphical Display of Tables

Barplot

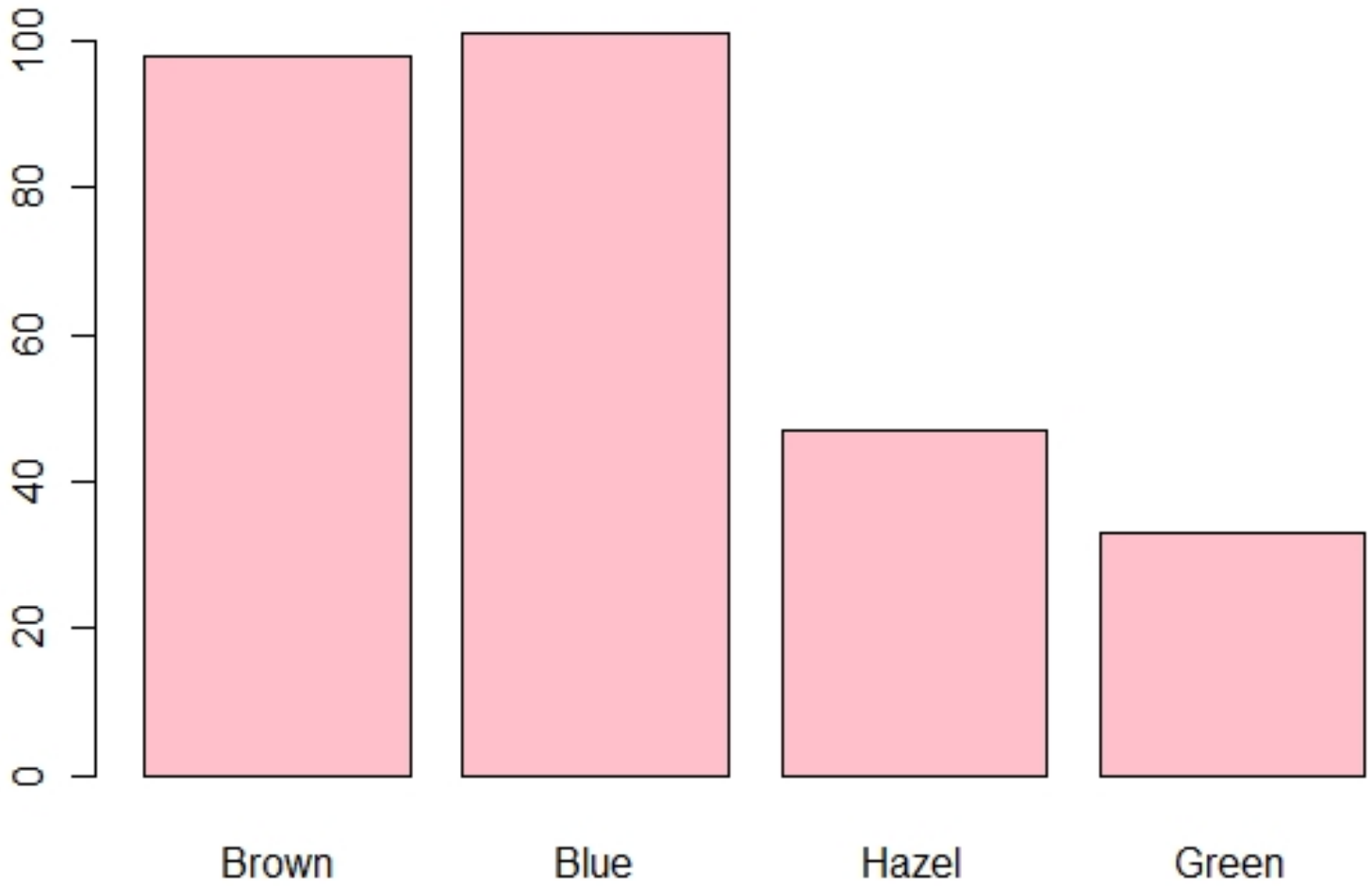
```
## barplot 1
barplot(D.H, col="pink") # students by eye

## barplot 2
Col <- c("black","brown","red",colors()[78])
barplot(D,col=Col, main="Students by Eye Color")
legend(3, 100,
c("Hair:Black","Hair:Brown","Hair:Red","Hair:Blond"),
col=Col, pch=15)
```

Note:

- Type ”colors()” to see all ready-made colors in R. (cf. http://research.stowers-institute.org/efg/R/Color/Chart/).
- ”pch” is point character, and ”pch=15” is a black square.

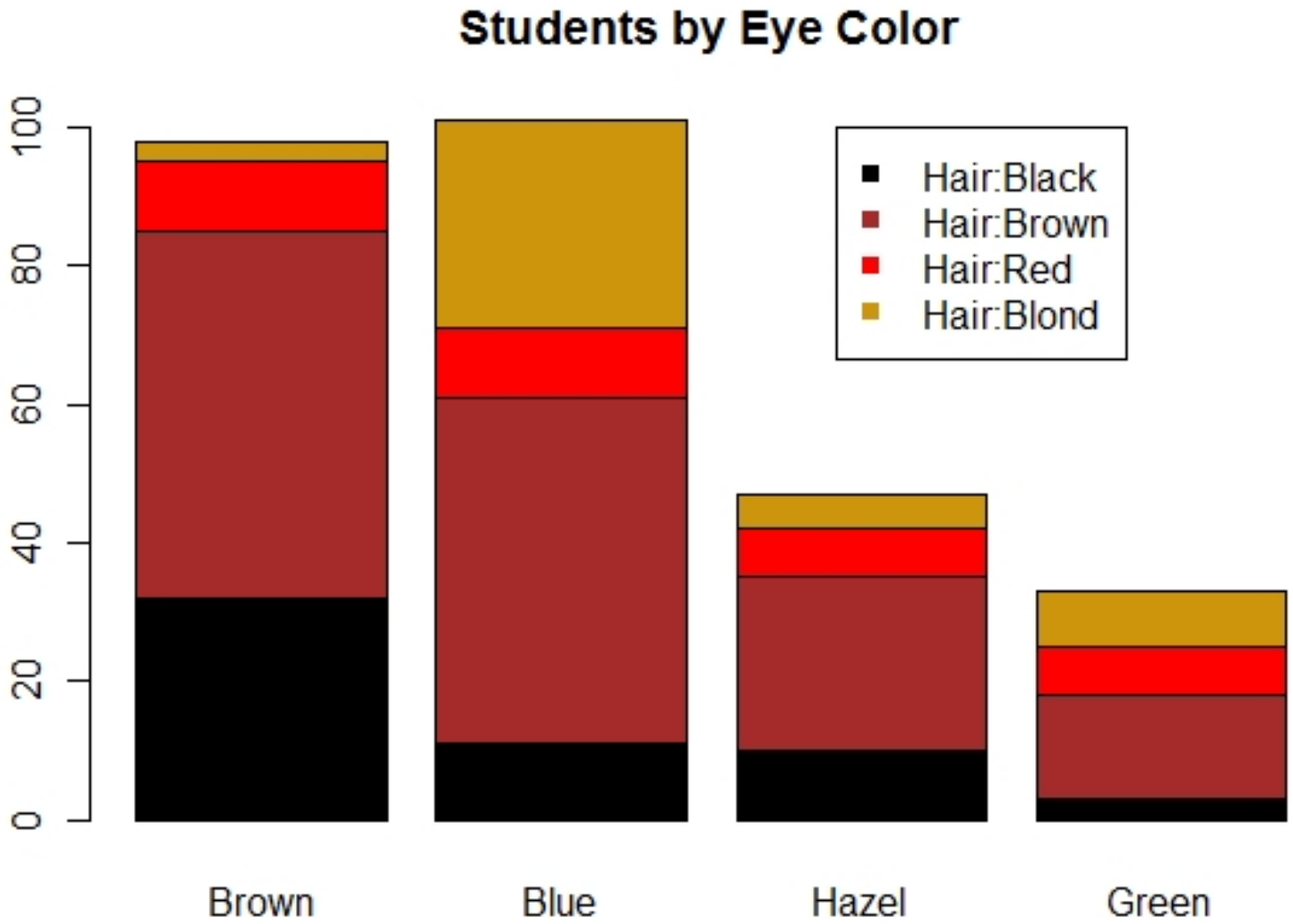
Graphical Display of Tables

Example: Barplot 1



Graphical Display of Tables

Example: Barplot 2



Graphical Display of Tables

Exercise

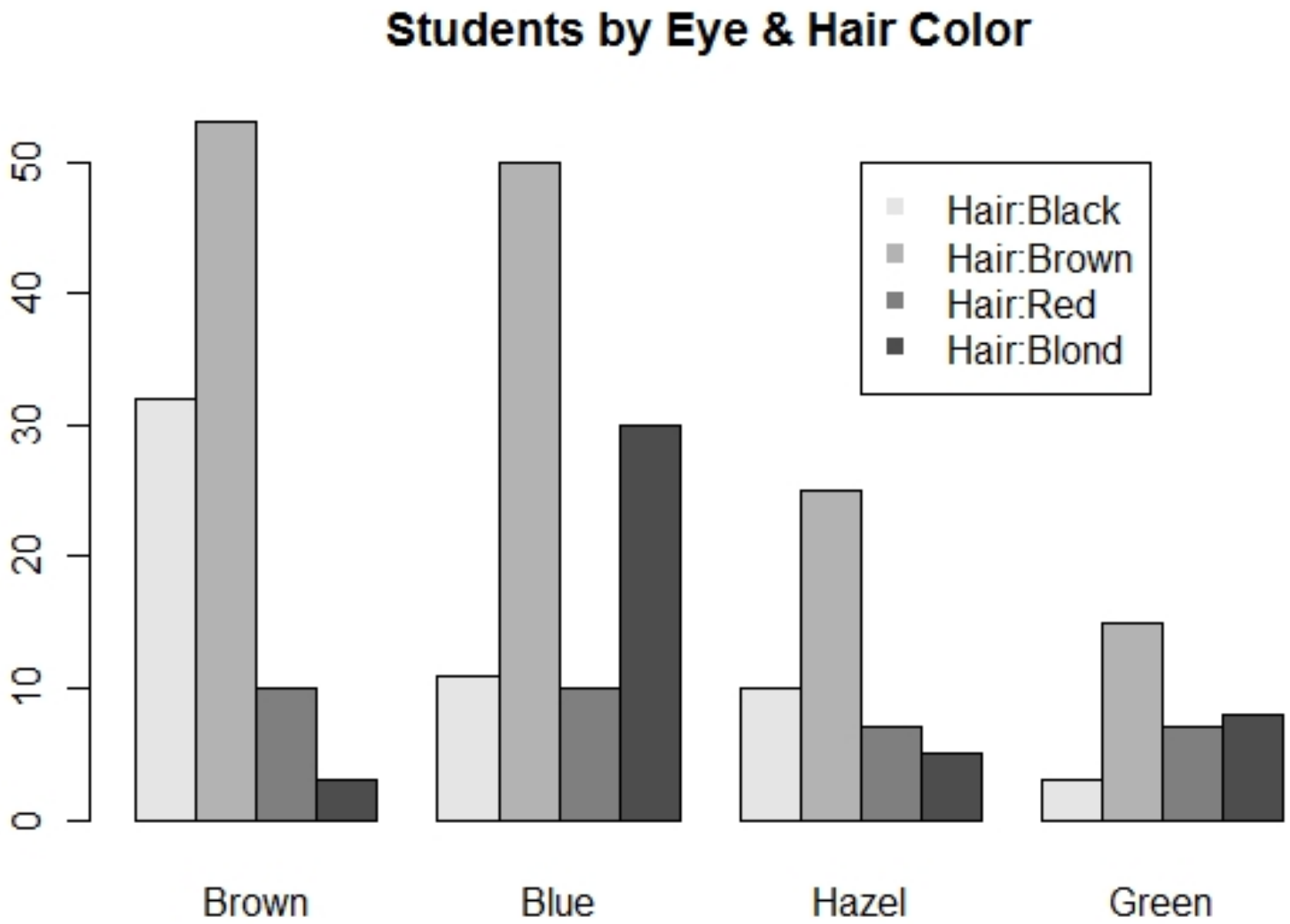
The ”HairEyeColor[,,2]” includes data for female. Create a barplot similar to the previous page, but interchange the role of eye and hair color.

Graphical Display of Tables

```
## barplot 3
Col2 <- c("grey90","grey70","grey50","grey30")
barplot(D,col=Col2, xlab="Eye",
beside=T, main="Students by Eye & Hair Color")
legend(13, 50,
c("Hair:Black","Hair:Brown","Hair:Red","Hair:Blond"),
col= Col2, pch=15)
```

Graphical Display of Tables

Example: Barplot 3



Graphical Display of Tables

Pie Chart

```
pie(D.E,
  col=c("brown","blue", colors()[146],"green"),
  main="Students by Eye Color")
```

Note:

- To start at 12 o'clock and move clockwise, add an option "clockwise=T".

Graphical Display of Tables