Quantitative Methods in Linguistics – Lecture 3

Adrian Brasoveanu*

March 30, 2014

Contents

This set of notes is primarily based on Gries (2009).

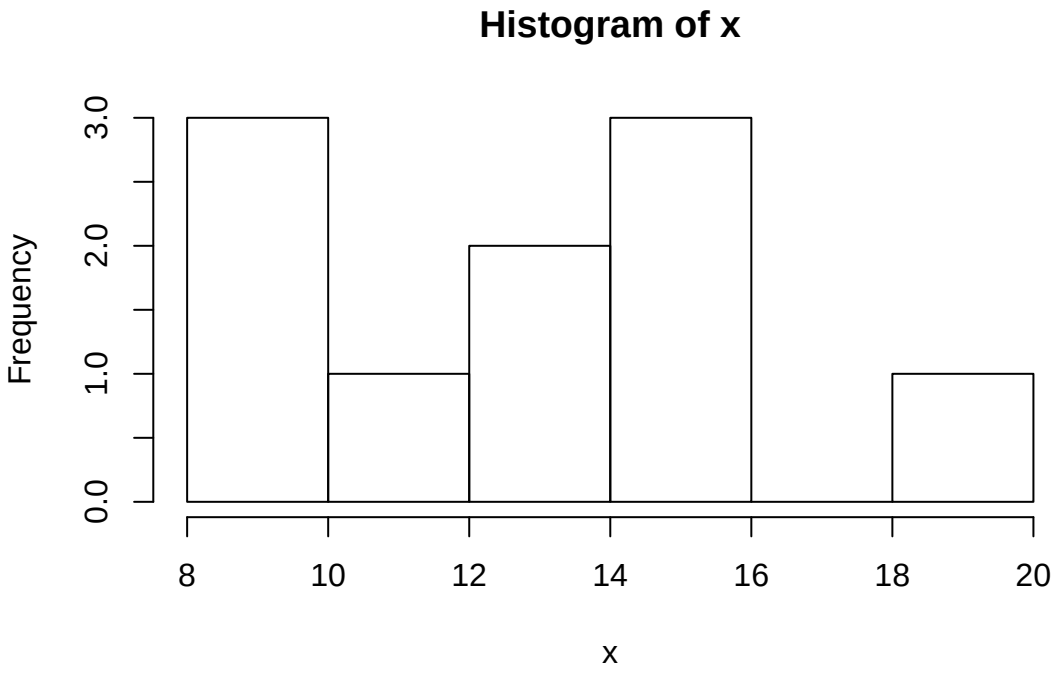
1 Basic graphics

```
> x <- c(12, 15, 13, 20, 14, 16, 10, 10, 8, 15)
```

A histogram:

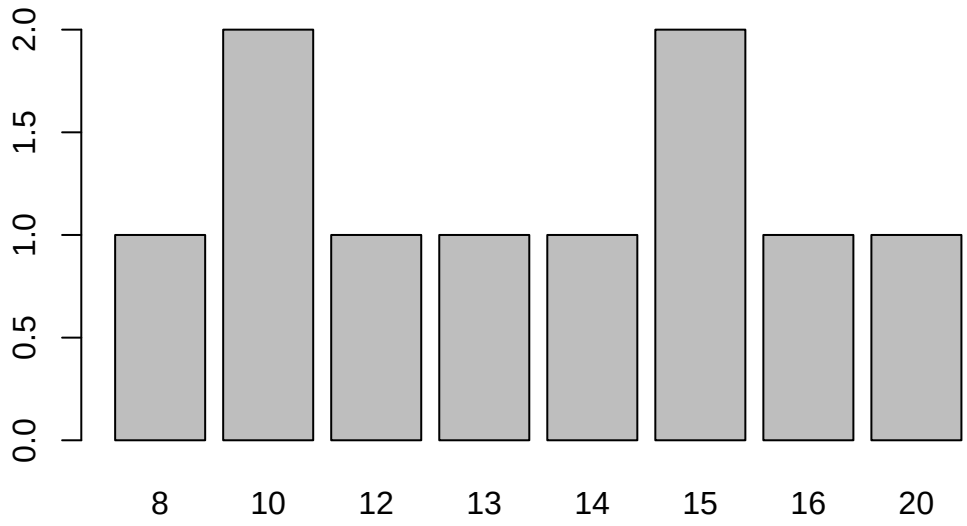
```
> hist(x)
```

*These notes have been generated with the 'knitr' package (Xie 2013) and are based on many sources, including but not limited to: Abelson (1995), Miles and Shevlin (2001), Faraway (2004), De Veaux et al. (2005), Braun and Murdoch (2007), Gelman and Hill (2007), Baayen (2008), Johnson (2008), Wright and London (2009), Gries (2009), Kruschke (2011), Diez et al. (2013), Gries (2013).



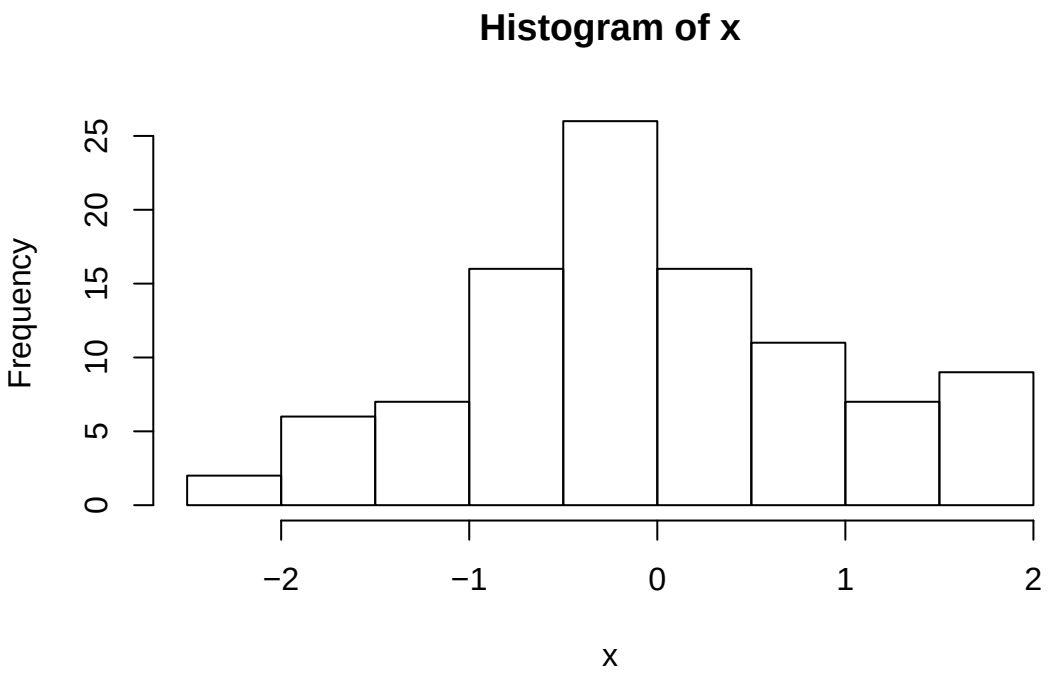
A barplot:

```
> barplot(table(x))
```

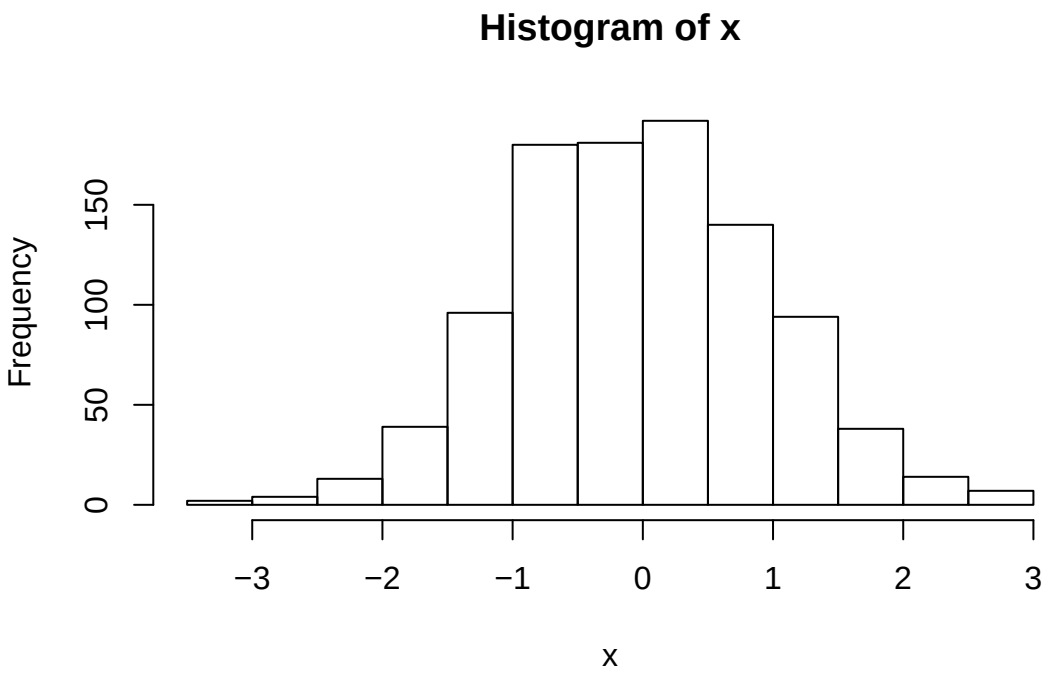


More examples, with random draws from a standard normal distribution (mean 0, standard deviation 1):

```
> (x <- rnorm(100))

  [1]  0.845301 -0.199772  0.106276 -0.976770 -0.367035 -0.578024 -0.495093
  [8]  1.030432  0.262994 -0.271282  0.465825  0.098610 -0.321201 -1.388742
 [15]  0.426482 -1.529441 -0.801854  0.070597 -0.259960 -1.793569  1.037764
 [22] -0.207403 -0.018258 -0.474986  1.440564 -0.776740  1.241672 -0.150820
 [29] -0.100256  1.202047 -0.426188 -0.517240 -0.408189 -1.341080 -0.826932
 [36]  1.884596 -0.839588 -0.903386 -1.778865  0.646681 -0.594170  0.298103
 [43] -0.592905 -0.493830 -0.160872 -0.743094  0.373379 -1.148508  1.664974
 [50] -1.931484  0.295470  0.539638 -0.713888 -0.400541 -2.008621  0.724535
 [57] -0.433387 -0.175161 -0.372436  0.059841 -0.225299  0.756914 -0.327854
 [64]  1.010017  0.347977  1.535635 -1.983544  0.551478 -0.735119  1.880514
 [71] -0.160039  0.593391  1.725170  0.236377 -0.738011  0.439917 -1.024118
 [78] -1.048861 -0.242941  1.943652 -1.940562  0.102366  1.772543 -1.247889
 [85]  0.989730  1.126204  1.543079  0.085151 -0.007284  0.571254  0.037215
 [92] -0.864348 -0.350355  0.868444 -0.875253 -2.321914  0.820119 -1.011135
 [99]  1.971407 -0.014549

> hist(x)
```

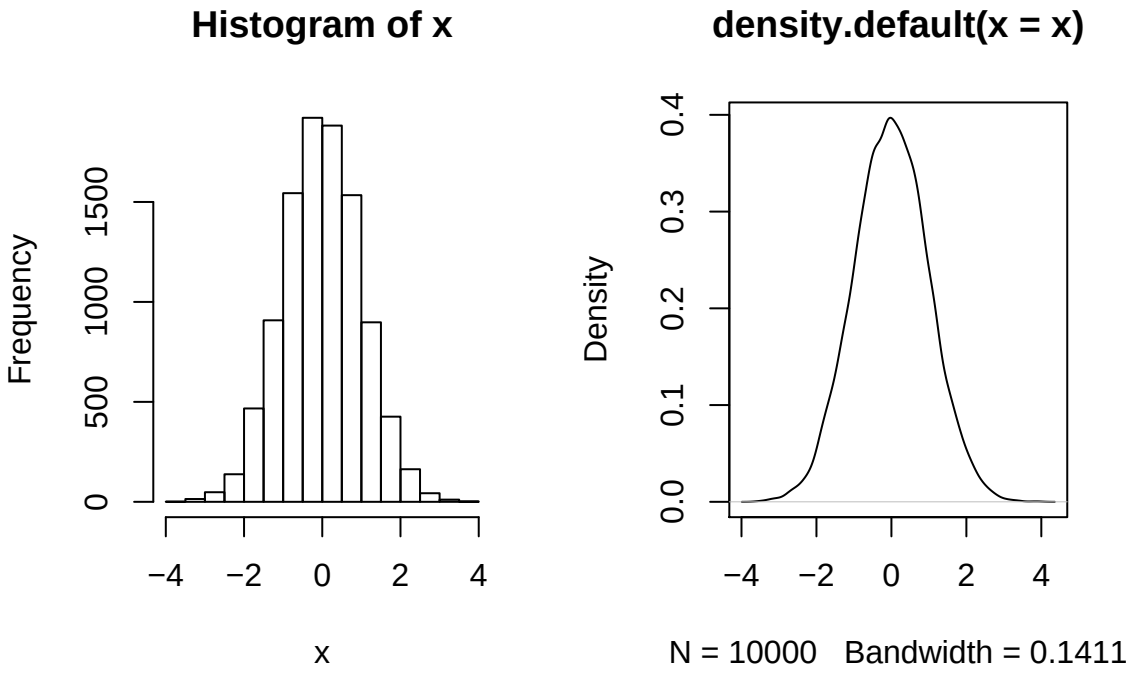



```
> x <- rnorm(1000)
> hist(x)
```



A histogram and the corresponding density plot:

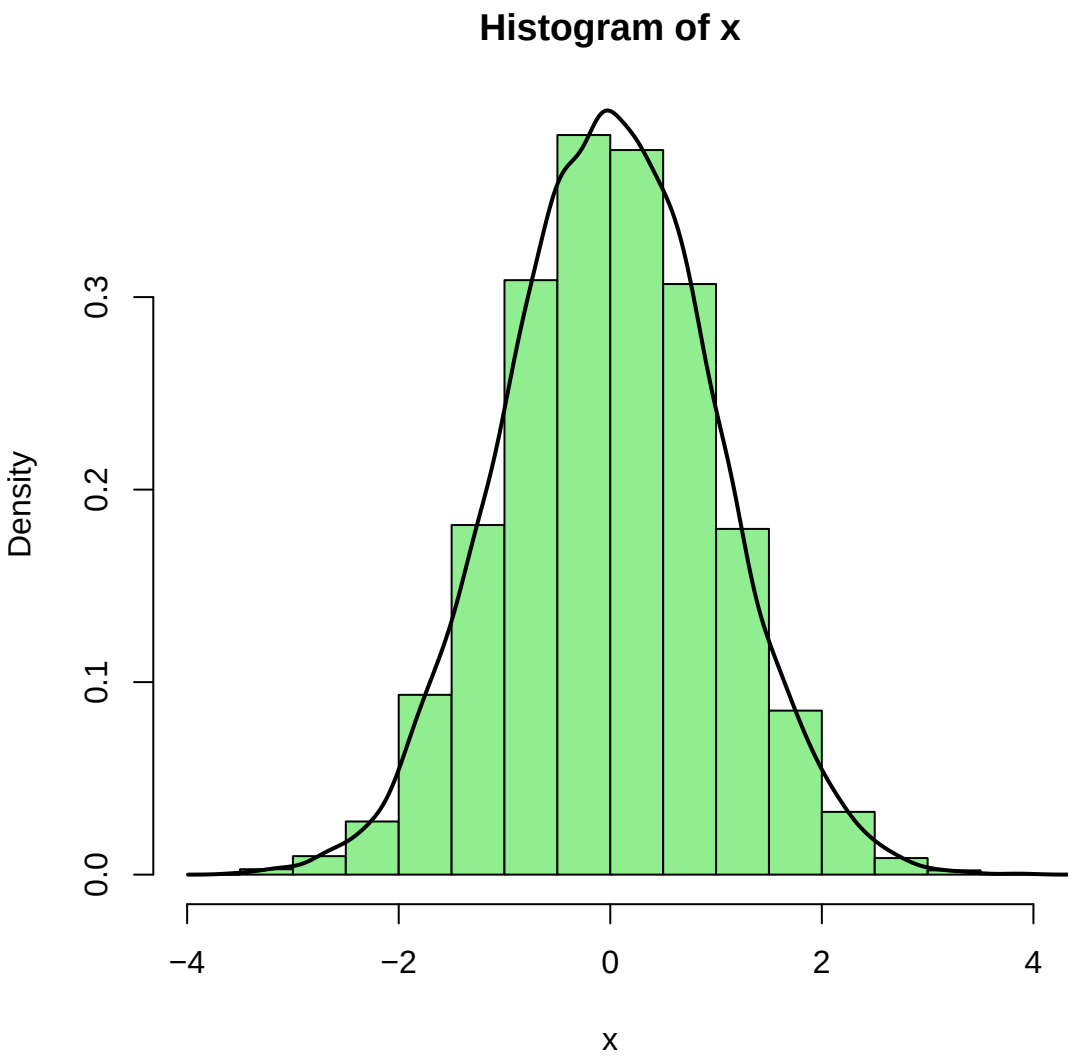
```
> par(mfrow = c(1, 2))
> x <- rnorm(10000)
> hist(x)
> plot(density(x))
```




```
> # ?density
> par(mfrow = c(1, 1))
```

The two together – note the `freq=F` option passed to the `hist` function:

```
> hist(x, col = "lightgreen", freq = F)
> lines(density(x), lwd = 2)
```

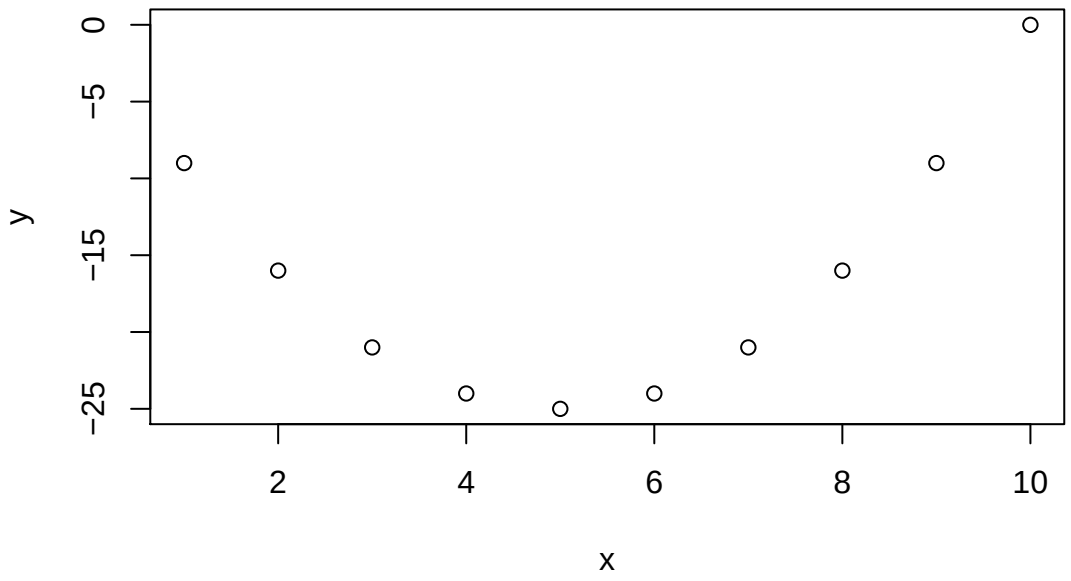



```
> # ?lines
```

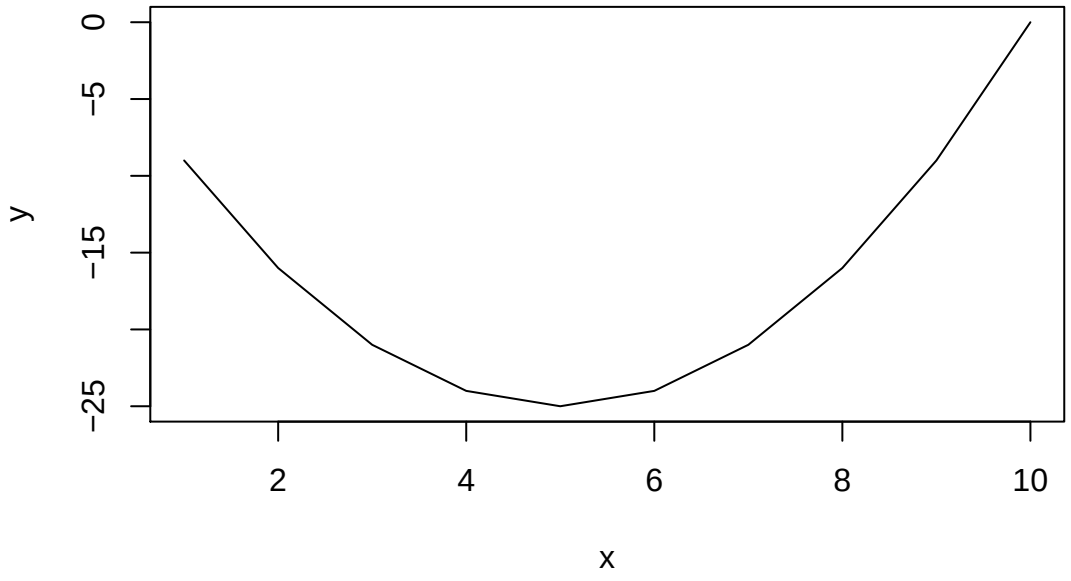
A scatterplot:

```
> (x <- seq(1, 10))

 [1]  1  2  3  4  5  6  7  8  9 10

> (y <- (x^2) - (10 * x))

 [1]  -9 -16 -21 -24 -25 -24 -21 -16  -9   0
```

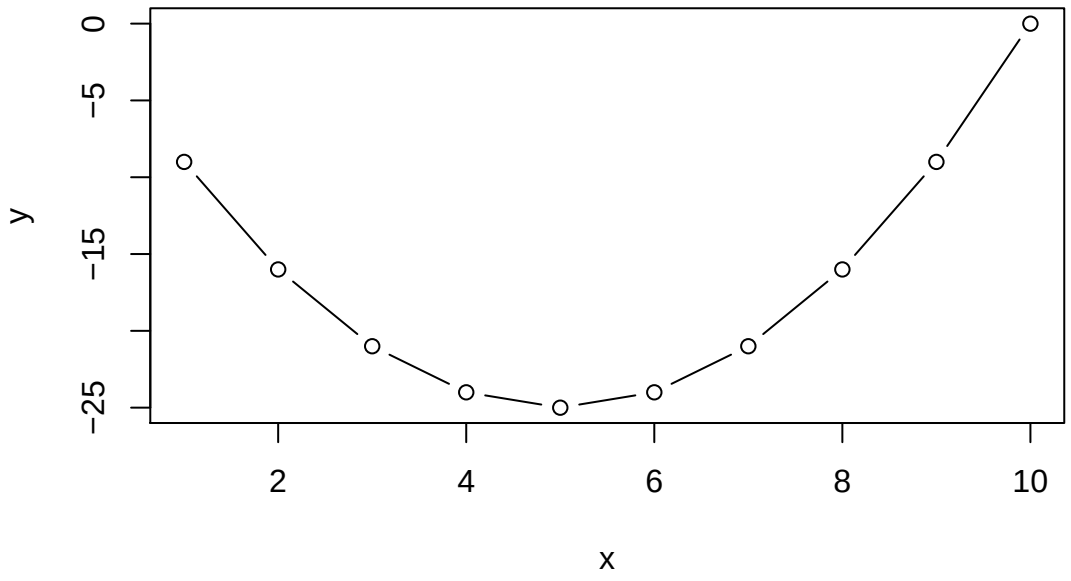
```
> plot(x, y)
```




```
> plot(x, y, type = "l")
```

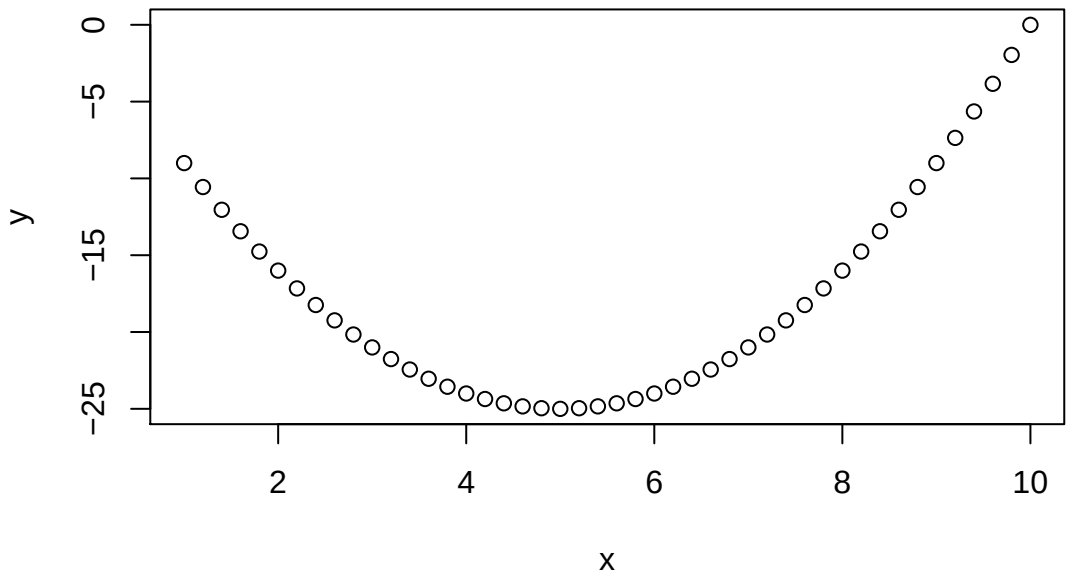


```
> plot(x, y, type = "b")
```

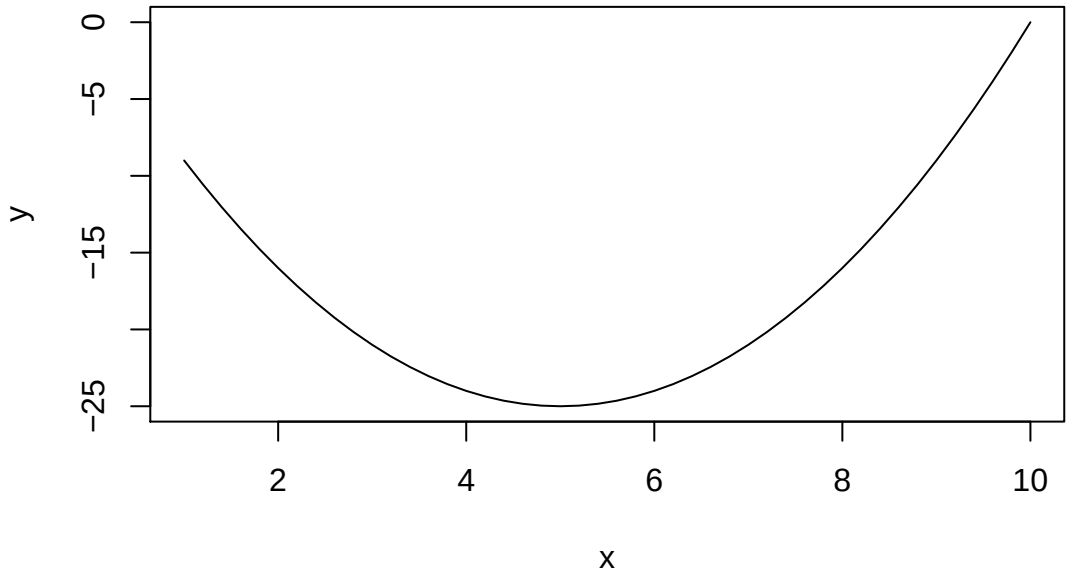



```
> (x <- seq(1, 10, by = 0.2))

 [1]  1.0  1.2  1.4  1.6  1.8  2.0  2.2  2.4  2.6  2.8  3.0  3.2  3.4  3.6
[15]  3.8  4.0  4.2  4.4  4.6  4.8  5.0  5.2  5.4  5.6  5.8  6.0  6.2  6.4
[29]  6.6  6.8  7.0  7.2  7.4  7.6  7.8  8.0  8.2  8.4  8.6  8.8  9.0  9.2
[43]  9.4  9.6  9.8 10.0

> (y <- (x^2) - (10 * x))

 [1]  -9.00 -10.56 -12.04 -13.44 -14.76 -16.00 -17.16 -18.24 -19.24 -20.16
[11] -21.00 -21.76 -22.44 -23.04 -23.56 -24.00 -24.36 -24.64 -24.84 -24.96
[21] -25.00 -24.96 -24.84 -24.64 -24.36 -24.00 -23.56 -23.04 -22.44 -21.76
[31] -21.00 -20.16 -19.24 -18.24 -17.16 -16.00 -14.76 -13.44 -12.04 -10.56
[41]  -9.00  -7.36  -5.64  -3.84  -1.96   0.00
```

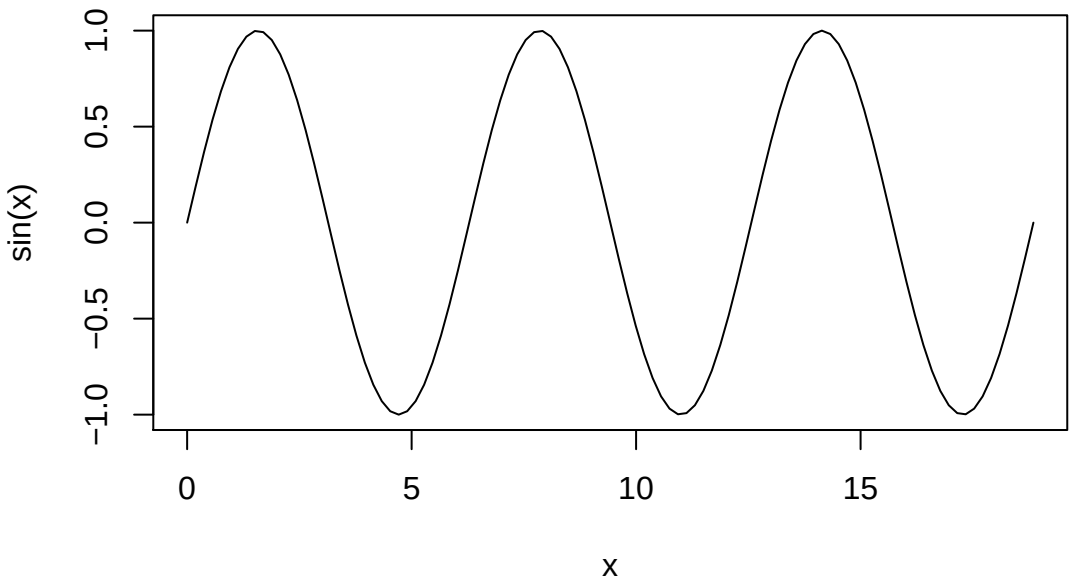
```
> plot(x, y)
```



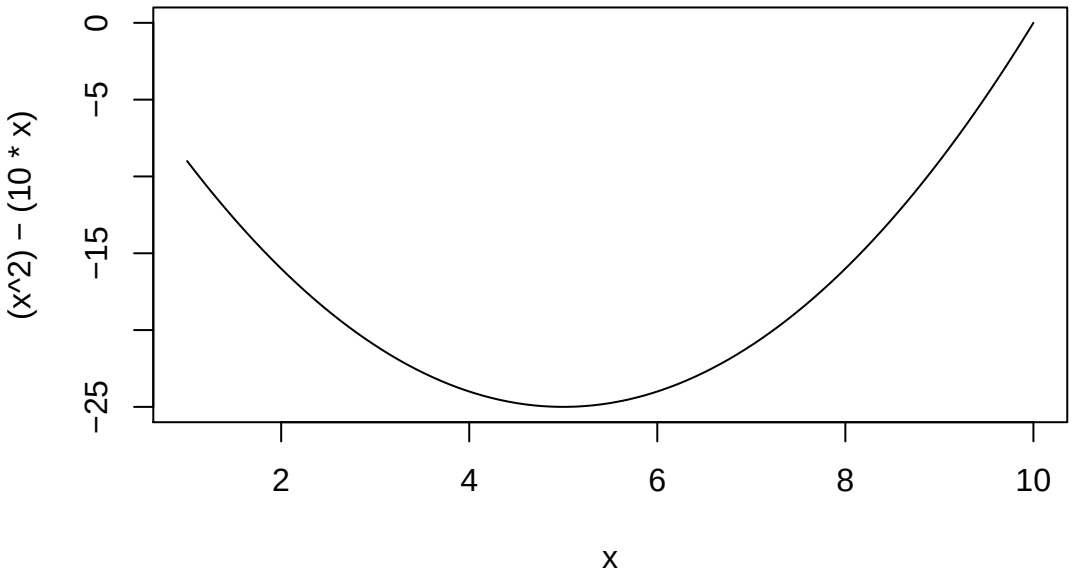

```
> plot(x, y, type = "l")
```



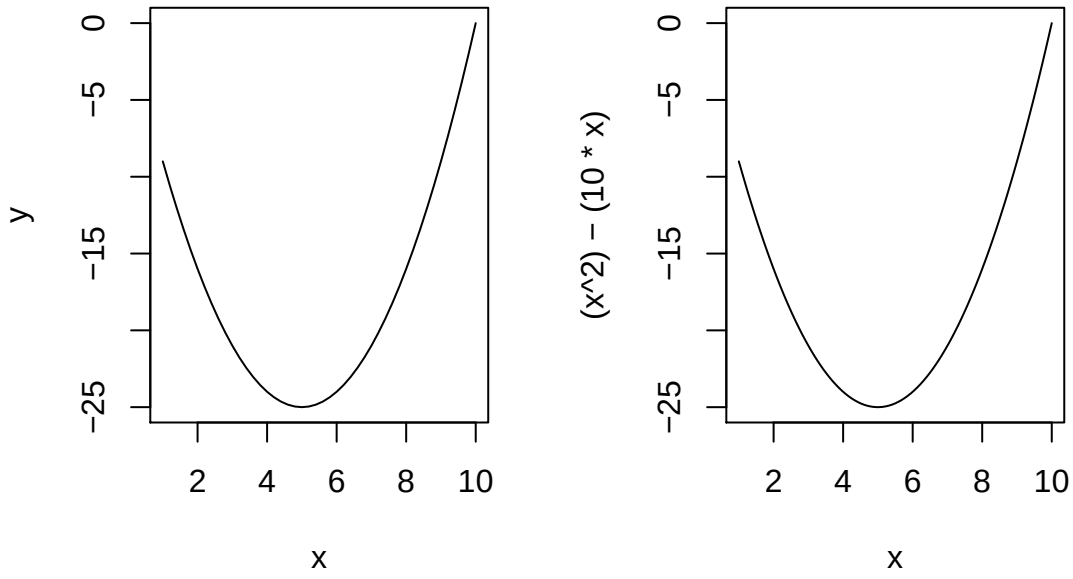
```
> curve(expr = sin, from = 0, to = 6 * pi)
```




```
> curve((x^2) - (10 * x), from = 1, to = 10)
```

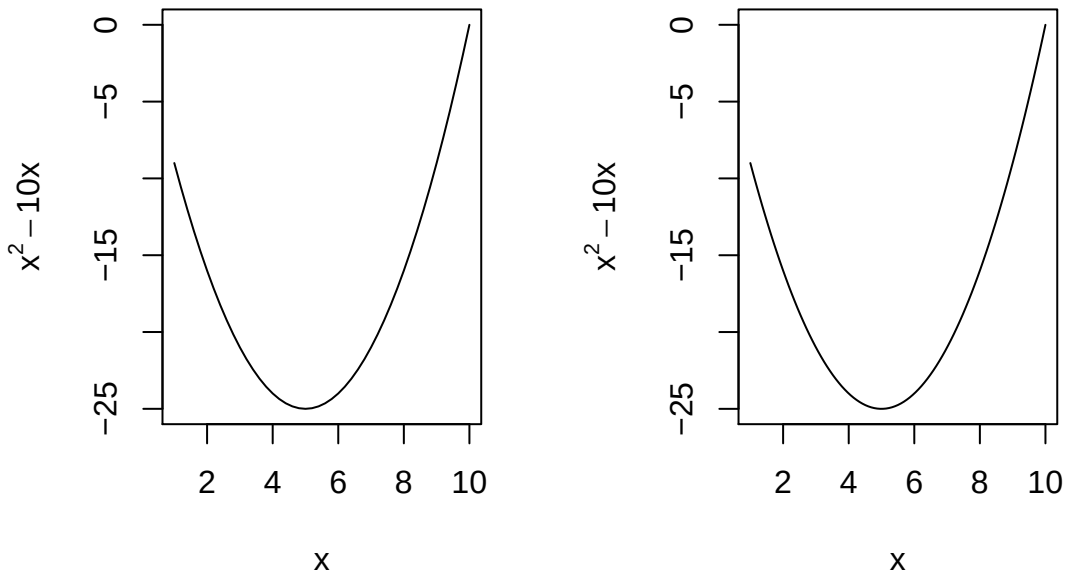


```
> par(mfrow = c(1, 2))
> (x <- seq(1, 10, by = 0.2))

 [1]  1.0  1.2  1.4  1.6  1.8  2.0  2.2  2.4  2.6  2.8  3.0  3.2  3.4  3.6
[15]  3.8  4.0  4.2  4.4  4.6  4.8  5.0  5.2  5.4  5.6  5.8  6.0  6.2  6.4
[29]  6.6  6.8  7.0  7.2  7.4  7.6  7.8  8.0  8.2  8.4  8.6  8.8  9.0  9.2
[43]  9.4  9.6  9.8 10.0

> (y <- (x^2) - (10 * x))

 [1]  -9.00 -10.56 -12.04 -13.44 -14.76 -16.00 -17.16 -18.24 -19.24 -20.16
[11] -21.00 -21.76 -22.44 -23.04 -23.56 -24.00 -24.36 -24.64 -24.84 -24.96
[21] -25.00 -24.96 -24.84 -24.64 -24.36 -24.00 -23.56 -23.04 -22.44 -21.76
[31] -21.00 -20.16 -19.24 -18.24 -17.16 -16.00 -14.76 -13.44 -12.04 -10.56
[41]  -9.00  -7.36  -5.64  -3.84  -1.96   0.00

> plot(x, y, type = "l")
> curve((x^2) - (10 * x), from = 1, to = 10)
```




```
> par(mfrow = c(1, 1))
```

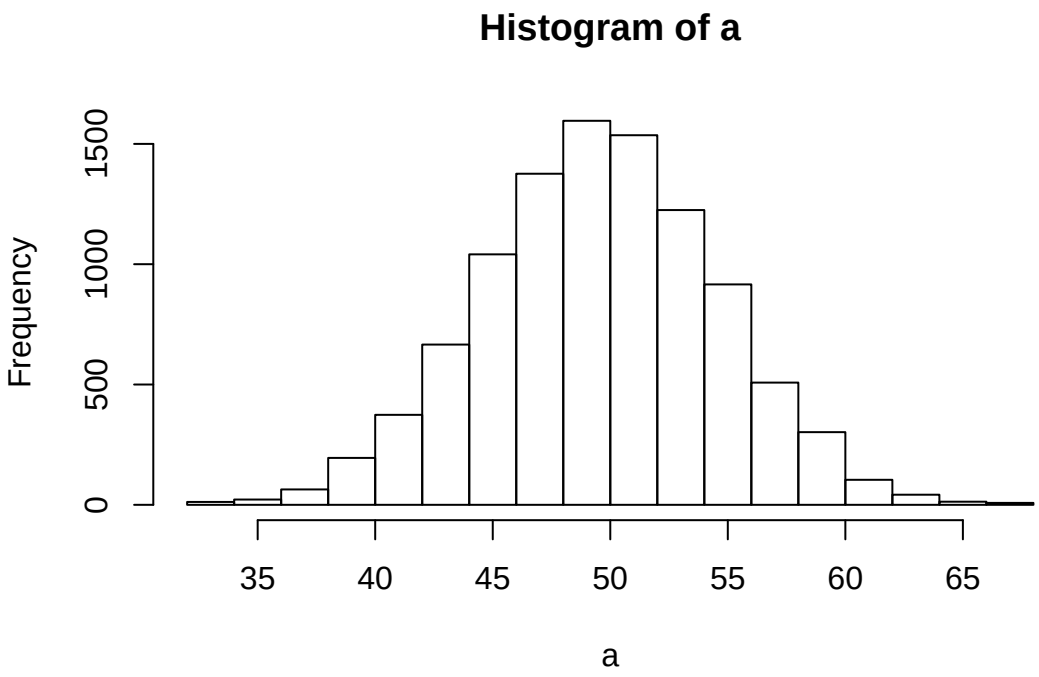
```
> par(mfrow = c(1, 2), mai = c(1.02, 0.92, 0.82, 0.42))
> plot(x, y, type = "l", ylab = expression(x^2 - 10 * x))
> curve((x^2) - (10 * x), from = 1, to = 10, ylab = expression(x^2 -
+     10 * x))
```



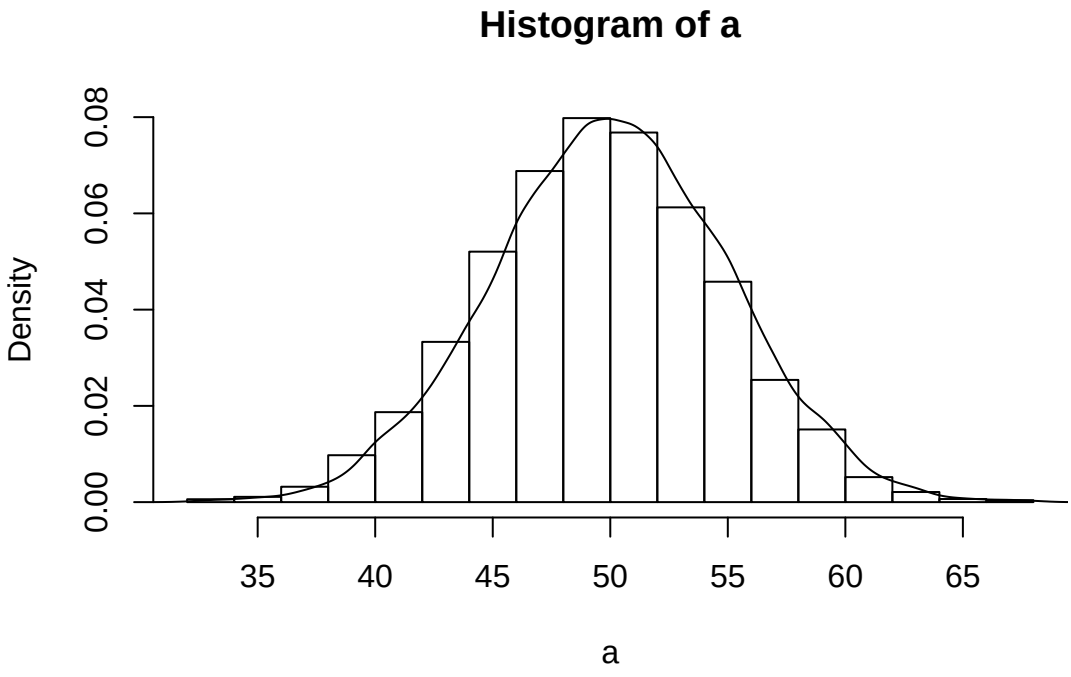

```
> par(mfrow = c(1, 1), mai = c(1.02, 0.82, 0.82, 0.42))
```

```
> # ?rbinom
> (a <- rbinom(100, 1, 0.5))

  [1] 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 1 0 0 1 1 1 1 0 0 1 0 0 0 1 1 0 0 1
 [36] 0 0 1 1 0 1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 1 1 1 0 1 1 1 0 1 0 1
 [71] 0 1 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 0

> sum(a)

[1] 51
```

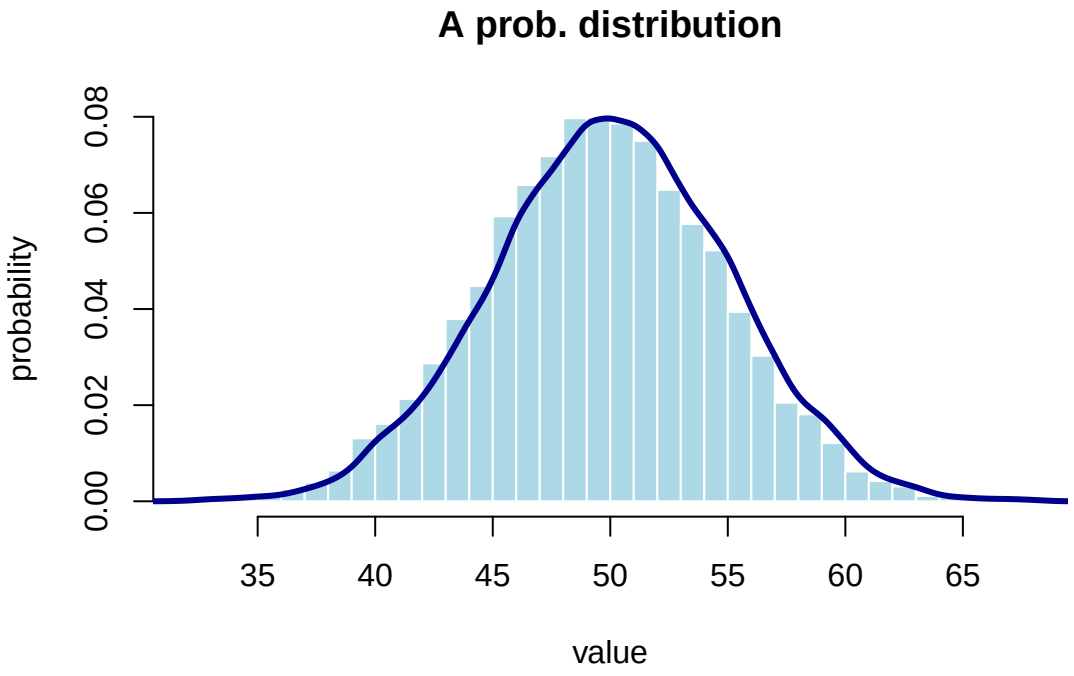
```
> a <- rbinom(10000, 100, 0.5)
> hist(a)
```




```
> hist(a, probability = TRUE)
> lines(density(a))
```



```
> hist(a, probability = TRUE, col = "lightblue", border = "white", main = "A prob. distribution",
+     xlab = "value", ylab = "probability", breaks = 30)
> # ?hist
> lines(density(a), col = "darkblue", lwd = 3)
```



2 Data frames

```
> rm(list = ls(all = T))  # clear workspace
> PartOfSpeech <- c("ADJ", "ADV", "N", "CONJ", "PREP")
> TokenFrequency <- c(421, 337, 1411, 458, 455)
> TypeFrequency <- c(271, 103, 735, 18, 37)
> Class <- c("open", "open", "open", "closed", "closed")
> x <- data.frame(PartOfSpeech, TokenFrequency, TypeFrequency, Class)
> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> y <- c(5, 4, 3, 2, 1)
> z <- data.frame(x, y)
> str(x)

'data.frame': 5 obs. of  4 variables:
```

```
 $ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4 3 5
 $ TokenFrequency: num  421 337 1411 458 455
 $ TypeFrequency : num  271 103 735 18 37
 $ Class         : Factor w/ 2 levels "closed","open": 2 2 2 1 1

> summary(x)

 PartOfSpeech TokenFrequency TypeFrequency    Class
 ADJ :1       Min.   : 337   Min.   : 18   closed:2
 ADV :1       1st Qu.: 421   1st Qu.: 37   open  :3
 CONJ:1       Median : 455   Median :103
 N   :1       Mean   : 616   Mean   :233
 PREP:1       3rd Qu.: 458   3rd Qu.:271
              Max.   :1411   Max.   :735

> x$PartOfSpeech

[1] ADJ  ADV  N    CONJ PREP
Levels: ADJ ADV CONJ N PREP

> str(x$PartOfSpeech)

 Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4 3 5

> summary(x$PartOfSpeech)

 ADJ  ADV CONJ    N PREP
   1    1    1    1    1

> x$foo <- c(5, 4, 3, 2, 1)
> x

  PartOfSpeech TokenFrequency TypeFrequency  Class foo
1          ADJ            421           271   open   5
2          ADV            337           103   open   4
3            N           1411           735   open   3
4         CONJ            458            18 closed   2
5         PREP            455            37 closed   1

> x$bar <- x$foo * 2
> x

  PartOfSpeech TokenFrequency TypeFrequency  Class foo bar
1          ADJ            421           271   open   5  10
2          ADV            337           103   open   4   8
3            N           1411           735   open   3   6
4         CONJ            458            18 closed   2   4
5         PREP            455            37 closed   1   2

> (x.2 <- data.frame(TokenFrequency, TypeFrequency, Class, row.names = PartOfSpeech))

     TokenFrequency TypeFrequency  Class
ADJ             421           271   open
ADV             337           103   open
N              1411           735   open
CONJ            458            18 closed
PREP            455            37 closed
```

```
> str(x.2)

'data.frame': 5 obs. of  3 variables:
 $ TokenFrequency: num  421 337 1411 458 455
 $ TypeFrequency : num  271 103 735 18 37
 $ Class         : Factor w/ 2 levels "closed","open": 2 2 2 1 1

> x.2$PartOfSpeech

NULL

> row.names(x.2)

[1] "ADJ"  "ADV"  "N"    "CONJ" "PREP"

> names(x.2)

[1] "TokenFrequency" "TypeFrequency"  "Class"
```

2.1 Saving a data frame to a file

```
> PartOfSpeech <- c("ADJ", "ADV", "N", "CONJ", "PREP")
> TokenFrequency <- c(421, 337, 1411, 458, 455)
> TypeFrequency <- c(271, 103, 735, 18, 37)
> Class <- c("open", "open", "open", "closed", "closed")
> (x <- data.frame(PartOfSpeech, TokenFrequency, TypeFrequency, Class))

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> write.table(x, file = "dataframe.txt", append = F, sep = "\t", eol = "\n",
+     na = "NA", dec = ".", quote = F, row.names = F, col.names = T)
> rm(list = ls(all = T))
> ls()

character(0)

> (x <- read.table(file = "dataframe.txt", header = T, sep = "\t", comment.char = ""))

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> # View(x)
> str(x)
```

```
'data.frame': 5 obs. of  4 variables:
 $ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4 3 5
 $ TokenFrequency: int  421 337 1411 458 455
 $ TypeFrequency : int  271 103 735 18 37
 $ Class         : Factor w/ 2 levels "closed","open": 2 2 2 1 1

> x$TokenFrequency

[1]  421  337 1411  458  455

> x$Class

[1] open   open   open   closed closed
Levels: closed open

> write.csv(x, "dataframe.csv", row.names = F)
> rm(list = ls(all = T))
> ls()

character(0)

> (x <- read.csv("dataframe.csv"))

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> str(x)

'data.frame': 5 obs. of  4 variables:
 $ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4 3 5
 $ TokenFrequency: int  421 337 1411 458 455
 $ TypeFrequency : int  271 103 735 18 37
 $ Class         : Factor w/ 2 levels "closed","open": 2 2 2 1 1
```

2.2 Attaching and detaching data frames

Warning: attach data frames sparingly – if ever!

```
> attach(x)
> Class

[1] open   open   open   closed closed
Levels: closed open

> TokenFrequency[4] <- 20
> TokenFrequency

[1]  421  337 1411   20  455

> x$TokenFrequency

[1]  421  337 1411  458  455
```

```
> TokenFrequency[4] <- 458
> detach(x)
> Class

Error:  object 'Class' not found

> x$Class

[1] open   open   open   closed closed
Levels: closed open
```

3 Subsetting data frames

```
> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> x[2, 3]

[1] 103

> x[2, ]

  PartOfSpeech TokenFrequency TypeFrequency Class
2          ADV            337           103  open

> x[, 3]

[1] 271 103 735  18  37

> x$TypeFrequency

[1] 271 103 735  18  37

> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> x[2:3, 4]

[1] open open
Levels: closed open

> x[3:4, 4]
```

```
[1] open   closed
Levels: closed open

> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> x[c(1, 3), c(2, 4)]

  TokenFrequency Class
1            421  open
3           1411  open

> which(x[, 2] > 450)

[1] 3 4 5

> which(x[2, ] > 2)

Warning:  > not meaningful for factors
Warning:  > not meaningful for factors

[1] 2 3

> (y <- x[, 2])

[1]  421  337 1411  458  455

> x$TokenFrequency

[1]  421  337 1411  458  455

> y[which(x[, 2] > 450)]

[1] 1411  458  455

> x$TokenFrequency[which(x[, 2] > 450)]

[1] 1411  458  455

> x$TokenFrequency[which(x$TokenFrequency > 450)]

[1] 1411  458  455

> x[, 2][which(x[, 2] > 450)]

[1] 1411  458  455

> x[, 3][x[, 3] > 100]

[1] 271 103 735

> x$TypeFrequency[x$TypeFrequency > 100]
```

```
[1] 271 103 735

> (y <- x[which(x[, 4] == "open"), ])

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open
3            N           1411           735  open

> (y <- x[which(x$Class == "open"), ])

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open
3            N           1411           735  open

> (y <- subset(x, Class == "open"))

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open
3            N           1411           735  open

> (y <- subset(x, Class == "open" & TokenFrequency < 1000))

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open

> (y <- subset(x, PartOfSpeech %in% c("ADJ", "ADV")))

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open
```

3.1 Ordering data frames

```
> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> x$TokenFrequency

[1]  421  337 1411  458  455

> order(x$TokenFrequency)

[1] 2 1 5 4 3

> x$TokenFrequency[order(x$TokenFrequency)]
```

```
[1]  337  421  455  458 1411

> (ordering.index <- order(x$TokenFrequency))

[1] 2 1 5 4 3

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
2          ADV            337           103   open
1          ADJ            421           271   open
5         PREP            455            37 closed
4         CONJ            458            18 closed
3            N           1411           735   open

> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> x[order(x$TokenFrequency), ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
2          ADV            337           103   open
1          ADJ            421           271   open
5         PREP            455            37 closed
4         CONJ            458            18 closed
3            N           1411           735   open

> -x$TokenFrequency

[1]  -421  -337 -1411  -458  -455

> order(-x$TokenFrequency)

[1] 3 4 5 1 2

> x$TokenFrequency[order(-x$TokenFrequency)]

[1] 1411  458  455  421  337

> (ordering.index <- order(x$TypeFrequency))

[1] 4 5 2 1 3

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
2          ADV            337           103   open
1          ADJ            421           271   open
3            N           1411           735   open
```

```
> x$Class

[1] open   open   open   closed closed
Levels: closed open

> str(x$Class)

 Factor w/ 2 levels "closed","open": 2 2 2 1 1

> order(x$Class)

[1] 4 5 1 2 3

> x$Class[order(x$Class)]

[1] closed closed open   open   open
Levels: closed open

> (ordering.index <- order(x$Class))

[1] 4 5 1 2 3

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open

> (ordering.index <- order(x$Class, x$TokenFrequency))

[1] 5 4 2 1 3

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
5         PREP            455            37 closed
4         CONJ            458            18 closed
2          ADV            337           103   open
1          ADJ            421           271   open
3            N           1411           735   open

> (ordering.index <- order(x$Class, -x$TokenFrequency))

[1] 4 5 3 1 2

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
3            N           1411           735   open
1          ADJ            421           271   open
2          ADV            337           103   open

> x[order(x$Class, -x$TokenFrequency), ]
```

```
  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
3            N           1411           735   open
1          ADJ            421           271   open
2          ADV            337           103   open

> x[order(x[, 4], -x[, 2]), ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
3            N           1411           735   open
1          ADJ            421           271   open
2          ADV            337           103   open

> x

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> dim(x)

[1] 5 4

> (no.of.rows <- dim(x)[1])

[1] 5

> (no.of.columns <- dim(x)[2])

[1] 4

> (ordering.index <- sample(no.of.rows))

[1] 2 5 1 4 3

> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
2          ADV            337           103   open
5         PREP            455            37 closed
1          ADJ            421           271   open
4         CONJ            458            18 closed
3            N           1411           735   open

> x[sample(dim(x)[1]), ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
2          ADV            337           103   open
3            N           1411           735   open
1          ADJ            421           271   open
5         PREP            455            37 closed
4         CONJ            458            18 closed
```

```
> x$Class

[1] open   open   open   closed closed
Levels: closed open

> sort(x$Class)

[1] closed closed open   open   open
Levels: closed open

> x$PartOfSpeech

[1] ADJ  ADV  N    CONJ PREP
Levels: ADJ ADV CONJ N PREP

> sort(x$PartOfSpeech)

[1] ADJ  ADV  CONJ N    PREP
Levels: ADJ ADV CONJ N PREP

> x$PartOfSpeech

[1] ADJ  ADV  N    CONJ PREP
Levels: ADJ ADV CONJ N PREP

> rank(x$PartOfSpeech)

[1] 1 2 4 3 5

> order(rank(x$PartOfSpeech))

[1] 1 2 4 3 5

> x$PartOfSpeech[order(rank(x$PartOfSpeech))]

[1] ADJ  ADV  CONJ N    PREP
Levels: ADJ ADV CONJ N PREP

> ordering.index <- order(rank(x$PartOfSpeech))
> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
4         CONJ            458            18 closed
3            N           1411           735   open
5         PREP            455            37 closed

> x$PartOfSpeech

[1] ADJ  ADV  N    CONJ PREP
Levels: ADJ ADV CONJ N PREP

> rank(x$PartOfSpeech)

[1] 1 2 4 3 5

> -rank(x$PartOfSpeech)
```

```
[1] -1 -2 -4 -3 -5

> x$PartOfSpeech[order(-rank(x$PartOfSpeech))]

[1] PREP N    CONJ ADV  ADJ
Levels: ADJ ADV CONJ N PREP

> sort(x$PartOfSpeech, decreasing = T)

[1] PREP N    CONJ ADV  ADJ
Levels: ADJ ADV CONJ N PREP

> ordering.index <- order(-rank(x$PartOfSpeech))
> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
5         PREP            455            37 closed
3            N           1411           735   open
4         CONJ            458            18 closed
2          ADV            337           103   open
1          ADJ            421           271   open

> ordering.index <- order(rank(x$Class), rank(x$PartOfSpeech))
> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
4         CONJ            458            18 closed
5         PREP            455            37 closed
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open

> ordering.index <- order(-rank(x$Class), -rank(x$PartOfSpeech))
> x[ordering.index, ]

  PartOfSpeech TokenFrequency TypeFrequency  Class
3            N           1411           735   open
2          ADV            337           103   open
1          ADJ            421           271   open
5         PREP            455            37 closed
4         CONJ            458            18 closed
```

4 Lists

```
> rm(list = ls(all = T))
>
> (a.vector <- c(1:10))

 [1]  1  2  3  4  5  6  7  8  9 10

> (a.dataframe <- read.table("dataframe.txt", header = T, sep = "\t",
+     comment.char = ""))  # insert here the data frame you saved above
```

```
  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> (another.vector <- c("This", "may", "be", "a", "sentence", "from",
+     "a", "corpus", "file", "."))

 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> (a.list <- list(a.vector, a.dataframe, another.vector))

[[1]]
 [1]  1  2  3  4  5  6  7  8  9 10

[[2]]
  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

[[3]]
 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> str(a.list)

List of 3
 $ : int [1:10] 1 2 3 4 5 6 7 8 9 10
 $ :'data.frame': 5 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4 3 5
  ..$ TokenFrequency: int [1:5] 421 337 1411 458 455
  ..$ TypeFrequency : int [1:5] 271 103 735 18 37
  ..$ Class         : Factor w/ 2 levels "closed","open": 2 2 2 1 1
 $ : chr [1:10] "This" "may" "be" "a" ...

> names(a.list) <- c("Part1", "Part2", "Part3")
> a.list

$Part1
 [1]  1  2  3  4  5  6  7  8  9 10

$Part2
  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed
```

```
$Part3
 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> (a.list <- list(Part1 = a.vector, Part2 = a.dataframe, Part3 = another.vector))

$Part1
 [1]  1  2  3  4  5  6  7  8  9 10

$Part2
  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

$Part3
 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> a.list[[1]]

 [1]  1  2  3  4  5  6  7  8  9 10

> a.list[[2]]

  PartOfSpeech TokenFrequency TypeFrequency  Class
1          ADJ            421           271   open
2          ADV            337           103   open
3            N           1411           735   open
4         CONJ            458            18 closed
5         PREP            455            37 closed

> a.list[[3]]

 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> a.list[1]

$Part1
 [1]  1  2  3  4  5  6  7  8  9 10

> is.list(a.list[1])

[1] TRUE

> is.list(a.list[[1]])

[1] FALSE

> is.vector(a.list[[1]])

[1] TRUE

> a.list$Part1
```

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

> a.list[["Part1"]]

 [1]  1  2  3  4  5  6  7  8  9 10

> a.list["Part1"]

$Part1
 [1]  1  2  3  4  5  6  7  8  9 10

> a.list[c(1, 3)]

$Part1
 [1]  1  2  3  4  5  6  7  8  9 10


$Part3
 [1] "This"     "may"      "be"       "a"        "sentence" "from"
 [7] "a"        "corpus"   "file"     "."

> a.list[[1]][3]

[1] 3

> a.list[[1]][3:5]

[1] 3 4 5

> a.list[[1]][c(3, 5)]

[1] 3 5

> a.list[[2]][3, 2]

[1] 1411

> a.list[[2]][3, 2:4]

  TokenFrequency TypeFrequency Class
3           1411           735  open

> a.list[[2]][3, c(2, 4)]

  TokenFrequency Class
3           1411  open

> x <- a.list[[2]]
> y <- split(x, x$Class)
> str(y)

List of 2
 $ closed:'data.frame': 2 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 3 5
  ..$ TokenFrequency: int [1:2] 458 455
  ..$ TypeFrequency : int [1:2] 18 37
  ..$ Class         : Factor w/ 2 levels "closed","open": 1 1
 $ open  :'data.frame': 3 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1 2 4
  ..$ TokenFrequency: int [1:3] 421 337 1411
  ..$ TypeFrequency : int [1:3] 271 103 735
  ..$ Class         : Factor w/ 2 levels "closed","open": 2 2 2
```

```
> y$open

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open
2          ADV            337           103  open
3            N           1411           735  open

> y <- split(x, x$PartOfSpeech)
> str(y)

List of 5
 $ ADJ :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1
  ..$ TokenFrequency: int 421
  ..$ TypeFrequency : int 271
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ ADV :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 2
  ..$ TokenFrequency: int 337
  ..$ TypeFrequency : int 103
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ CONJ:'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 3
  ..$ TokenFrequency: int 458
  ..$ TypeFrequency : int 18
  ..$ Class         : Factor w/ 2 levels "closed","open": 1
 $ N   :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 4
  ..$ TokenFrequency: int 1411
  ..$ TypeFrequency : int 735
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ PREP:'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 5
  ..$ TokenFrequency: int 455
  ..$ TypeFrequency : int 37
  ..$ Class         : Factor w/ 2 levels "closed","open": 1

> y$ADJ

  PartOfSpeech TokenFrequency TypeFrequency Class
1          ADJ            421           271  open

> y$N

  PartOfSpeech TokenFrequency TypeFrequency Class
3            N           1411           735  open

> y$PREP

  PartOfSpeech TokenFrequency TypeFrequency  Class
5         PREP            455            37 closed

> y <- split(x, list(x$Class, x$PartOfSpeech))
> str(y)
```

```
List of 10
 $ closed.ADJ :'data.frame': 0 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..:
  ..$ TokenFrequency: int(0)
  ..$ TypeFrequency : int(0)
  ..$ Class         : Factor w/ 2 levels "closed","open":
 $ open.ADJ   :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 1
  ..$ TokenFrequency: int 421
  ..$ TypeFrequency : int 271
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ closed.ADV :'data.frame': 0 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..:
  ..$ TokenFrequency: int(0)
  ..$ TypeFrequency : int(0)
  ..$ Class         : Factor w/ 2 levels "closed","open":
 $ open.ADV   :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 2
  ..$ TokenFrequency: int 337
  ..$ TypeFrequency : int 103
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ closed.CONJ:'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 3
  ..$ TokenFrequency: int 458
  ..$ TypeFrequency : int 18
  ..$ Class         : Factor w/ 2 levels "closed","open": 1
 $ open.CONJ  :'data.frame': 0 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..:
  ..$ TokenFrequency: int(0)
  ..$ TypeFrequency : int(0)
  ..$ Class         : Factor w/ 2 levels "closed","open":
 $ closed.N   :'data.frame': 0 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..:
  ..$ TokenFrequency: int(0)
  ..$ TypeFrequency : int(0)
  ..$ Class         : Factor w/ 2 levels "closed","open":
 $ open.N     :'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 4
  ..$ TokenFrequency: int 1411
  ..$ TypeFrequency : int 735
  ..$ Class         : Factor w/ 2 levels "closed","open": 2
 $ closed.PREP:'data.frame': 1 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..: 5
  ..$ TokenFrequency: int 455
  ..$ TypeFrequency : int 37
  ..$ Class         : Factor w/ 2 levels "closed","open": 1
 $ open.PREP  :'data.frame': 0 obs. of  4 variables:
  ..$ PartOfSpeech  : Factor w/ 5 levels "ADJ","ADV","CONJ",..:
  ..$ TokenFrequency: int(0)
  ..$ TypeFrequency : int(0)
  ..$ Class         : Factor w/ 2 levels "closed","open":
```

5 Character / String Processing

```
> example.1 <- c("I", "do", "not", "know")
> nchar(example.1)

[1] 1 2 3 4

> substr("internationalization", 6, 13)

[1] "national"

> substr(example.1, 2, 3)

[1] ""   "o"  "ot" "no"

> some.first.vector <- c("abcd", "efgh")
> some.other.vector <- c("ijkl", "mnop")
> substr(c(some.first.vector, some.other.vector), c(1, 2, 3, 4), c(2,
+     3, 4, 4))

[1] "ab" "fg" "kl" "p"

> tolower(example.1)

[1] "i"    "do"   "not"  "know"

> toupper(example.1)

[1] "I"    "DO"   "NOT"  "KNOW"

> chartr("o", "x", example.1)

[1] "I"    "dx"   "nxt"  "knxw"

> paste("I", "do", "not", "know", sep = " ")

[1] "I do not know"

> paste("I", "do", "not", "know")

[1] "I do not know"

> paste("I", "do", "not", "know", collapse = " ")

[1] "I do not know"

> paste("I", "do", "not", "know", sep = " ", collapse = " ")

[1] "I do not know"

> paste(example.1, sep = " ")

[1] "I"    "do"   "not"  "know"

> paste(example.1)

[1] "I"    "do"   "not"  "know"

> paste(example.1, collapse = " ")
```

```
[1] "I do not know"

> (list.1 <- as.list(example.1))

[[1]]
[1] "I"

[[2]]
[1] "do"

[[3]]
[1] "not"

[[4]]
[1] "know"

> paste(list.1, sep = " ")

[1] "I"    "do"   "not"  "know"

> paste(list.1)

[1] "I"    "do"   "not"  "know"

> paste(list.1, collapse = " ")

[1] "I do not know"

> example.2 <- "I do not know"
> strsplit(example.2, " ")

[[1]]
[1] "I"    "do"   "not"  "know"

> unlist(strsplit(example.2, " "))

[1] "I"    "do"   "not"  "know"

> strsplit(example.2, "")

[[1]]
 [1] "I" " " " " "d" "o" " " " " "n" "o" "t" " " " " "k" "n" "o" "w"

> example.3 <- c("Hello Elmo", "Bye bye binky")
> example.3

[1] "Hello Elmo"    "Bye bye binky"

> strsplit(example.3, " ")

[[1]]
[1] "Hello" "Elmo"

[[2]]
[1] "Bye"   "bye"   "binky"

> text.1 <- "This is the first sentence. This is the second sentence."
> strsplit(text.1, ".")
```

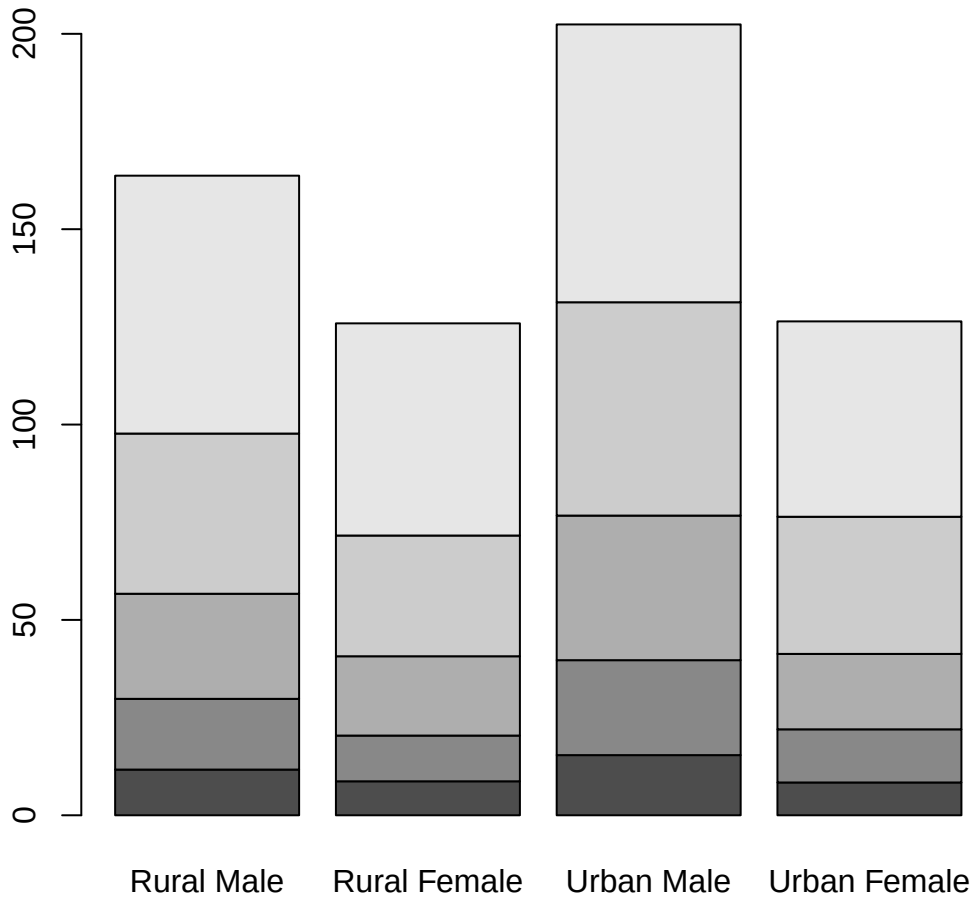
```
[[1]]
 [1] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
[24] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
[47] "" "" "" "" "" "" "" "" "" ""

> text.1.sentences <- strsplit(text.1, "\\. ")
> text.1.sentences

[[1]]
[1] "This is the first sentence"   "This is the second sentence."

> text.1.sentences <- unlist(text.1.sentences)
> text.1.sentences

[1] "This is the first sentence"   "This is the second sentence."

> text.1.words <- strsplit(text.1.sentences, " ")
> text.1.words

[[1]]
[1] "This"     "is"       "the"      "first"    "sentence"

[[2]]
[1] "This"      "is"        "the"       "second"    "sentence."

> text.1.words <- unlist(text.1.words)
> text.1.words

 [1] "This"      "is"        "the"       "first"     "sentence"
 [6] "This"      "is"        "the"       "second"    "sentence."

> which(text.1.words == "")

integer(0)

> which(text.1.words == "the")

[1] 3 8

> str(list.1)

List of 4
 $ : chr "I"
 $ : chr "do"
 $ : chr "not"
 $ : chr "know"

> strsplit(list.1, " ")

Error:  non-character argument

> strsplit(list.1, "")

Error:  non-character argument

> str(unlist(list.1))

 chr [1:4] "I" "do" "not" "know"
```

```
> strsplit(unlist(list.1), " ")

[[1]]
[1] "I"

[[2]]
[1] "do"

[[3]]
[1] "not"

[[4]]
[1] "know"

> strsplit(unlist(list.1), "")

[[1]]
[1] "I"

[[2]]
[1] "d" "o"

[[3]]
[1] "n" "o" "t"

[[4]]
[1] "k" "n" "o" "w"
```

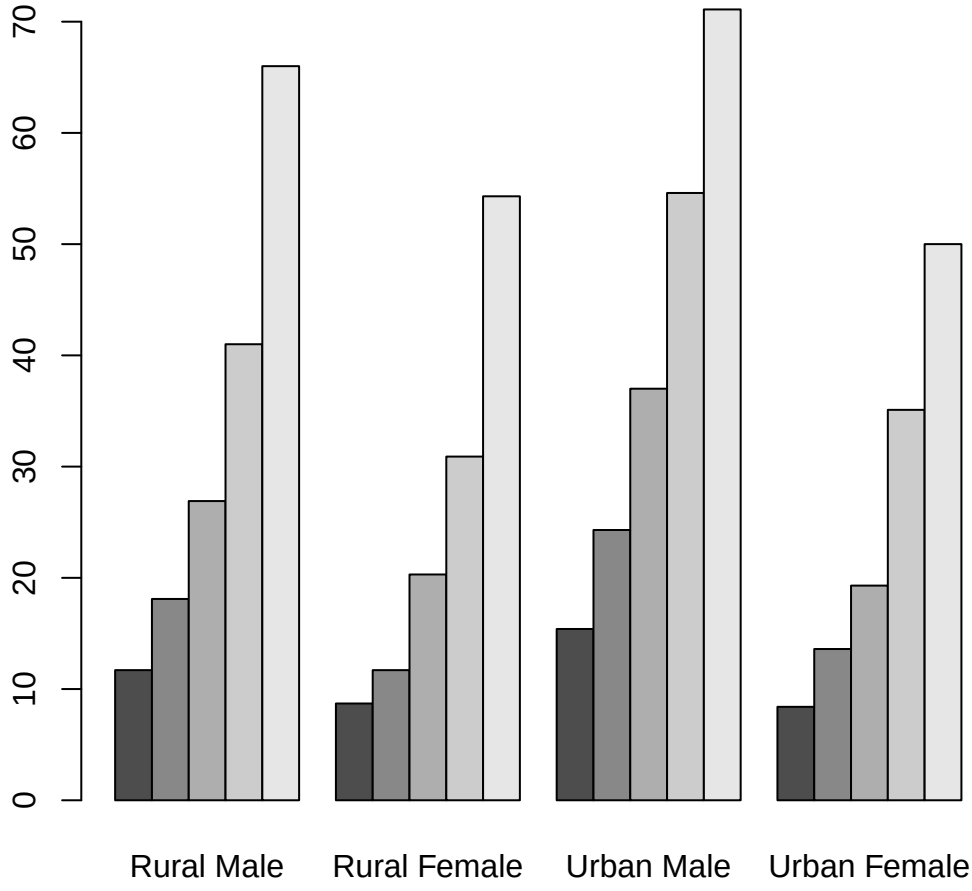
6 More graphics

```
> VADeaths

      Rural Male Rural Female Urban Male Urban Female
50-54       11.7          8.7       15.4          8.4
55-59       18.1         11.7       24.3         13.6
60-64       26.9         20.3       37.0         19.3
65-69       41.0         30.9       54.6         35.1
70-74       66.0         54.3       71.1         50.0

> # ?VADeaths
```

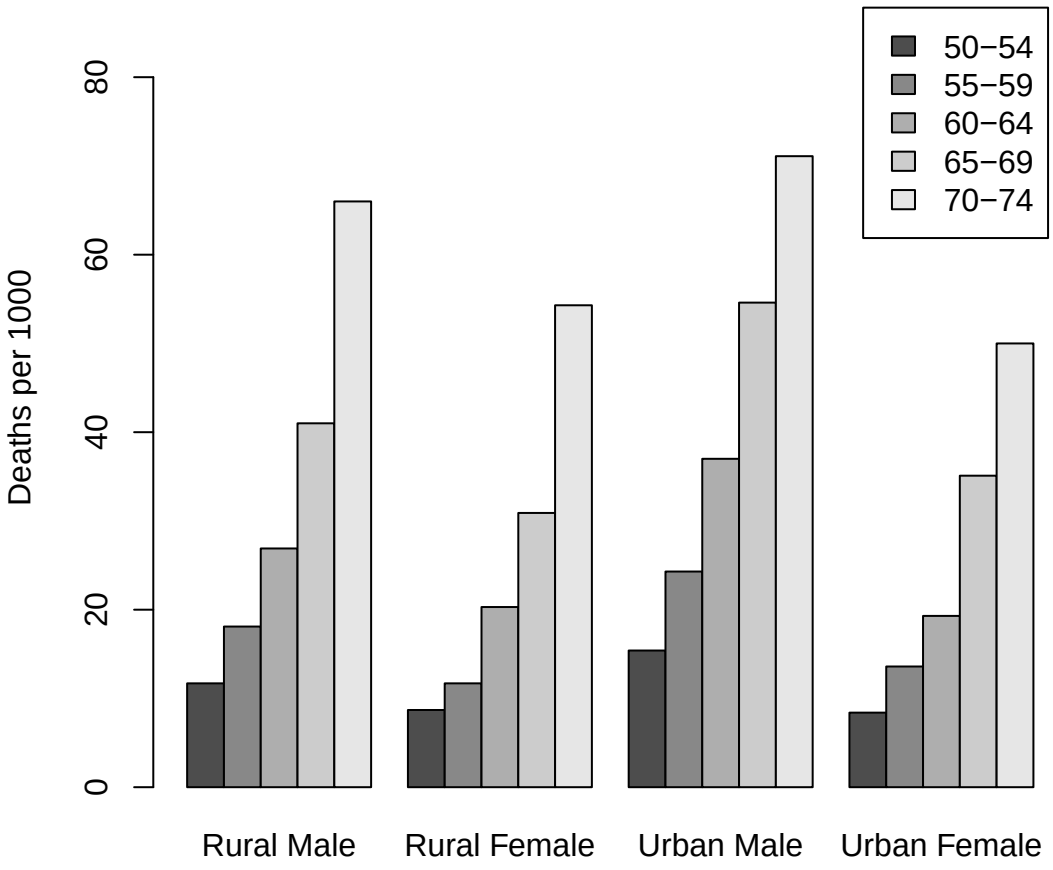
```
> barplot(VADeaths)
```



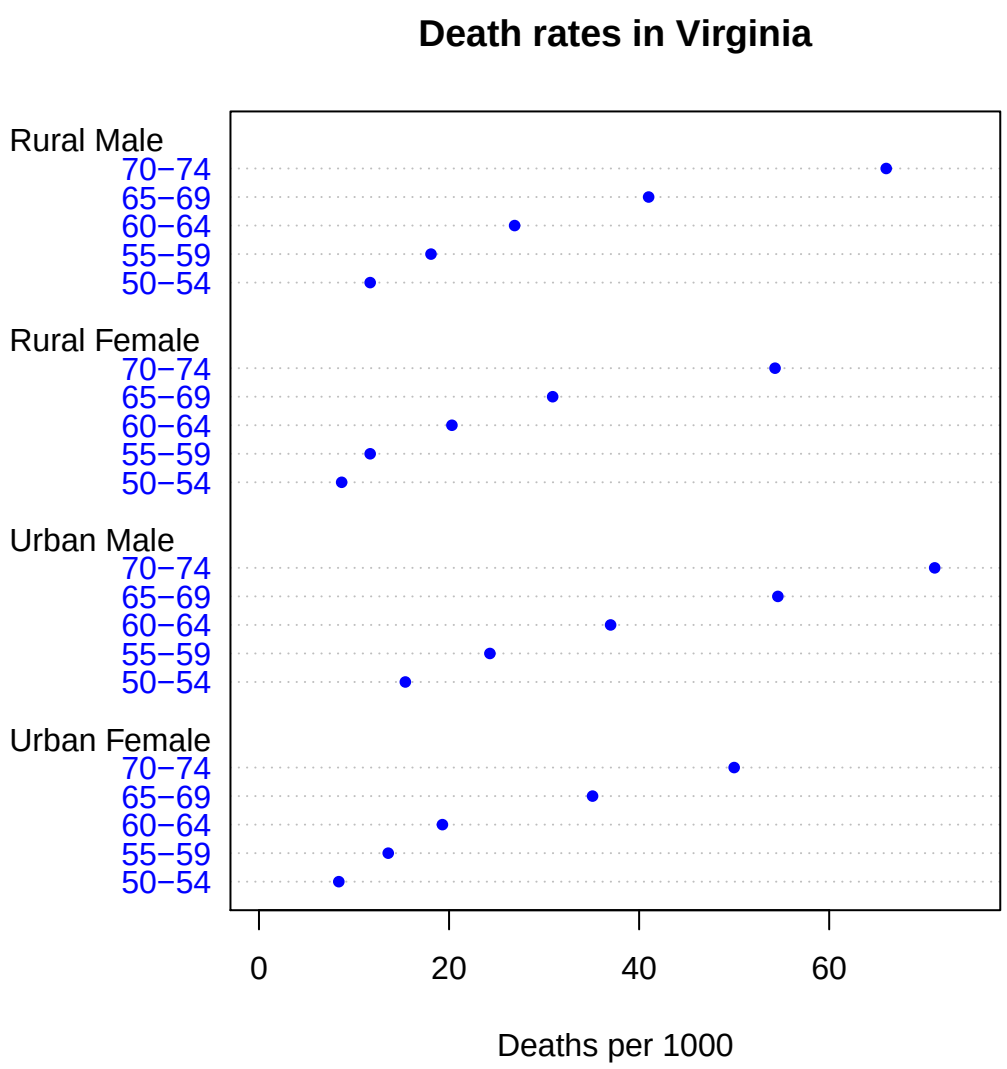

```
> barplot(VADeaths, beside = TRUE)
```




```
> barplot(VADeaths, beside = TRUE, legend = TRUE, ylim = c(0, 90), ylab = "Deaths per 1000",
+     main = "Death rates in Virginia")
```

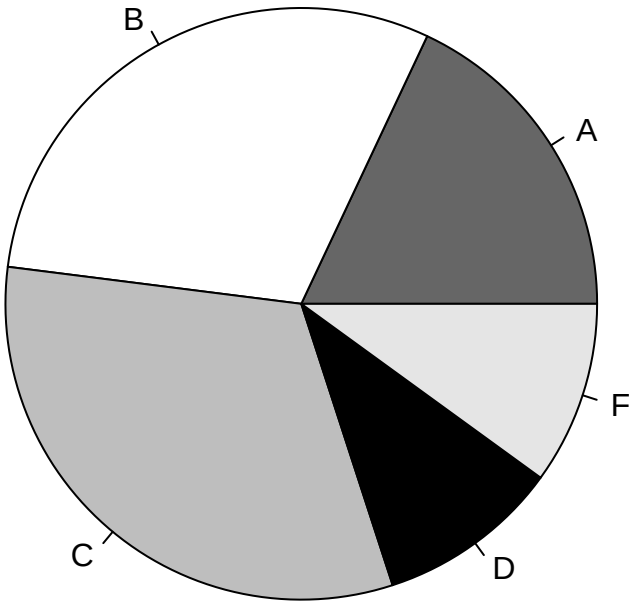



```
> dotchart(VADeaths, xlim = c(0, 75), xlab = "Deaths per 1000", main = "Death rates in Virginia",
+     pch = 20, col = "blue")
```




```
> groupsizes <- c(18, 30, 32, 10, 10)
> grades <- c("A", "B", "C", "D", "F")
```

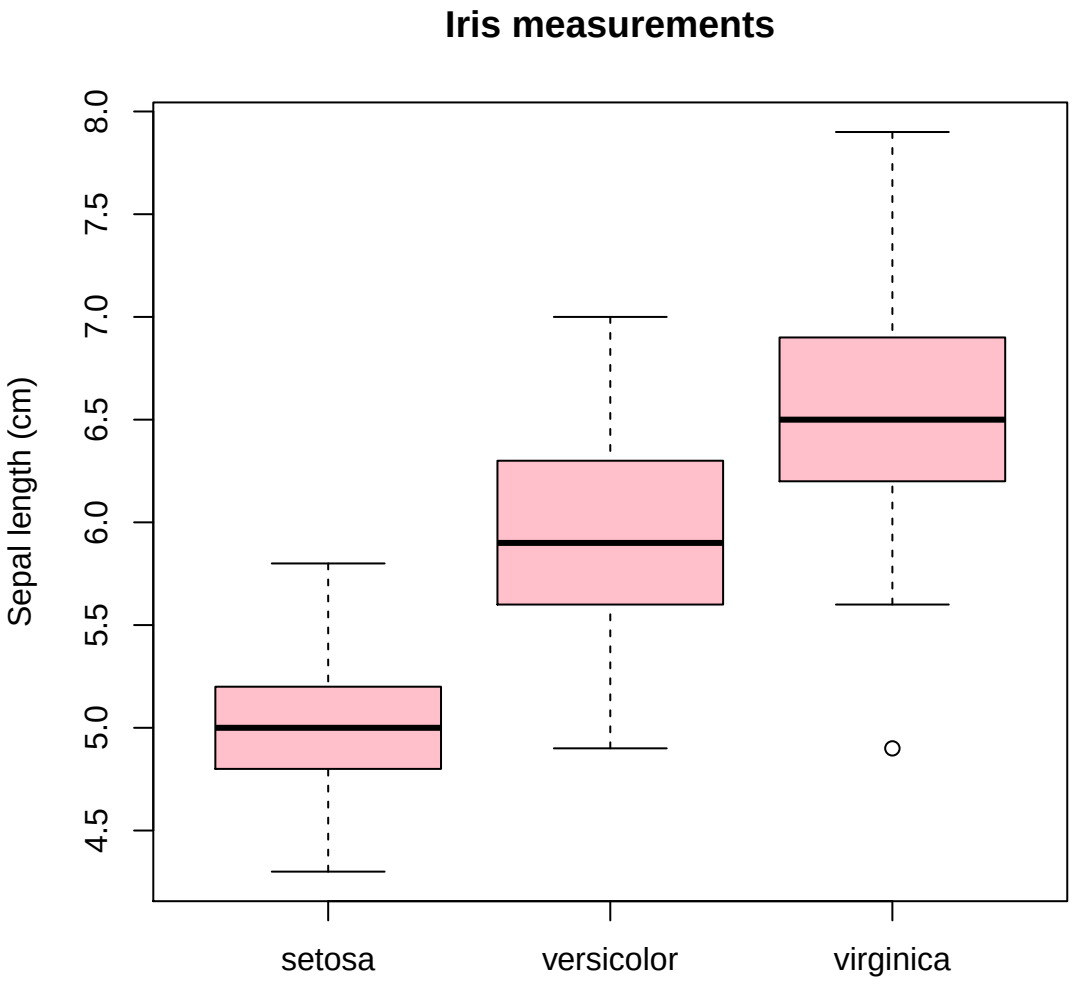
```
> pie(groupsizes, grades, col = c("grey40", "white", "grey", "black",
+     "grey90"))
```




```
> # ?iris
> head(iris)

  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
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa

> str(iris)

'data.frame': 150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
```

```
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

> summary(iris)

  Sepal.Length   Sepal.Width    Petal.Length   Petal.Width
 Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1
 1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3
 Median :5.80   Median :3.00   Median :4.35   Median :1.3
 Mean   :5.84   Mean   :3.06   Mean   :3.76   Mean   :1.2
 3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8
 Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

```
> boxplot(Sepal.Length ~ Species, data = iris, ylab = "Sepal length (cm)",
+     main = "Iris measurements", col = "pink")
```



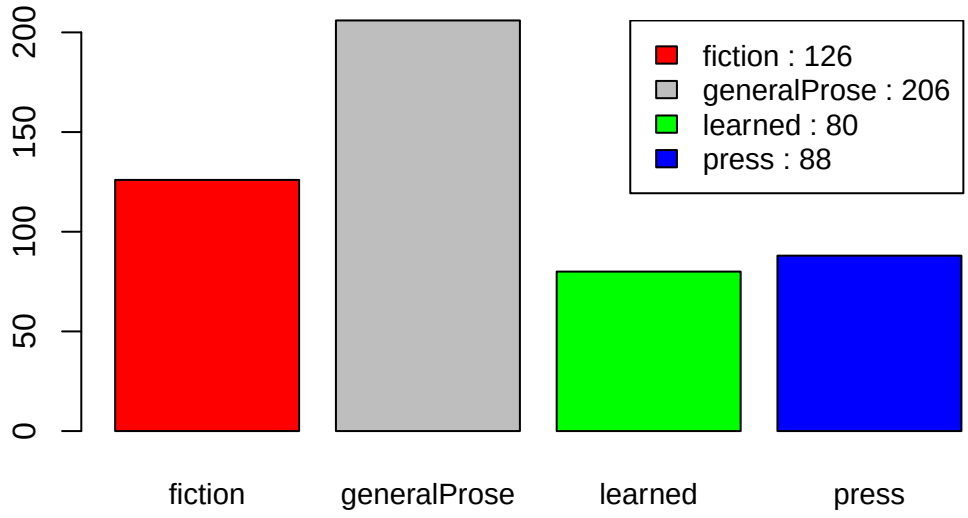

```
> # ?boxplot
```

7 The Brown Corpus

We introduce the tagged version of the Brown Corpus (Francis and Kucera 1979) in this section:

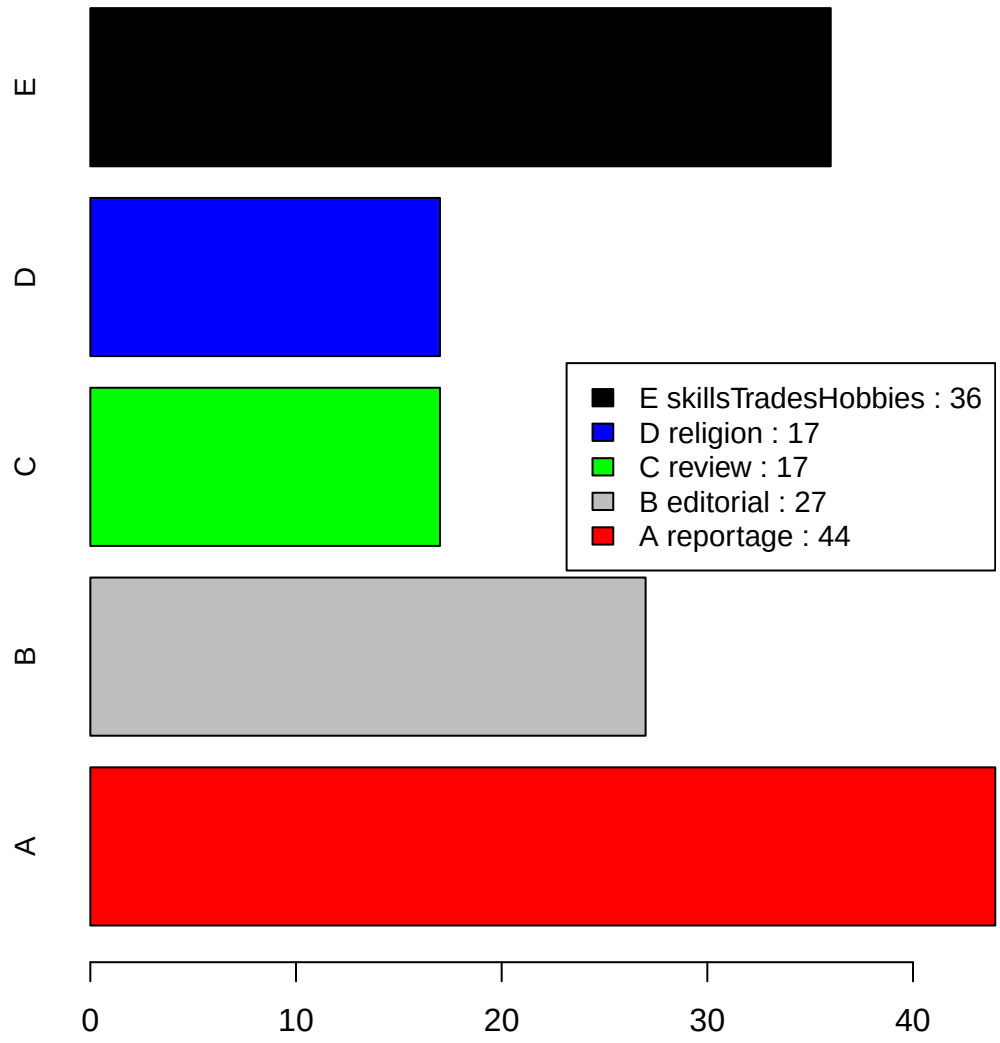
```
> brown.corpus.structure <- read.table(file = "brown_text_categories.txt",
+     header = T, sep = "\t", comment.char = "")
>
> str(brown.corpus.structure)

'data.frame': 15 obs. of  5 variables:
 $ ProseType        : Factor w/ 2 levels "imaginative",..: 2 2 2 2 2 2 2 2 2 1 ...
 $ GenreGroup       : Factor w/ 4 levels "fiction","generalProse",..: 4 4 4 2 2 2 2 2 3 1 ...
 $ Category         : Factor w/ 15 levels "A","B","C","D",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
 $ ContentOfCategory: Factor w/ 15 levels "adventureWestern",..: 10 3 11 9 15 8 2 6 13 4 ...
 $ NumberOfTexts    : int  44 27 17 17 36 48 75 30 80 29 ...

> brown.corpus.structure

     ProseType   GenreGroup Category              ContentOfCategory
1  informative        press        A                      reportage
2  informative        press        B                      editorial
3  informative        press        C                         review
4  informative generalProse        D                       religion
5  informative generalProse        E             skillsTradesHobbies
6  informative generalProse        F                     popularLore
7  informative generalProse        G belleslettresBiographiesEssays
8  informative generalProse        H                    miscellaneous
9  informative      learned        J                          science
10 imaginative      fiction        K                  generalFiction
11 imaginative      fiction        L         misteryDetectiveFiction
12 imaginative      fiction        M                  scienceFiction
13 imaginative      fiction        N                adventureWestern
14 imaginative      fiction        P                romanceLoveStory
15 imaginative      fiction        R                           humor
   NumberOfTexts
1             44
2             27
3             17
4             17
5             36
6             48
7             75
8             30
9             80
10            29
11            24
12             6
13            29
14            29
15             9

> # View(brown.corpus.structure)
>
> (number.of.texts.by.genre <- tapply(brown.corpus.structure$NumberOfTexts,
+     brown.corpus.structure$GenreGroup, sum))

     fiction generalProse      learned        press
         126          206           80           88

>
> barplot(number.of.texts.by.genre, cex.names = 0.9, col = c("red",
+     "gray", "green", "blue"))
> legend(3, max(number.of.texts.by.genre), paste(names(number.of.texts.by.genre),
+     ":", number.of.texts.by.genre), cex = 0.9, fill = c("red", "gray",
+     "green", "blue"))
```

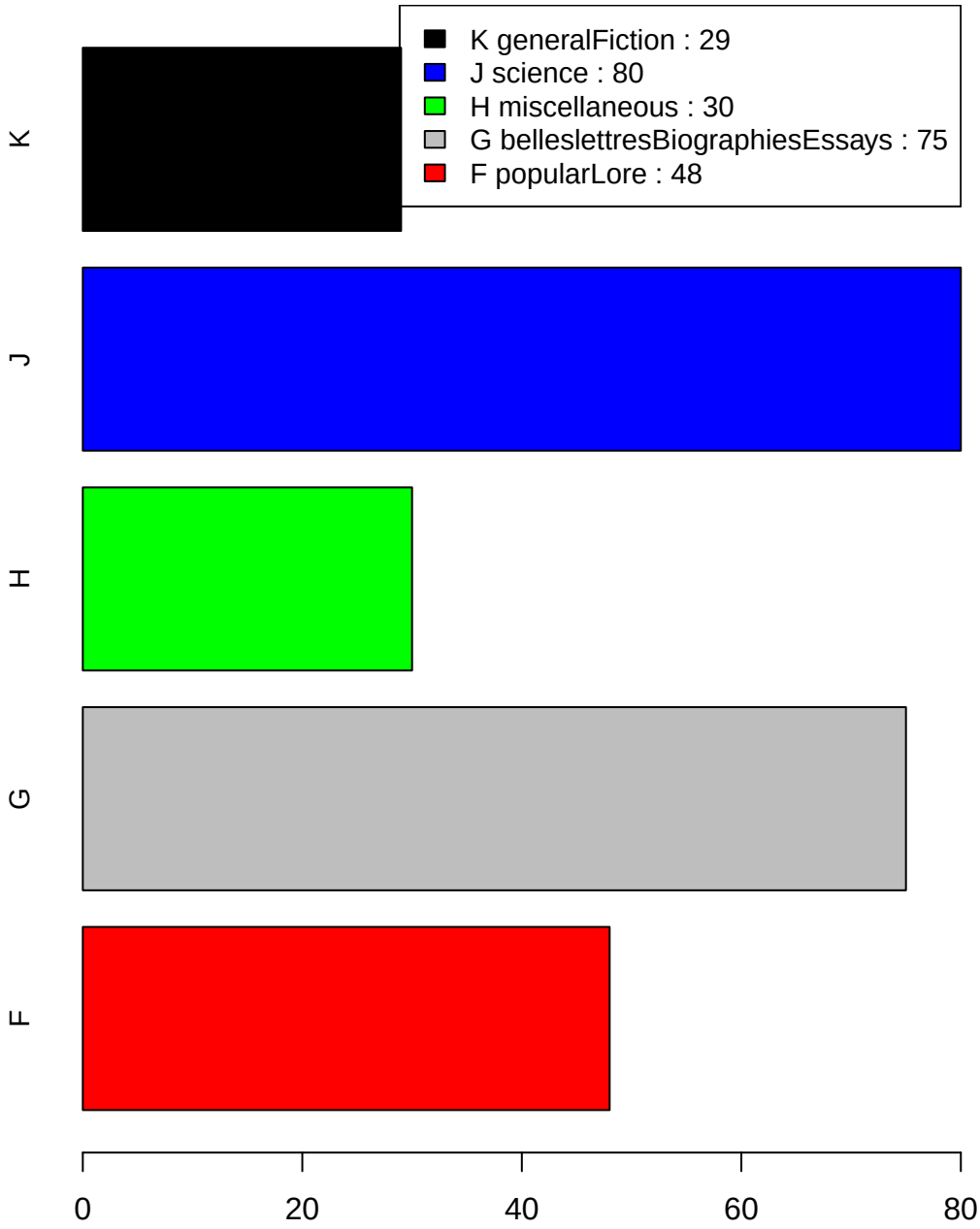



```
>
> (number.of.texts.by.category <- tapply(brown.corpus.structure$NumberOfTexts,
+     brown.corpus.structure$Category, sum))

 A  B  C  D  E  F  G  H  J  K  L  M  N  P  R
44 27 17 17 36 48 75 30 80 29 24  6 29 29  9
```

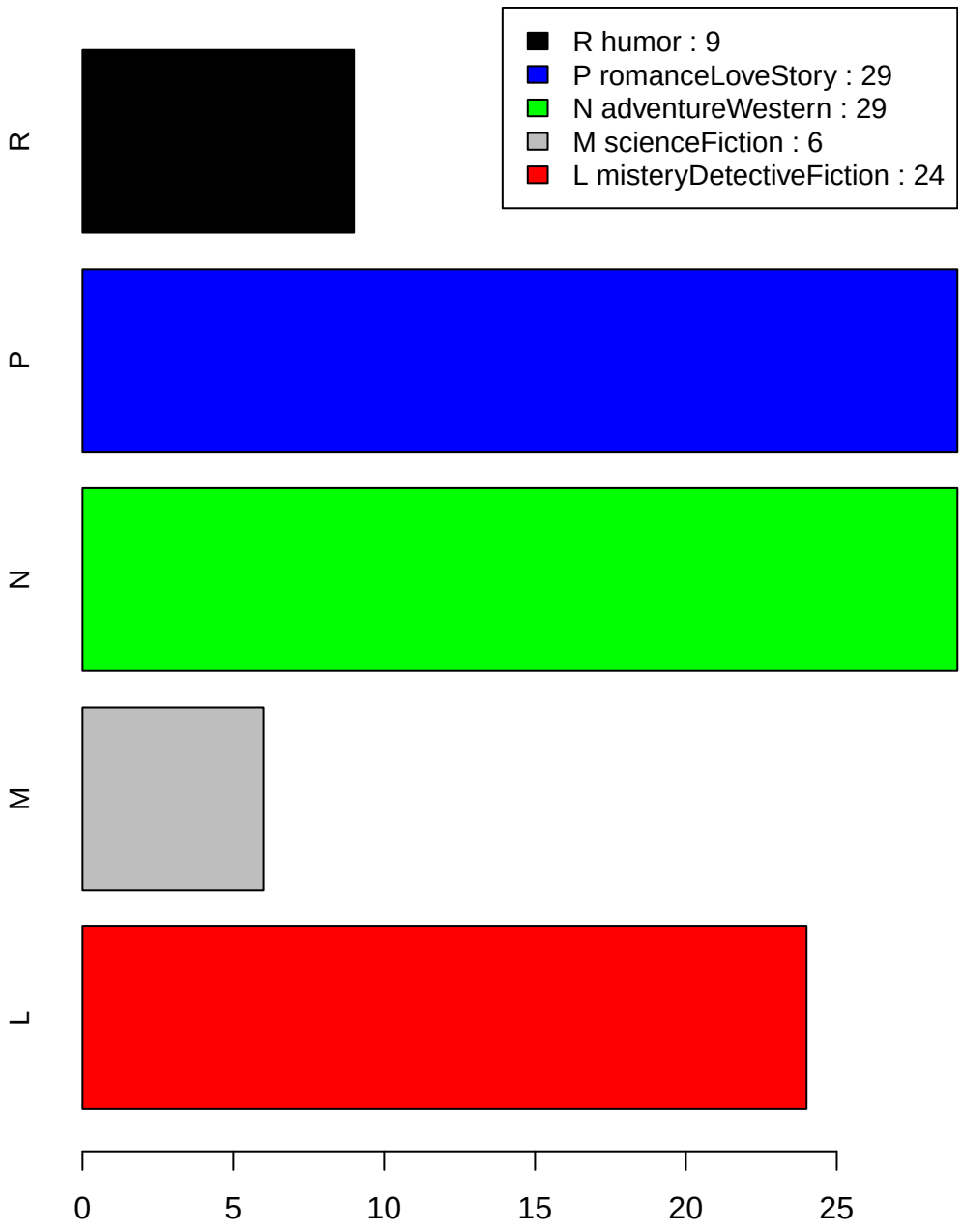
```
> barplot(number.of.texts.by.category[1:5], horiz = TRUE, cex.names = 0.9,
+     col = c("red", "gray", "green", "blue", "black"))
> legend("right", paste(rev(brown.corpus.structure$Category[1:5]), rev(brown.corpus.structure$ContentOf(
+     ":", rev(number.of.texts.by.category[1:5])), cex = 0.9, fill = rev(c("red",
+     "gray", "green", "blue", "black")))
```




```
> barplot(number.of.texts.by.category[6:10], horiz = TRUE, cex.names = 0.9,
+     col = c("red", "gray", "green", "blue", "black"))
> legend("topright", paste(rev(brown.corpus.structure$Category[6:10]),
+     rev(brown.corpus.structure$ContentOfCategory[6:10]), ":", rev(number.of.texts.by.category[6:10])),
+     cex = 0.9, fill = rev(c("red", "gray", "green", "blue", "black")))
```




```
> barplot(number.of.texts.by.category[11:15], horiz = TRUE, cex.names = 0.9,
+     col = c("red", "gray", "green", "blue", "black"))
> legend("topright", paste(rev(brown.corpus.structure$Category[11:15]),
+     rev(brown.corpus.structure$ContentOfCategory[11:15]), ":", rev(number.of.texts.by.category[11:15])
+     cex = 0.9, fill = rev(c("red", "gray", "green", "blue", "black")))
```




```
> brown.tagged <- scan(file = "BTL.TXT", what = "char", sep = "\n",
+     comment.char = "")
> system.time(brown.tagged <- scan(file = "BTL.TXT", what = "char",
```

```
+     sep = "\n", comment.char = ""))

   user  system elapsed
  0.608   0.004   0.614

> head(brown.tagged)

[1] "|SA01:1 the_AT Fulton_NP County_NN Grand_JJ Jury_NN said_VBD Friday_NR an_AT investigation_NN of_IN
[2] "|SA01:2 the_AT jury_NN further_RBR said_VBD in_IN term-end_NN presentments_NNS that_CS the_AT City_
[3] "|SA01:3 the_AT September-October_NP term_NN jury_NN had_HVD been_BEN charged_VBN by_IN Fulton_NP Su
[4] "|SA01:4 only_RB a_AT relative_JJ handful_NN of_IN such_JJ reports_NNS was_BEDZ received_VBN ,_, the
[5] "|SA01:5 the_AT jury_NN said_VBD it_PPS did_DOD find_VB that_CS many_AP of_IN Georgia's_NP$ registra
[6] "|SA01:6 it_PPS recommended_VBD that_CS Fulton_NP legislators_NNS act_VB to_TO have_HV these_DTS law

> tail(brown.tagged)

[1] "|SR09:93 as_CS you_PPSS can_MD count_VB on_IN me_PPO to_TO do_DO the_AT same_AP ._."
[2] "|SR09:94 compassionately_RB yours_PP$$ ,_, S._NP J._NP Perelman_NP revulsion_NN in_IN the_AT desert
[3] "|SR09:95 the_AT doors_NNS of_IN the_AT J_NP train_NN slid_VBD shut_VBN ,_, and_CC as_CS I_PPSS drop
[4] "|SR09:96 she_PPS was_BEDZ a_AT living_VBG doll_NN and_CC no_AT mistake_NN --_-- the_AT blue-black_J
[5] "|SR09:97 from_IN what_WDT I_PPSS was_BEDZ able_JJ to_IN gauge_NN in_IN a_AT swift_JJ ,_, greedy_JJ
[6] "|S:1 ._"

> length(brown.tagged)

[1] 57067

> brown.tag.set <- read.delim(file = "brown_tag_set.txt", header = T,
+     sep = "\t", comment.char = "")
> # View(brown.tag.set)
> head(brown.tag.set)

  Tag          Description  Examples
1 --                  dash         --
2   ( opening parenthesis          (
3   ) closing parenthesis          )
4   *             negator    not n't
5   ,               comma          ,
6   . sentence terminator  . ? ; ! :

> tail(brown.tag.set)

         Tag                                                  Description
221   WRB+DO   WH-adverb + verb "to do", present, not 3rd person singular
222  WRB+DOD                          WH-adverb + verb "to do", past tense
223 WRB+DOD*                WH-adverb + verb "to do", past tense, negated
224  WRB+DOZ WH-adverb + verb "to do", present tense, 3rd person singular
225   WRB+IN                                    WH-adverb + preposition
226   WRB+MD                                 WH-adverb + modal auxillary
        Examples
221         howda
222 where'd how'd
223        whyn't
224         how's
225         why'n
226       where'd
```

```
> dim(brown.tag.set)

[1] 226   3

> str(brown.tag.set)

'data.frame': 226 obs. of  3 variables:
 $ Tag        : Factor w/ 226 levels "-- ",",",",":",..: 1 5 6 7 2 4 3 8 9 10 ...
 $ Description: Factor w/ 225 levels "adjective ","adjective, comparative ",..: 24 127 19 101 21 158 20
 $ Examples   : Factor w/ 220 levels "--",",",",":",". ? ; ! :",..: 1 5 6 129 2 4 3 143 12 29 ...

> # brown.tag.set
> brown.words.1 <- strsplit(brown.tagged, " ")
> head(brown.words.1)

[[1]]
 [1] "|SA01:1"            "the_AT"             "Fulton_NP"
 [4] "County_NN"          "Grand_JJ"           "Jury_NN"
 [7] "said_VBD"           "Friday_NR"          "an_AT"
[10] "investigation_NN"   "of_IN"              "Atlanta's_NP$"
[13] "recent_JJ"          "primary_NN"         "election_NN"
[16] "produced_VBD"       "no_AT"              "evidence_NN"
[19] "that_CS"            "any_DTI"            "irregularities_NNS"
[22] "took_VBD"           "place_NN"           "._."

[[2]]
 [1] "|SA01:2"          "the_AT"           "jury_NN"
 [4] "further_RBR"      "said_VBD"         "in_IN"
 [7] "term-end_NN"      "presentments_NNS" "that_CS"
[10] "the_AT"           "City_NN"          "Executive_JJ"
[13] "Committee_NN"     ",_,"              "which_WDT"
[16] "had_HVD"          "over-all_JJ"      "charge_NN"
[19] "of_IN"            "the_AT"           "election_NN"
[22] ",_,"              "deserves_VBZ"     "the_AT"
[25] "praise_NN"        "and_CC"           "thanks_NNS"
[28] "of_IN"            "the_AT"           "City_NN"
[31] "of_IN"            "Atlanta_NP"       "for_IN"
[34] "the_AT"           "manner_NN"        "in_IN"
[37] "which_WDT"        "the_AT"           "election_NN"
[40] "was_BEDZ"         "conducted_VBN"    "._."

[[3]]
 [1] "|SA01:3"             "the_AT"              "September-October_NP"
 [4] "term_NN"             "jury_NN"             "had_HVD"
 [7] "been_BEN"            "charged_VBN"         "by_IN"
[10] "Fulton_NP"           "Superior_JJ"         "Court_NN"
[13] "Judge_NN"            "Durwood_NP"          "Pye_NP"
[16] "to_TO"               "investigate_VB"      "reports_NNS"
[19] "of_IN"               "possible_JJ"         "irregularities_NNS"
[22] "in_IN"               "the_AT"              "hard-fought_JJ"
[25] "primary_NN"          "which_WDT"           "was_BEDZ"
[28] "won_VBN"             "by_IN"               "Mayor-nominate_NN"
[31] "Ivan_NP"             "Allen_NP"            "Jr._NP"
[34] "._."
```

```
[[4]]
 [1] "|SA01:4"        "only_RB"        "a_AT"           "relative_JJ"
 [5] "handful_NN"     "of_IN"          "such_JJ"        "reports_NNS"
 [9] "was_BEDZ"       "received_VBN"   ",_,"            "the_AT"
[13] "jury_NN"        "said_VBD"       ",_,"            "considering_IN"
[17] "the_AT"         "widespread_JJ"  "interest_NN"    "in_IN"
[21] "the_AT"         "election_NN"    ",_,"            "the_AT"
[25] "number_NN"      "of_IN"          "voters_NNS"     "and_CC"
[29] "the_AT"         "size_NN"        "of_IN"          "this_DT"
[33] "city_NN"        "._."

[[5]]
 [1] "|SA01:5"         "the_AT"          "jury_NN"
 [4] "said_VBD"        "it_PPS"          "did_DOD"
 [7] "find_VB"         "that_CS"         "many_AP"
[10] "of_IN"           "Georgia's_NP$"   "registration_NN"
[13] "and_CC"          "election_NN"     "laws_NNS"
[16] "are_BER"         "outmoded_JJ"     "or_CC"
[19] "inadequate_JJ"   "and_CC"          "often_RB"
[22] "ambiguous_JJ"    "._."

[[6]]
 [1] "|SA01:6"         "it_PPS"          "recommended_VBD"
 [4] "that_CS"         "Fulton_NP"       "legislators_NNS"
 [7] "act_VB"          "to_TO"           "have_HV"
[10] "these_DTS"       "laws_NNS"        "studied_VBN"
[13] "and_CC"          "revised_VBN"     "to_IN"
[16] "the_AT"          "end_NN"          "of_IN"
[19] "modernizing_VBG" "and_CC"          "improving_VBG"
[22] "them_PPO"        "._."

> tail(brown.words.1)

[[1]]
 [1] "|SR09:93" "as_CS"    "you_PPSS" "can_MD"   "count_VB" "on_IN"
 [7] "me_PPO"   "to_TO"    "do_DO"    "the_AT"   "same_AP"  "._."

[[2]]
 [1] "|SR09:94"           "compassionately_RB" "yours_PP$$"
 [4] ",_,"                "S._NP"              "J._NP"
 [7] "Perelman_NP"        "revulsion_NN"       "in_IN"
[10] "the_AT"             "desert_NN"

[[3]]
 [1] "|SR09:95"     "the_AT"       "doors_NNS"    "of_IN"
 [5] "the_AT"       "J_NP"         "train_NN"     "slid_VBD"
 [9] "shut_VBN"     ",_,"          "and_CC"       "as_CS"
[13] "I_PPSS"       "dropped_VBD"  "into_IN"      "a_AT"
[17] "seat_NN"      "and_CC"       ",_,"          "exhaling_VBG"
[21] ",_,"          "looked_VBD"   "up_RP"        "across_IN"
[25] "the_AT"       "aisle_NN"     ",_,"          "the_AT"
[29] "whole_JJ"     "aviary_NN"    "in_IN"        "my_PP$"
[33] "head_NN"      "burst_VBD"    "into_IN"      "song_NN"
[37] "._."
```

```
[[4]]
 [1] "|SR09:96"          "she_PPS"           "was_BEDZ"
 [4] "a_AT"              "living_VBG"        "doll_NN"
 [7] "and_CC"            "no_AT"             "mistake_NN"
[10] "--_--"             "the_AT"            "blue-black_JJ"
[13] "bang_NN"           ",_,"               "the_AT"
[16] "wide_JJ"           "cheekbones_NNS"    ",_,"
[19] "olive-flushed_JJ" ",_,"                "that_WPS"
[22] "betrayed_VBD"      "the_AT"            "Cherokee_NP"
[25] "strain_NN"         "in_IN"             "her_PP$"
[28] "Midwestern_JJ"     "lineage_NN"        ",_,"
[31] "and_CC"            "the_AT"            "mouth_NN"
[34] "whose_WP$"         "only_AP"           "fault_NN"
[37] ",_,"               "in_IN"             "the_AT"
[40] "novelist's_NN$"    "carping_VBG"       "phrase_NN"
[43] ",_,"               "was_BEDZ"          "that_CS"
[46] "the_AT"            "lower_JJR"         "lip_NN"
[49] "was_BEDZ"          "a_AT"              "trifle_NN"
[52] "too_QL"            "voluptuous_JJ"     "._."

[[5]]
 [1] "|SR09:97"          "from_IN"           "what_WDT"
 [4] "I_PPSS"            "was_BEDZ"          "able_JJ"
 [7] "to_IN"             "gauge_NN"          "in_IN"
[10] "a_AT"              "swift_JJ"          ",_,"
[13] "greedy_JJ"         "glance_NN"         ",_,"
[16] "the_AT"            "figure_NN"         "inside_IN"
[19] "the_AT"            "coral-colored_JJ" "boucle_NN"
[22] "dress_NN"          "was_BEDZ"          "stupefying_VBG"
[25] "._."

[[6]]
[1] "|S:1" "._"
```

```
> length(unlist(brown.words.1))
```

```
[1] 1194534
```

```
> words.per.sentence.1 <- sapply(brown.words.1, length)
> str(words.per.sentence.1)
```

```
 int [1:57067] 24 42 34 34 23 23 42 3 25 24 ...
```

```
> sum(words.per.sentence.1)
```

```
[1] 1194534
```

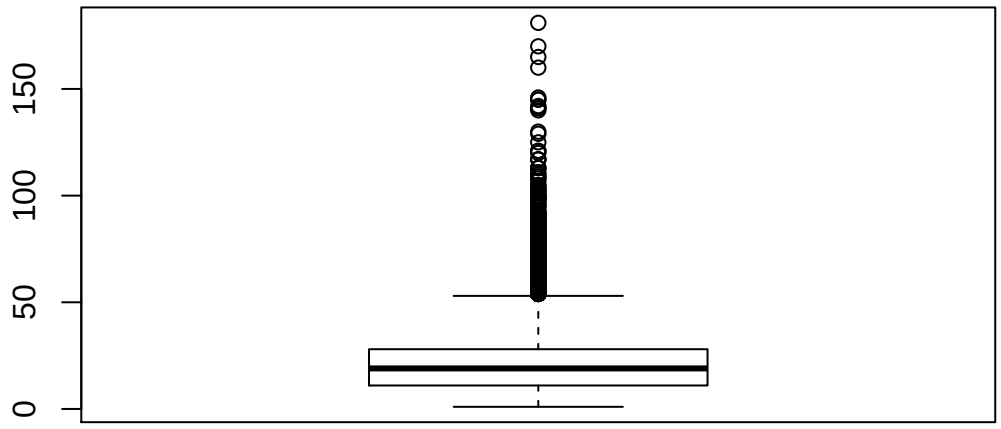
```
> mean(words.per.sentence.1)
```

```
[1] 20.93
```

```
> summary(words.per.sentence.1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    11.0    19.0    20.9    28.0   181.0
```

```
> boxplot(words.per.sentence.1)
```



References

Abelson, R.P. (1995). *Statistics as Principled Argument*. L. Erlbaum Associates.

Baayen, R. Harald (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Braun, J. and D.J. Murdoch (2007). *A First Course in Statistical Programming with R*. Cambridge University Press.

De Veaux, R.D. et al. (2005). *Stats: Data and Models*. Pearson Education, Limited.

Diez, D. et al. (2013). *OpenIntro Statistics: Second Edition*. CreateSpace Independent Publishing Platform. URL: http://www.openintro.org/stat/textbook.php.

Faraway, J.J. (2004). *Linear Models With R*. Chapman & Hall Texts in Statistical Science Series. Chapman & Hall/CRC.

Francis, W. N. and H. Kucera (1979). *Brown Corpus Manual*. Tech. rep. Department of Linguistics, Brown University, Providence, Rhode Island, US. URL: http://icame.uib.no/brown/bcm.html.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Gries, S.T. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Taylor & Francis.

— (2013). *Statistics for Linguistics with R: A Practical Introduction, 2nd Edition*. Mouton De Gruyter.

Johnson, K. (2008). *Quantitative methods in linguistics*. Blackwell Pub.

Kruschke, John K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier.

Miles, J. and M. Shevlin (2001). *Applying Regression and Correlation: A Guide for Students and Researchers*. SAGE Publications.

Wright, D.B. and K. London (2009). *Modern regression techniques using R: A practical guide for students and researchers*. SAGE.

Xie, Yihui (2013). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC.