

# Introduction to R

Martin T. Morgan\*

27-28 February 2014

## 1 R

---

### Exercise 1

*This exercise uses data describing 128 microarray samples as a basis for exploring R functions. Covariates such as age, sex, type, stage of the disease, etc., are in a data file `pData.csv`.*

*The following command creates a variable `pdataFiles` that is the location of a comma-separated value ('csv') file to be used in the exercise. A csv file can be created using, e.g., 'Save as...' in spreadsheet software.*

```
pdataFile <- "~/extdata/pData.csv"
```

*Make sure the file exists (you've specified the correct location) before continuing!*

```
stopifnot(file.exists(pdataFile))
```

*Input the csv file using `read.table`, assigning the input to a variable `pdata`. Use `dim` to find out the dimensions (number of rows, number of columns) in the object. Are there 128 rows? Use `names` or `colnames` to list the names of the columns of `pdata`. Use `summary` to summarize each column of the data. What are the data types of each column in the data frame?*

*A data frame is a list of equal length vectors. Select the 'sex' column of the data frame using `[[` or `$`. Pause to explain to your neighbor why this sub-setting works. Since a data frame is a list, use `sapply` to ask about the class of each column in the data frame. Explain to your neighbor what this produces, and why.*

*Use `table` to summarize the number of males and females in the sample. Consult the help page `?table` to figure out additional arguments required to include NA values in the tabulation.*

*The `mol.biol` column summarizes molecular biological attributes of each sample. Use `table` to summarize the different molecular biology levels in the sample. Use `%in%` to create a logical vector of the samples that are either BCR/ABL or NEG. Subset the original phenotypic data to contain those samples that are BCR/ABL or NEG.*

*After sub-setting, what are the levels of the `mol.biol` column? Set the levels to be BCR/ABL and NEG, i.e., the levels in the subset.*

*One would like covariates to be similar across groups of interest. Use `t.test` to assess whether BCR/ABL and NEG have individuals with similar age. To do this, use a formula that describes the response age in terms of the predictor `mol.biol`. If age is not independent of molecular biology, what complications might this introduce into subsequent analysis? Use the `boxplot` function to visualize the relationship between age and `mol.biol`.*

**Solution:** Here we input the data and explore basic properties.

```
pdata <- read.table(pdataFile)
dim(pdata)
```

---

\*[mtmorgan@fhcrc.org](mailto:mtmorgan@fhcrc.org)

```
## [1] 128 21

names(pdata)

## [1] "cod"          "diagnosis"    "sex"          "age"          "BT"
## [6] "remission"    "CR"           "date.cr"      "t.4.11."      "t.9.22."
## [11] "cyto.normal"  "citog"        "mol.biol"     "fusion.protein" "mdr"
## [16] "kinet"        "ccr"          "relapse"      "transplant"   "f.u"
## [21] "date.last.seen"

summary(pdata)

##      cod      diagnosis      sex      age      BT      remission
## 10005 : 1  1/15/1997 : 2  F   :42  Min. : 5.0  B2   :36  CR   :99
## 1003  : 1  1/29/1997 : 2  M   :83  1st Qu.:19.0 B3   :23  REF  :15
## 1005  : 1  11/15/1997: 2  NA's: 3  Median :29.0 B1   :19  NA's:14
## 1007  : 1  2/10/1998 : 2           Mean  :32.4 T2   :15
## 1010  : 1  2/10/2000 : 2           3rd Qu.:45.5 B4   :12
## 11002 : 1  (Other)   :116          Max.   :58.0 T3   :10
## (Other):122 NA's      : 2           NA's   :5  (Other):13
##      CR      date.cr      t.4.11.      t.9.22.      cyto.normal
## CR      :96  11/11/1997: 3  Mode :logical  Mode :logical  Mode :logical
## DEATH IN CR      : 3  1/21/1998 : 2  FALSE:86  FALSE:67  FALSE:69
## DEATH IN INDUCTION: 7  10/18/1999: 2  TRUE :7  TRUE :26  TRUE :24
## REF      :15  12/7/1998 : 2  NA's :35  NA's :35  NA's :35
## NA's      : 7  1/17/1997 : 1
##      (Other)   :87
##      NA's      :31
##      citog      mol.biol      fusion.protein      mdr      kinet
## normal      :24  ALL1/AF4:10  p190      :17  NEG :101  dyploid:94
## simple alt. :15  BCR/ABL :37  p190/p210: 8  POS : 24  hyperd.:27
## t(9;22)     :12  E2A/PBX1: 5  p210      : 8  NA's: 3  NA's : 7
## t(9;22)+other:11  NEG      :74  NA's      :95
## complex alt. :10  NUP-98 : 1
## (Other)     :21  p15/p16 : 1
## NA's        :35
##      ccr      relapse      transplant      f.u      date.last.seen
## Mode :logical  Mode :logical  Mode :logical  REL      :61  1/7/1998 : 2
## FALSE:74      FALSE:35      FALSE:91      CCR      :23  12/15/1997: 2
## TRUE :26      TRUE :65      TRUE :9       BMT / DEATH IN CR: 4  12/31/2002: 2
## NA's :28      NA's :28      NA's :28      BMT / CCR      : 3  3/29/2001 : 2
##      DEATH IN CR      : 2  7/11/1997 : 2
##      (Other)         : 7  (Other)   :83
##      NA's           :28  NA's      :35
```

A data frame can be subset as if it were a matrix, or a list of column vectors.

```
head(pdata[, "sex"], 3)

## [1] M M F
## Levels: F M

head(pdata$sex, 3)

## [1] M M F
## Levels: F M

head(pdata[["sex"]], 3)
```

```
## [1] M M F
## Levels: F M

sapply(pdata, class)

##          cod      diagnosis      sex      age      BT      remission
##      "factor"    "factor"    "factor"  "integer"  "factor"  "factor"
##          CR      date.cr      t.4.11.    t.9.22.  cyto.normal  citog
##      "factor"    "factor"    "logical"  "logical"  "logical"  "factor"
##      mol.biol fusion.protein      mdr      kinet      ccr      relapse
##      "factor"    "factor"    "factor"    "factor"  "logical"  "logical"
##      transplant      f.u date.last.seen
##      "logical"    "factor"    "factor"
```

The number of males and females, including NA, is

```
table(pdata$sex, useNA = "ifany")

##
##      F      M <NA>
##      42     83     3
```

An alternative version of this uses the `with` function: `with(pdata, table(sex, useNA="ifany"))`.

The `mol.biol` column contains the following samples:

```
with(pdata, table(mol.biol, useNA = "ifany"))

## mol.biol
## ALL1/AF4  BCR/ABL E2A/PBX1      NEG  NUP-98  p15/p16
##          10      37      5      74      1      1
```

A logical vector indicating that the corresponding row is either BCR/ABL or NEG is constructed as

```
ridx <- pdata$mol.biol %in% c("BCR/ABL", "NEG")
```

We can get a sense of the number of rows selected via `table` or `sum` (discuss with your neighbor what `sum` does, and why the answer is the same as the number of TRUE values in the result of the `table` function).

```
table(ridx)

## ridx
## FALSE  TRUE
##      17   111

sum(ridx)

## [1] 111
```

The original data frame can be subset to contain only BCR/ABL or NEG samples using the logical vector `ridx` that we created.

```
pdata1 <- pdata[ridx, ]
```

The levels of each factor reflect the levels in the original object, rather than the levels in the subset object, e.g.,

```
levels(pdata1$mol.biol)

## [1] "ALL1/AF4" "BCR/ABL" "E2A/PBX1" "NEG"      "NUP-98"  "p15/p16"
```

These can be re-coded by updating the new data frame to contain a factor with the desired levels.

```
pdata1$mol.biol <- factor(pdata1$mol.biol)
table(pdata1$mol.biol)
```

```
##
## BCR/ABL      NEG
##      37      74
```

To ask whether age differs between molecular biology types, we use a formula `age ~ mol.biol` to describe the relationship ('age as a function of molecular biology') that we wish to test

```
with(pdata1, t.test(age ~ mol.biol))

##
## Welch Two Sample t-test
##
## data:  age by mol.biol
## t = 4.817, df = 68.53, p-value = 8.401e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  7.135 17.224
## sample estimates:
## mean in group BCR/ABL      mean in group NEG
##                40.25                28.07
```

This summary can be visualize with, e.g., the `boxplot` function

```
## not evaluated
boxplot(age ~ mol.biol, pdata1)
```

Molecular biology seem to be strongly associated with age; individuals in the NEG group are considerably younger than those in the BCR/ABL group. We might wish to include age as a covariate in any subsequent analysis seeking to relate molecular biology to gene expression.