R2

Ashish Mahabal
AyBi199, Caltech
7 May 2009

Quick recap

- ? To get help
- # for comments
- -> or <- (or =) for assignments
- == for comparison

- Concatenate with c # x = c(a,b)
- seq to create sequence # z = seq(-30,30)
- is.na(x) to look for notavailableness

- summary(x)
- data(x)
- read.table(file)
- save(file)
- source(file) # input
- sink(file) # output

Common routines

- length(a) # length of vector
- max(a) # similarly min, mean
- sort(a), or sort(a, decreasing=T)
- unique(a)
- duplicated(a) # returns a logical array!

Tom Short's R ref card

http://www.rpad.org/Rpad/R-refcard.pdf

R Reference Card

by Tom Short, EPRI Solutions, Inc., tshort@episolutions.com 2005-07-12

Help and basics

Most R functions have online documentation.

help(topic) documentation on topic
?topic id.
help.search("topic") search the help system
apropos("topic") the names of all objects in the search list matching the regular expression "topic"
help.start() start the HTML version of help
str(a) display the internal *str*ucture of an R object
summary(a) gives a "summary" of a, usually a statistical summary but it is *generic* meaning it has different operations for different classes of a
ls() show objects in the search path; specify pat="pat" to search on a pattern
ls.str() str() for each variable in the search path
dir() show files in the current directory
methods(a) shows S3 methods of a
methods(class=class(a)) lists all the methods to handle objects of class a
options(...) set or examine many global options; common ones: width, digits, error
library(x) load add-on packages; library(help=x) lists datasets and functions in package x.
attach(x) database x to the R search path; x can be a list, data frame, or R data file created with save. Use search() to show the search path.
detach(x) x from the R search path; x can be a name or character string of an object previously attached or a package.

Input and output

load() load the datasets written with save
data(x) loads specified data sets
read.table(file) reads a file in table format and creates a data frame from it; the default separator sep="" is any whitespace; use header=TRUE to read the first line as a header of column names; use as.is=TRUE to prevent character vectors from being converted to factors; use comment.char="" to prevent "#" from being interpreted as a comment; use skip=n to skip n lines before reading data; see the help for options on row naming, NA treatment, and others
read.csv("filename",header=TRUE) id. but with defaults set for reading comma-delimited files
read.delim("filename",header=TRUE) id. but with defaults set for reading tab-delimited files
read.fwf(file,widths,header=FALSE,sep="",as.is=FALSE) read a table of *f*ixed width *f*ormatted data into a 'data.frame'; widths is an integer vector, giving the widths of the fixed-width fields
save(file,...) saves the specified objects (...) in the XDR platform-independent binary format
save.image(file) saves all objects
cat(..., file="", sep=" ") prints the arguments after coercing to character; sep is the character separator between arguments
print(a, ...) prints its arguments; generic, meaning it can have different methods for different objects
format(x,...) format an R object for pretty printing
write.table(x,file="",row.names=TRUE,col.names=TRUE, sep=" ") prints x after converting to a data frame; if quote is TRUE, character or factor columns are surrounded by quotes ("); sep is the field separator; eol is the end-of-line separator; na is the string for missing values; use col.names=NA to add a blank column header to get the column headers aligned correctly for spreadsheet input
sink(file) output to file, until sink()

Most of the I/O functions have a file argument. This can often be a character string naming a file or a connection. file="" means the standard input or output. Connections can include files, pipes, zipped files, and R variables.

On windows, the file connection can also be used with description = "clipboard". To read a table copied from Excel, use
x <- read.delim("clipboard")
To write a table to the clipboard for Excel, use
write.table(x,"clipboard",sep="\t",col.names=NA)
For database interaction, see packages RODBC, DBI, RMySQL, RPgSQL, and ROracle. See packages XML, hdf5, netCDF for reading other file formats.

Data creation

c(...) generic function to combine arguments with the default forming a vector; with recursive=TRUE descends through lists combining all elements into one vector
from:to generates a sequence; ":" has operator priority; 1:4 + 1 is "2,3,4,5"
seq(from,to) generates a sequence by= specifies increment; length= specifies desired length
seq(along=x) generates 1, 2, ..., length(x); useful for for loops
rep(x,times) replicate x times; use each= to repeat "each" element of x each times; rep(c(1,2,3),2) is 1 2 3 1 2 3; rep(c(1,2,3),each=2) is 1 1 2 2 3 3
data.frame(...) create a data frame of the named or unnamed arguments; data.frame(v=1:4,ch=c("a","B","c","d"),n=10); shorter vectors are recycled to the length of the longest
list(...) create a list of the named or unnamed arguments; list(a=c(1,2),b="hi",c=3i);
array(x,dim=) array with data x; specify dimensions like dim=c(3,4,2); elements of x recycle if x is not long enough
matrix(x,nrow=,ncol=) matrix; elements of x recycle
factor(x,levels=) encodes a vector x as a factor
gl(n,k,length=n*k,labels=1:n) generate levels (factors) by specifying the pattern of their levels; k is the number of levels, and n is the number of replications
expand.grid() a data frame from all combinations of the supplied vectors or factors
rbind(...) combine arguments by rows for matrices, data frames, and others
cbind(...) id. by columns

Slicing and extracting data

Indexing lists

x[n]	list with elements n
x[[n]]	n^{th} element of the list
x[["name"]]	element of the list named "name"
x$name	id.

Indexing vectors

x[n]	n^{th} element
x[-n]	all *but* the n^{th} element
x[1:n]	first n elements
x[-(1:n)]	elements from n+1 to the end
x[c(1,4,2)]	specific elements
x["name"]	element named "name"
x[x > 3]	all elements greater than 3
x[x > 3 & x < 5]	all elements between 3 and 5
x[x %in% c("a","and","the")]	elements in the given set

Indexing matrices

x[i,j]	element at row i, column j
x[i,]	row i
x[,j]	column j
x[,c(1,3)]	columns 1 and 3
x["name",]	row named "name"

Indexing data frames (matrix indexing plus the following)

x[["name"]]	column named "name"
x$name	id.

Variable conversion

as.array(x), as.data.frame(x), as.numeric(x), as.logical(x), as.complex(x), as.character(x), ... convert type; for a complete list, use methods(as)

Variable information

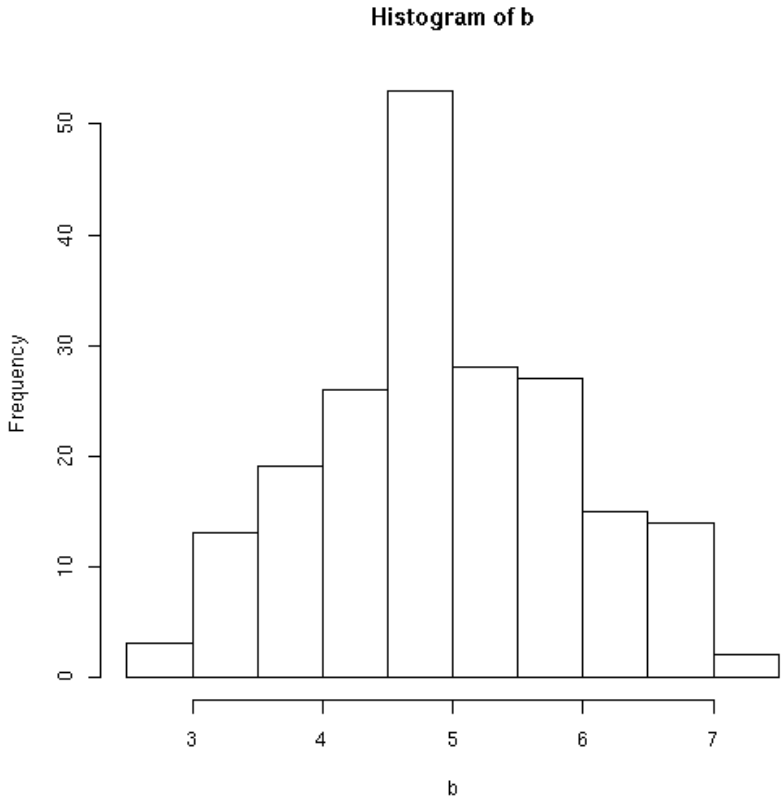
is.na(x), is.null(x), is.array(x), is.data.frame(x), is.numeric(x), is.complex(x), is.character(x), ... test for type; for a complete list, use methods(is)
length(x) number of elements in x
dim(x) Retrieve or set the dimension of an object; dim(x) <- c(3,2)
dimnames(x) Retrieve or set the dimension names of an object
nrow(x) number of rows; NROW(x) is the same but treats a vector as a one-row matrix
ncol(x) and **NCOL(x)** id. for columns
class(x) get or set the class of x; class(x) <- "myclass"
unclass(x) remove the class attribute of x
attr(x,which) get or set the attribute which of x
attributes(obj) get or set the list of attributes of obj

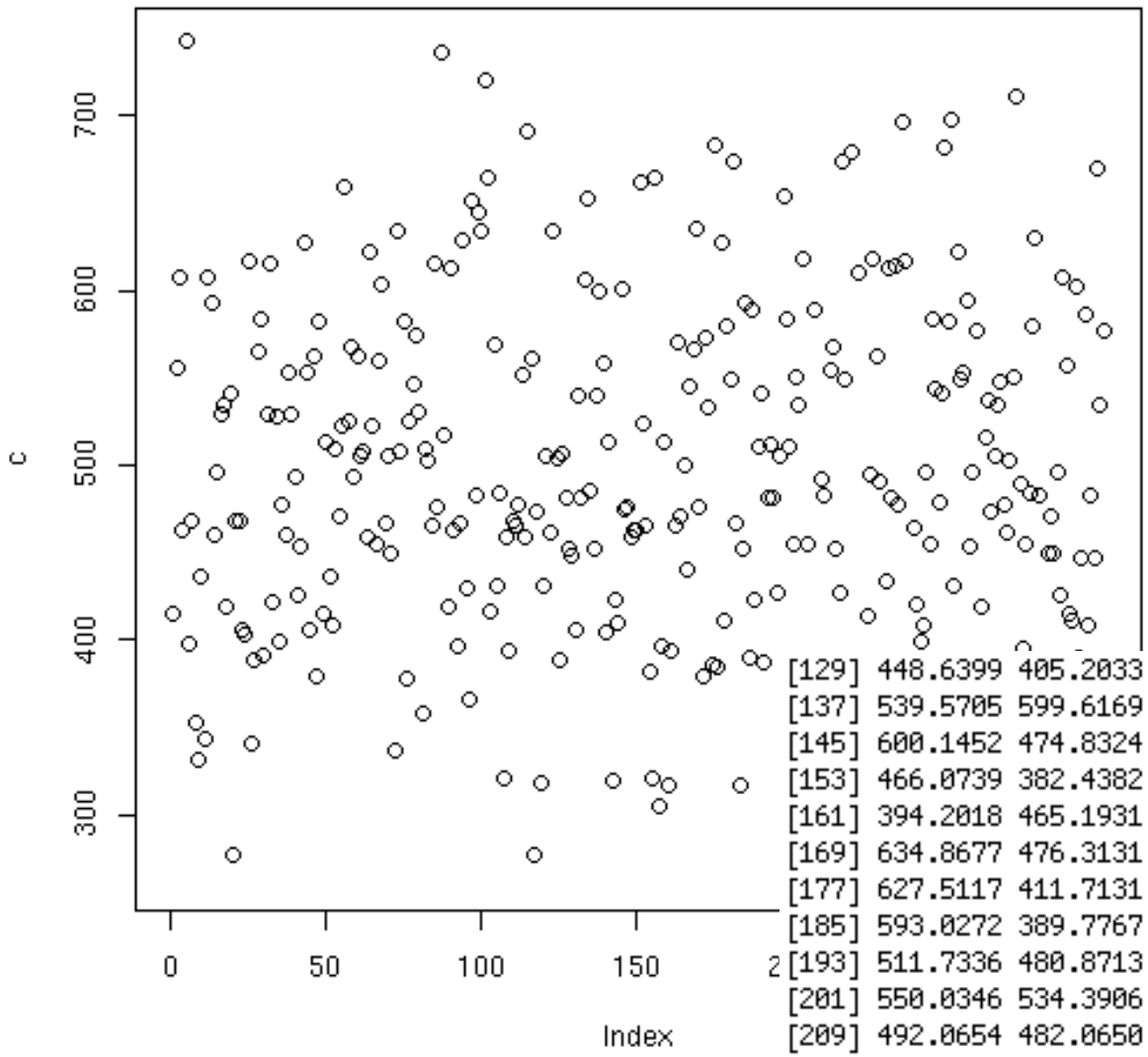
Data selection and manipulation

which.max(x) returns the index of the greatest element of **x**
which.min(x) returns the index of the smallest element of **x**
rev(x) reverses the elements of x
sort(x) sorts the elements of x in increasing order; to sort in decreasing order: rev(sort(x))
cut(x,breaks) divides x into intervals (factors); breaks is the number of cut intervals or a vector of cut points

Example

- a=rnorm(100,mean=5,sd=1)
- b=rnorm(200,mean=5,sd=1)
- c=c(100*a,100*b)
- length(c)
- c





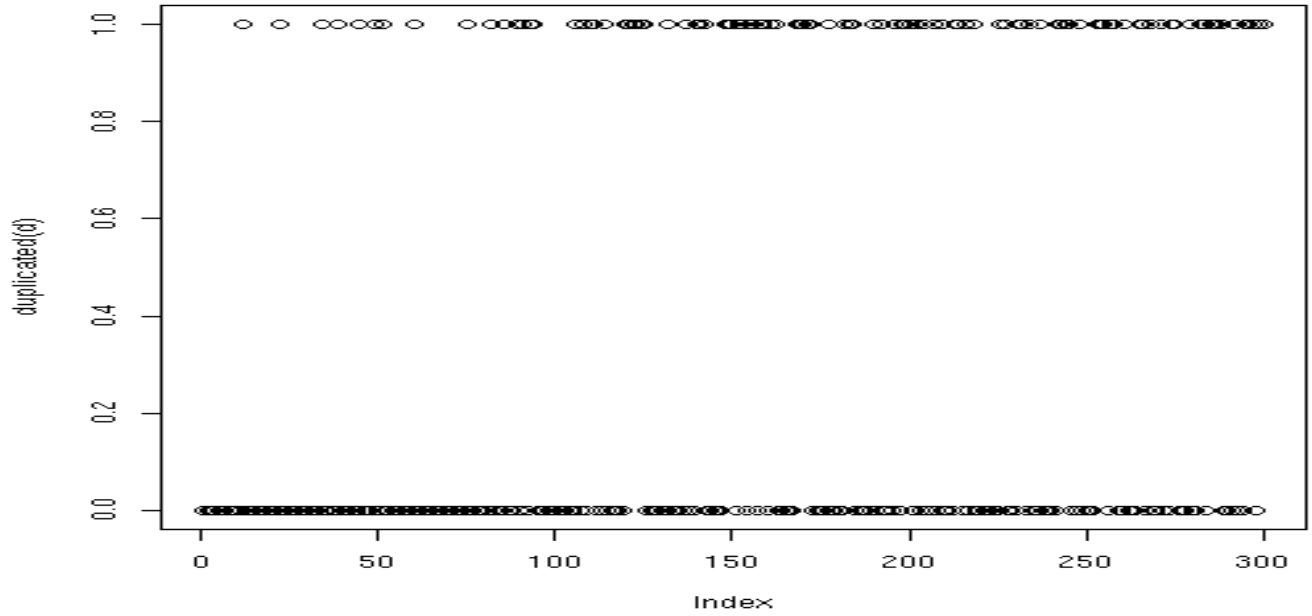

```
[129] 448.6399 405.2033 539.6339 481.5814 605.8925 652.4077 485.9140 452.6308
[137] 539.5705 599.6169 558.6897 403.9841 513.5366 320.0797 423.3735 409.3903
[145] 600.1452 474.8324 475.7574 458.8818 462.2530 462.1789 662.1314 524.2215
[153] 466.0739 382.4382 320.3050 664.2152 304.6201 396.7627 513.1186 316.9793
[161] 394.2018 465.1931 570.3309 471.2018 499.7137 439.7604 545.3611 565.8577
[169] 634.8677 476.3131 378.7135 573.0420 533.1097 386.3176 683.4459 383.8697
[177] 627.5117 411.7131 579.6656 548.9674 673.4191 466.5636 316.8249 451.9777
[185] 593.0272 389.7767 588.8479 422.8503 510.6999 541.5611 387.2153 481.9646
[193] 511.7336 480.8713 426.9614 504.6129 653.7176 583.9883 510.9250 454.2808
[201] 550.0346 534.3906 618.4641 389.1341 455.4347 351.0636 588.9437 373.3234
[209] 492.0654 482.0650 302.8964 554.8883 567.3488 451.8892 427.2502 673.1516
[217] 549.4872 340.9521 678.7345 324.3261 609.4050 271.7941 347.2002 413.5027
[225] 494.1476 618.3846 561.6571 490.4388 311.9638 433.9501 612.2427 481.6557
[233] 613.3592 477.4250 695.8766 616.0761 379.0668 307.3831 464.5840 420.0470
[241] 398.5331 408.9758 496.4556 454.2031 583.3175 543.1616 478.5739 541.4109
[249] 681.4943 581.5734 697.4596 431.3465 622.0025 549.1973 552.8481 594.3782
[257] 454.1116 496.1235 576.2675 369.9352 419.1967 515.2750 537.4536 473.0187
[265] 505.1537 534.9228 547.8853 476.7180 461.7504 502.4479 550.4487 710.9494
[273] 489.0445 394.7242 455.1865 484.4981 580.1016 630.0980 482.1229 264.0277
```

- d=as.integer(c)
- d
- mean(d); max(d); min(d); sd(d)

```
> a
  [1] 415 556 607 462 742 397 468 352 331 435 343 607 592 459 496 528 534 419
 [19] 541 277 467 467 406 403 617 340 388 565 583 391 529 614 421 528 399 476
 [37] 460 553 528 492 425 453 627 552 406 562 379 582 415 513 435 408 509 470
 [55] 522 658 524 567 493 562 504 507 458 622 521 454 559 603 466 505 449 336
 [73] 633 508 582 378 525 546 573 530 358 508 501 465 615 476 735 516 419 612
 [91] 462 396 467 627 429 366 650 483 643 634 719 664 416 569 430 483 320 458
[109] 394 467 465 477 551 458 690 561 276 472 317 430 505 460 633 504 388 506
[127] 481 451 448 405 539 481 605 652 485 452 539 599 558 403 513 320 423 409
[145] 600 474 475 458 462 462 662 524 466 382 320 664 304 396 513 316 394 465
[163] 570 471 499 439 545 565 634 476 378 573 533 386 683 383 627 411 579 548
[181] 673 466 316 451 593 389 588 422 510 541 387 481 511 480 426 504 653 583
[199] 510 454 550 534 618 389 455 351 588 373 492 482 302 554 567 451 427 673
[217] 549 340 678 324 609 271 347 413 494 618 561 490 311 433 612 481 613 477
[235] 695 616 379 307 464 420 398 408 496 454 583 543 478 541 681 581 697 431
[253] 622 549 552 594 454 496 576 369 419 515 537 473 505 534 547 476 461 502
[271] 550 710 489 394 455 484 580 630 482 264 359 449 470 450 496 425 607 556
[289] 414 410 601 389 446 585 408 482 446 670 533 576
```

```
[109] FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE F
[121]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE F
[133] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
[145] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE F
[157] FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE F
[169]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE F
[181] FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
[193] FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
[205] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
[217] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[229] FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE F
[241] FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE F
```

- unique(d)
- duplicated(d)
- plot(duplicated(d)) # coercion

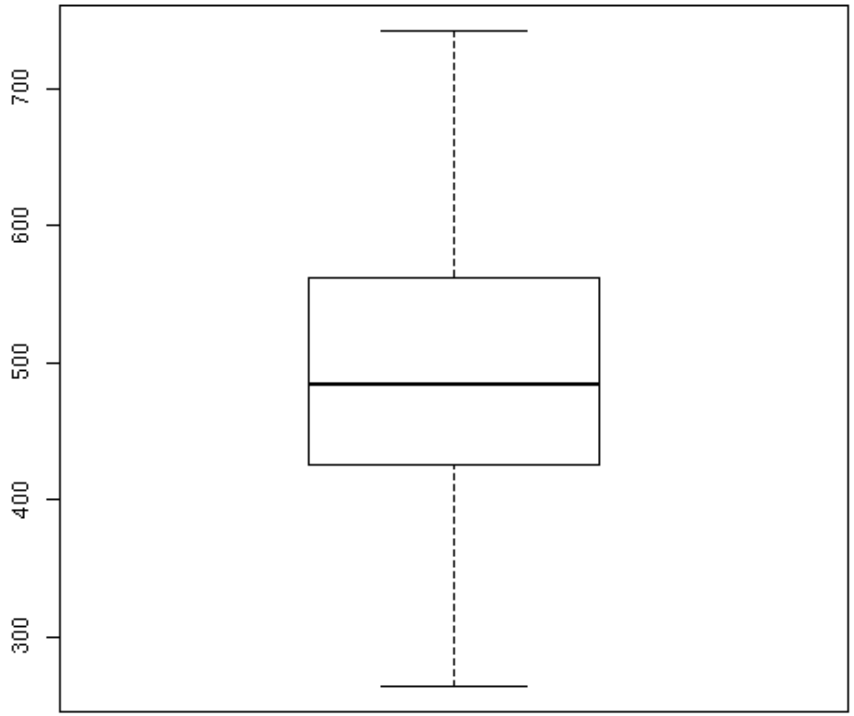


- duplicated(unique(d))
- five(d) # whats the five-of-d
- ?five # nothing? Do you know five?
- ??five # at least something that has five?
- fivenum(d) # ah, so its fivenum. I'll take that
- ?fivenum # what is it BTW?

… minimum, lower-hinge, median,upper-hinge, maximum …

- fivenum(x, na.rm = TRUE) # options!
- g=d
- g[duplicated(d)]<-NA
- fivenum(d)
  - [1] 264.0 425.5 483.5 561.5 742.0
- fivenum(g,na.rm = TRUE)
  - [1] 264.0 419.5 492.0 571.5 742.0

- boxplot.stats(x, coef = 1.5, do.conf = TRUE, do.out = TRUE)
- quantile, range, bxp



- which selects indices:
- a=which(d<500) # returns indices
- a=d[which(d<500)] # returns the elements
- which.max(d) # index of max element

```
 [1]   1   4   6   7   8   9  10  11  14  15
[19]  30  33  35  36  37  40  41  42  45  47
[37]  71  72  76  81  84  86  89  91  92  93
[55] 109 110 111 112 114 117 118 119 120 122 1
[73] 140 142 143 144 146 147 148 149 150 153 1
[91] 165 166 170 171 174 176 178 182 183 184 1
[109] 205 206 208 209 210 211 214 215 218 220 2
[127] 234 237 238 239 240 241 242 243 244 247 2
[145] 273 274 275 276 279 280 281 282 283 284 2
[163] 297
```

```
  [1] 415 462 397 468 352 331 435 343 459 496 419
 [19] 391 421 399 476 460 492 425 453 406 379 415
 [37] 449 336 378 358 465 476 419 462 396 467 429
 [55] 394 467 465 477 458 276 472 317 430 460 388
 [73] 403 320 423 409 474 475 458 462 462 466 382
 [91] 499 439 476 378 386 383 411 466 316 451 389
[109] 455 351 373 492 482 302 451 427 340 324 271
[127] 477 379 307 464 420 398 408 496 454 478 431
[145] 489 394 455 484 482 264 359 449 470 450 496
[163] 446
```

- match(x,y) # elements of x in y, else NA
- merge(a,b) # using common columns/rows

```
> t1=read.table('table1',header=TRUE)
> t1
   id    ra   dec
1 101 200.1  33.1
2 102 199.3 -13.3
3 103 200.2  19.1
> t2=read.table('table2',header=TRUE)
> t2
   id mag
1 101  17
2 102  18
3 104  19
```

```
> merge(t1,t2)
   id    ra   dec mag
1 101 200.1  33.1  17
2 102 199.3 -13.3  18
```

choose(n,k) # combinations of k from n

choose(5,3) returns 10

sample(x,size) # resamples size elements from x (with the option of replacement)

```
> cards= paste(c("C","D","H","S"), rep(1:13,times=4), sep="")
> cards
 [1] "C1"  "D2"  "H3"  "S4"  "C5"  "D6"  "H7"  "S8"  "C9"  "D10" "H11" "S12"
[13] "C13" "D1"  "H2"  "S3"  "C4"  "D5"  "H6"  "S7"  "C8"  "D9"  "H10" "S11"
[25] "C12" "D13" "H1"  "S2"  "C3"  "D4"  "H5"  "S6"  "C7"  "D8"  "H9"  "S10"
[37] "C11" "D12" "H13" "S1"  "C2"  "D3"  "H4"  "S5"  "C6"  "D7"  "H8"  "S9"
[49] "C10" "D11" "H12" "S13"
```

- > sample(cards,5)

[1] "S9" "H4" "D5" "C3" "C9"

- > sample(cards,2)

[1] "H10" "D10"

- Conj(4+3i)
- fft(x)
- solve(a) # matrix inverse of a
- solve(a,b) # solves a %*% x = b for x

```
> tmat2
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    9   16   25
[2,]    2    8   18   32   50
[3,]    3   12   27   48   75
[4,]    4   16   36   64  100
[5,]    5   20   45   80  125
> solve(tmat2)
Error in solve.default(tmat2) :
  Lapack routine dgesv: system is exactly singular
> tmat3=tmat2*rnorm(25)
> tmat3
           [,1]      [,2]       [,3]      [,4]        [,5]
[1,] -0.2227551 -6.751126 -12.490741  24.65620   -3.117128
[2,] -2.0841821  2.421399  11.489664 -57.46727  -17.618221
[3,]  3.5686843  6.241676  52.513037  53.94490  -15.989971
[4,] -0.1881898 10.699605  -4.583948 -68.40203   82.022038
[5,] -0.1673437 25.650273 -26.332502 104.38119 -133.083879
```

```
> solve(tmat3)
             [,1]        [,2]        [,3]        [,4]         [,5]
[1,] -2.35764651 -0.97551325 -0.47699801 -0.63287369 -0.148376429
[2,]  0.05424243  0.02748805  0.02324214  0.05100759  0.023734912
[3,]  0.08538674  0.04414917  0.03231509  0.02145215  0.001494099
[4,]  0.08539455  0.02308904  0.02026885  0.02211618  0.006138554
[5,]  0.06350137  0.01589843  0.01458283  0.02372857  0.001766126
> rcorr(tmat3)
     [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 1.00  0.06  0.67  0.48  0.00
[2,] 0.06  1.00 -0.31  0.44 -0.54
[3,] 0.67 -0.31  1.00 -0.09  0.27
[4,] 0.48  0.44 -0.09  1.00 -0.82
[5,] 0.00 -0.54  0.27 -0.82  1.00
```

```
> solve(tmat3,tmat2)
            [,1]        [,2]       [,3]        [,4]        [,5]
[1,] -9.0130440 -36.0521758 -81.117396 -144.208703 -225.326099
[2,]  0.5016499   2.0065995   4.514849    8.026398   12.541247
[3,]  0.3639094   1.4556378   3.275185    5.822551    9.097736
[4,]  0.3115367   1.2461467   2.803830    4.984587    7.788417
[5,]  0.2427916   0.9711665   2.185125    3.884666    6.069791
> s=solve(tmat3,tmat2)
> tmat3 %*% s
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    9   16   25
[2,]    2    8   18   32   50
[3,]    3   12   27   48   75
[4,]    4   16   36   64  100
[5,]    5   20   45   80  125
```

Strings and date/time

- paste(…) # concatenating vectors
- substr(x,start,stop) # can also assign
- grep(pattern,x)
- as.Date(s)

Graphics devices

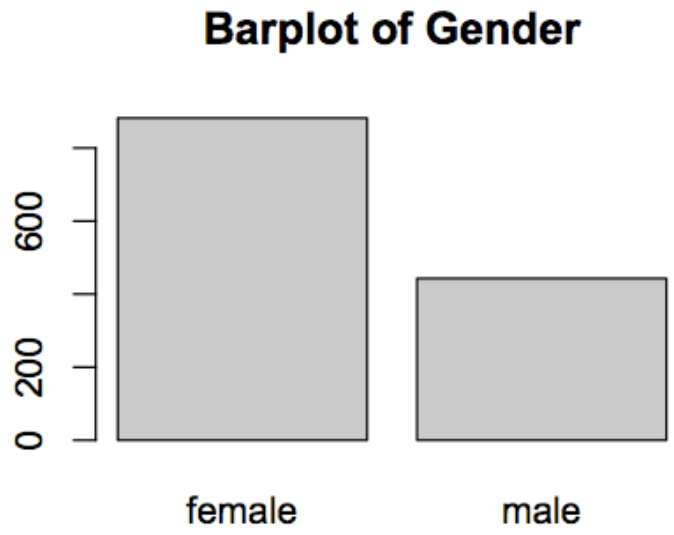
- x11(), windows()
- postscript(file)
- png, pdf, jpeg, …

Next few slides from http://scc.stat.ucla.edu/mini-courses

Frequency bar plots

Display counts of each category next to each other for easy comparison.

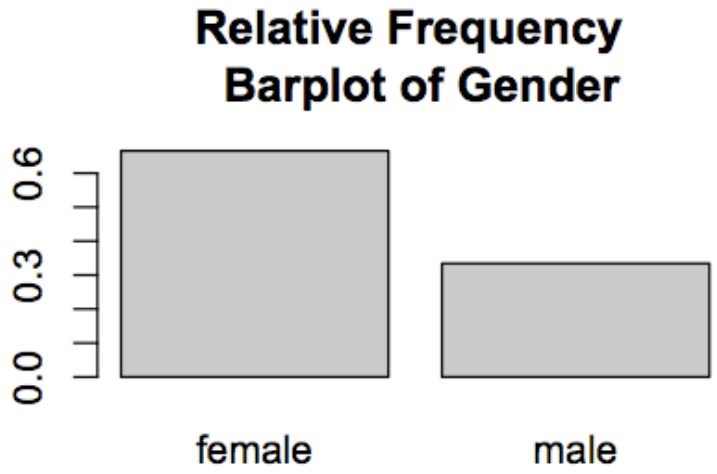
```
1 barplot(table(gender), main = "Barplot of
      Gender")
```



Relative frequency bar plots

Display relative proportions of each category.

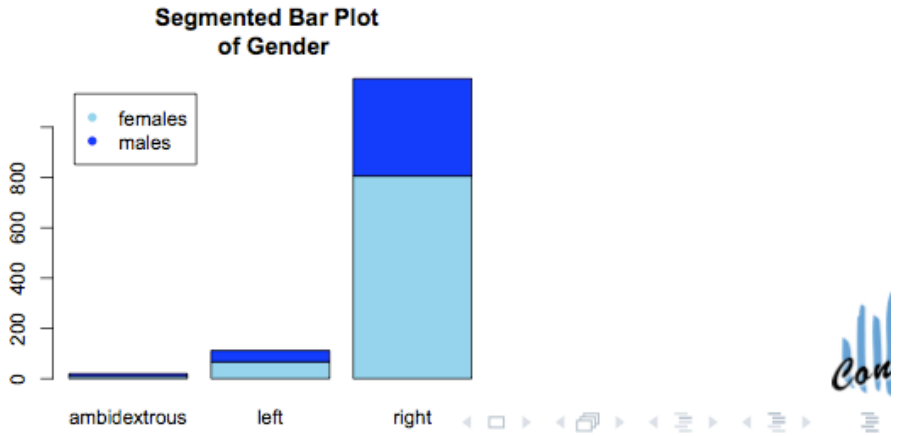
```
1 barplot(table(gender)/length(gender), main
    Relative Frequency \n Barplot of Gender
```



Segmented bar charts

Displays two categorical variables at a time.

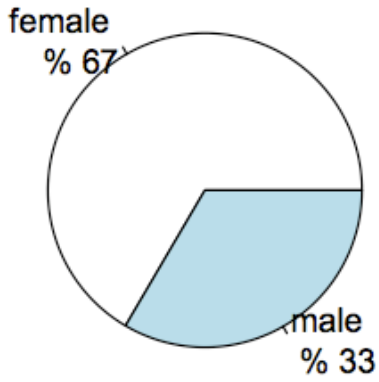
```
1 barplot(table(gender, hand), col = c("skyblue"
    , "blue"), main = "Segmented Bar Plot \n
    of Gender")
2 legend("topleft", c("females","males"), col =
    c("skyblue", "blue"), pch = 16, inset =
    0.05)
```



Pie charts

Pie charts display counts as percentages of individuals in each category.

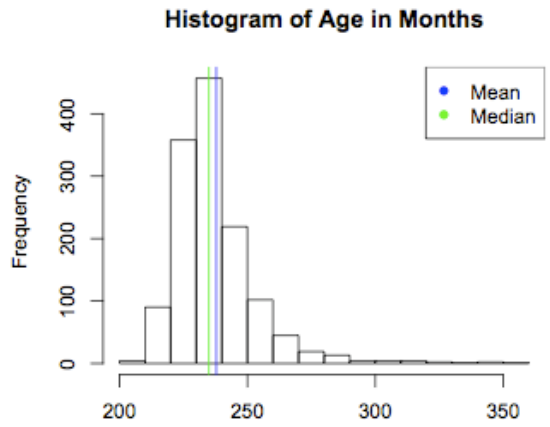
```
1 pct = round(table(gender) / length(gender) *
      100)
2 lbls = paste(names(table(gender)), "\n", "%",
      pct)
3 pie(table(gender), labels = lbls)
```



Adding measures to plots

Adding mean and median to a histogram.

```
1 hist(ageinmonths, main = "Histogram of Age in
      Months")
2 abline(v = mean(ageinmonths), col = "blue")
3 abline(v = median(ageinmonths), col = "green")
4 legend("topright", c("Mean", "Median"), pch =
      16, col = c("blue", "green"))
```

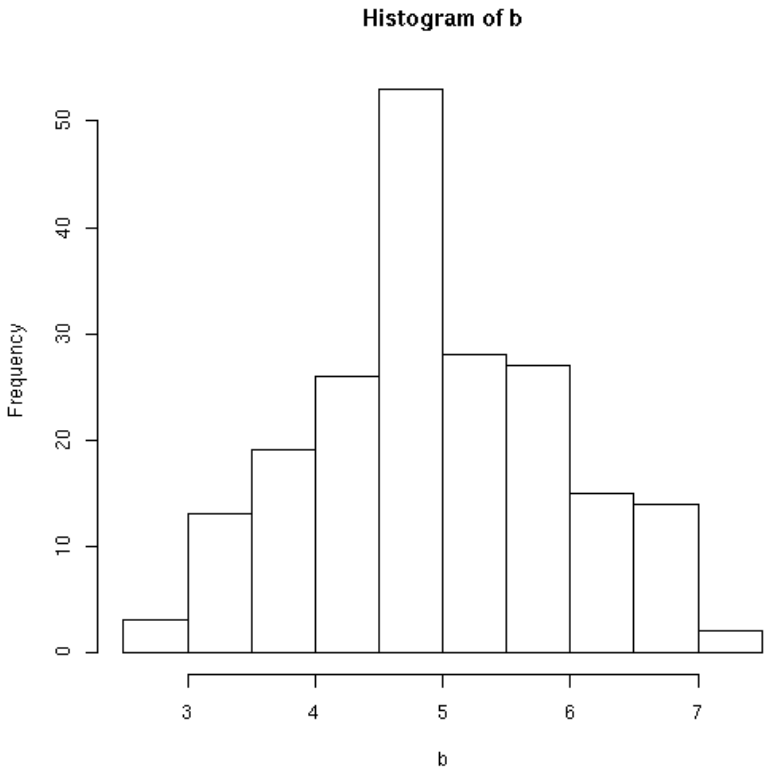


Distributions

`rnorm(n, mean=0, sd=1)` Gaussian (normal)
`rexp(n, rate=1)` exponential
`rgamma(n, shape, scale=1)` gamma
`rpois(n, lambda)` Poisson
`rweibull(n, shape, scale=1)` Weibull
`rcauchy(n, location=0, scale=1)` Cauchy
`rbeta(n, shape1, shape2)` beta
`rt(n, df)` 'Student' (t)
`rf(n, df1, df2)` Fisher–Snedecor (F) (χ^2)
`rchisq(n, df)` Pearson

Replace r by d,p,q to get prob density, cumulative prob density and the value of the quantile

- pnorm(4,5,1)
  - # how many below 4 for mean 5 and sd 1 (as a fraction)
  - [1] 0.1586553
- pnorm(6,5,1) – pnorm(4,5,1)
  - # between 4 and 6
  - [1] 0.6826895
- qnorm(0.9,5,1)
  - # 90th percentile?
  - [1] 6.281552



Geometric distribution

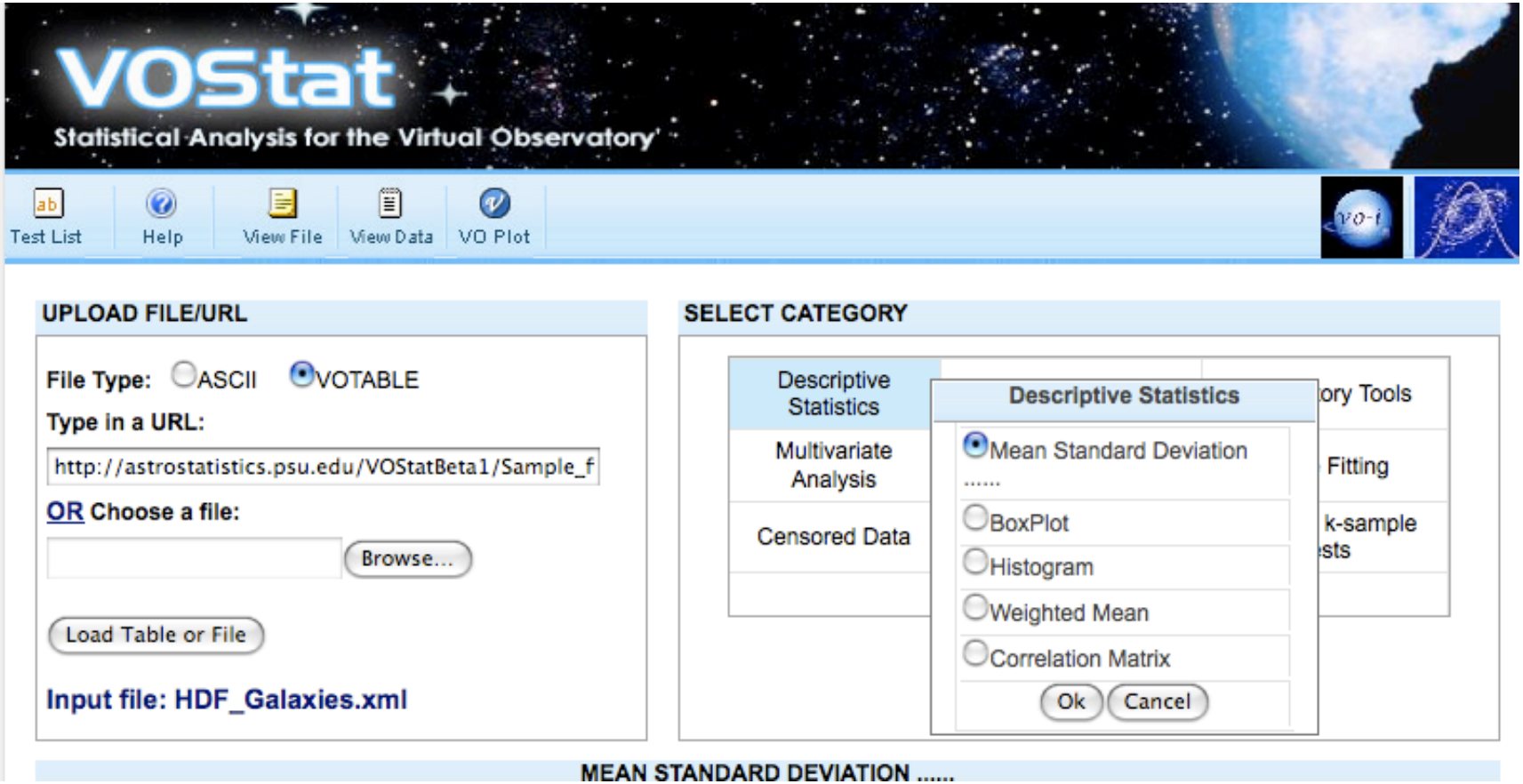
- dgeom(4,0.35)
  - Probability of first success on the 5^{th} trial given that the probability of success on a given try is 0.35
  - [1] 0.06247719

  For a fair coin its 0.5, 0.25, 0.125 etc.

  pgeom(n,p) gives the probability of success until that point. So it will be 0.5, 0.75, 0.875, …

Binomial distribution

- dbinom(m,n,p) gives probability of success in m out of n attempts given that the success at any given attempt is p
  - dbinom(3 , 5 , 0.35) # success in 3 of 5
    - [1] 0.1811469
  - sum(dbinom(3:5 , 5 , 0.35)) # in 3 or 4 or 5 of 5
    - [1] 0.2351694


VOStat
Statistical Analysis for the Virtual Observatory
Test List
Help
View File
View Data
VO Plot
UPLOAD FILE/URL
File Type: ASCII VOTABLE
Type in a URL:
http://astrostatistics.psu.edu/VOStatBeta1/Sample_f
OR Choose a file:
Browse...
Load Table or File
Input file: HDF_Galaxies.xml
SELECT CATEGORY
Descriptive Statistics
Multivariate Analysis
Censored Data
Descriptive Statistics
Mean Standard Deviation .....
BoxPlot
Histogram
Weighted Mean
Correlation Matrix
Ok
Cancel
ory Tools
Fitting
k-sample
sts
MEAN STANDARD DEVIATION .....

Statistical tests

- shapiro.test Shapiro-Wilk's W test (u)
- ks.test Kolmogorov-Smirnov 2-sample test (b)
- kruskal.test Kruskal-Wallis k-sample test (b)
- wilcox.test Wilcoxon rank-sum test (b)
- cor.test Bivariate correlation coefficient (b)
- cov Covarience (b)
- rcorr Correlation Matrix (m)
- prcomp Principle Component Analysis (m)

Linear regression help (partial snapshot)

Regression, in general, is the problem of estimating a conditional expected

It is often erroneously thought that the reason the technique is called "linea
But in fact, if the model is

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$$

(in which case we have put the vector (x_i, x_i^2) in the role formerly played
by β), then the problem is still one of **linear** regression, even though the gr

Linear regression is called "linear" because the relation of the response to t
function of some parameters. Regression models which are not a linear fun
models. A neural network is an example of a nonlinear regression model.

- library(Hmisc)
- table <- read.table(fname,header=F)
- tposed = t(table)
- rcorr(tposed, type="pearson”)
- rcorr(tposed, type="spearman”)

MEAN STANDARD DEVIATION .....

Column Names	Mean	Standard Deviation	Median	Median Absolute Deviation	Histogram	BoxPlot
RAJ2000	189.2064	0.02218875	189.2062	0.0252042	RAJ2000	RAJ2000
DEJ2000	1.036918	0.0002201704	1.036873	0.0002271947	DEJ2000	DEJ2000
Xpos	1126.767	568.3535	1183.5	739.8174	Xpos	Xpos
Ypos	1204.481	549.7623	1223	750.9369	Ypos	Ypos
Imag	23.69011	0.943918	23.92	0.978516	Imag	Imag
U_B	-0.4879699	0.7573193	-0.74	0.496671	U_B	U_B
B_V	0.7631481	0.5199708	0.625	0.437367	B_V	B_V
V_I	0.9223333	0.4341621	0.865	0.415128	V_I	V_I
recno	135.5630	78.18918	135.5	100.0755	recno	recno

Correlation with option Spearman

	RAJ2000	DEJ2000	Xpos	Ypos
RAJ2000	1.00000000000	0.3187517128	0.09078245882	0.11576158751
DEJ2000	0.31875171281	1.0000000000	-0.02557942879	-0.24798193285
Xpos	0.09078245882	-0.0255794288	1.00000000000	-0.03265218964
Ypos	0.11576158751	-0.2479819328	-0.03265218964	1.00000000000
Imag	0.00141798094	0.0204151591	0.00100888117	0.02376762228
U_B	-0.00960168334	0.1376018373	-0.18660312960	0.06837582322
B_V	-0.06606288772	-0.0990899686	-0.03850825371	-0.00249747380
V_I	-0.03861179659	-0.0830582013	0.01631353326	-0.04112840135
recno	0.00174061539	0.0197769383	0.00134310964	0.02399277050

Correlation with option Pearson

$r

	RAJ2000	DEJ2000	Xpos	Ypos
RAJ2000	1.00000000000	0.2843183050	0.13905806036	0.0696222736
DEJ2000	0.28431830505	1.0000000000	0.04215667230	-0.1992923014
Xpos	0.13905806036	0.0421566723	1.00000000000	-0.0357669735
Ypos	0.06962227361	-0.1992923014	-0.03576697346	1.0000000000
Imag	0.00611365673	0.0291307939	0.01823175639	0.0313959000
U_B	-0.07470960244	0.0620628837	-0.17230925885	0.0754271916
B_V	-0.09422345021	-0.1525281203	-0.04881731975	0.0158323456
V_I	-0.05097537179	-0.1221161202	0.02935269575	-0.0413020859
recno	0.00627972174	0.0219099629	0.00398853955	0.0270030626

Hmisc: Harrell Miscellaneous

The Hmisc library contains many functions useful for data analysis, high-leve
computing sample size and power, importing datasets, imputing missing valu
character string manipulation, conversion of S objects to LaTeX code, and re
'http://biostat.mc.vanderbilt.edu/trac/Hmisc'.

Version:	3.6-0
Depends:	R (≥ 2.4.0), methods
Imports:	lattice, cluster, survival
Suggests:	lattice, grid, nnet, foreign, chron, acepack, TeachingDemos, De
Published:	2009-04-29
Author:	Frank E Harrell Jr, with contributions from many other users.
Maintainer:	Charles Dupont <charles.dupont at vanderbilt.edu>
License:	GPL (≥ 2)
URL:	http://biostat.mc.vanderbilt.edu/s/Hmisc, http://biostat.mc.vand http://biostat.mc.vanderbilt.edu/twiki/pub/Main/StatReport/sun /trac/Hmisc

- library()
- R CMD INSTALL -l libpath ~/Hmisc_3.6-0.tar.gz
- Include this in your .Rprofile file

```
> .First <- function() {
    options(prompt="$ ", continue="+\t")   # $ is the prompt
    options(digits=5, length=999)          # custom numbers and printout
    x11()                                  # for graphics
    par(pch = "+")                         # plotting character
    source(file.path(Sys.getenv("HOME"), "R", "mystuff.R"))
                                           # my personal functions
    library(MASS)                          # attach a package
  }
```

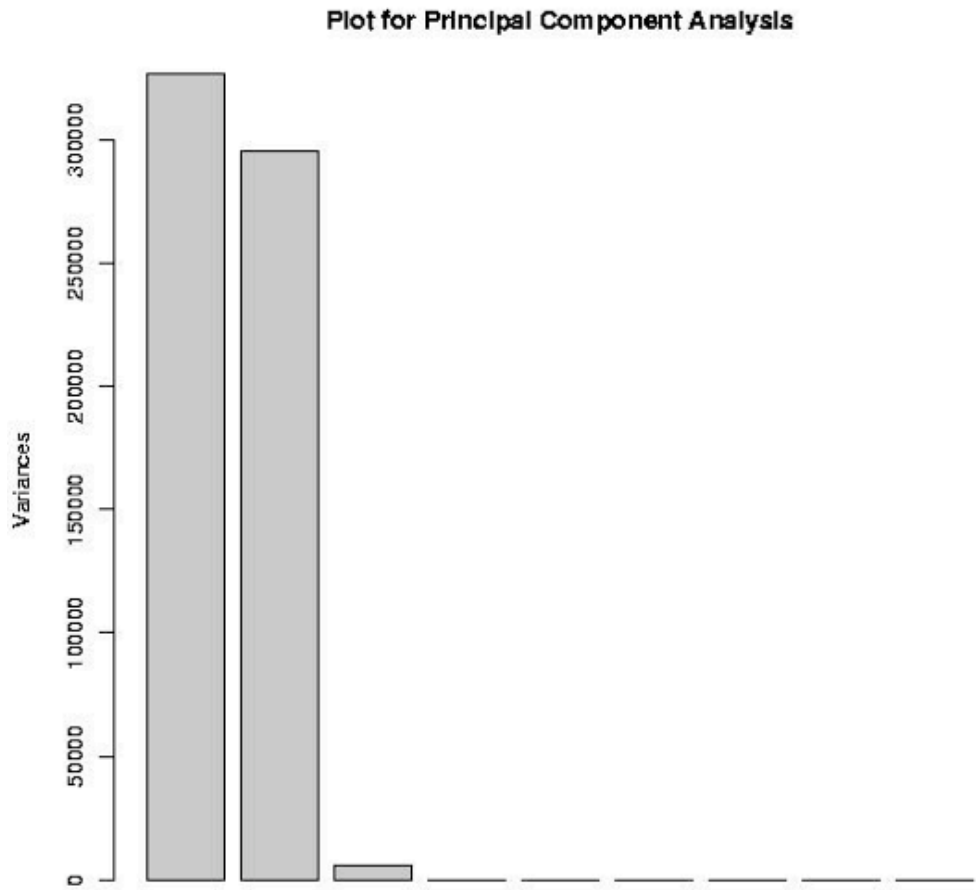
```
> .Last <- function() {
  graphics.off()                          # a small safety measure.
  cat(paste(date(),"\nAdios\n"))          # Is it time for lunch?
}
```

Principal Component Analysis

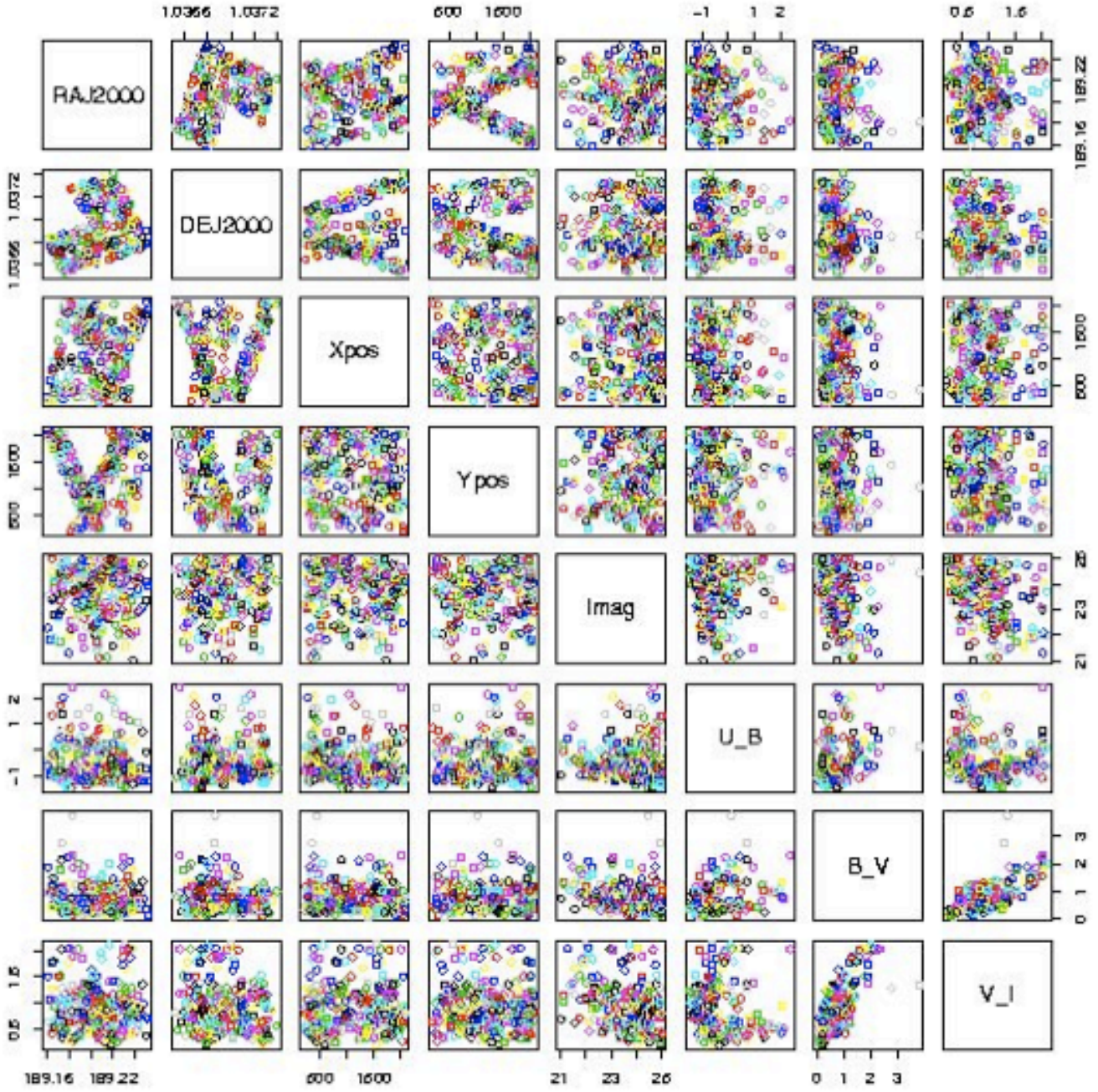
- library(mva)
- table <- read.table(”fname",header=T)
- prcomp(table)
- summary(prcomp(table))

```
Standard deviations:
 5.720782e+02 5.440414e+02 7.801390e+01 7.620405e-01 5.720174e-01
[6] 2.378011e-01 2.283860e-01 2.176200e-02 2.013178e-04


Rotation:
                  PC1           PC2           PC3           PC4
RAJ2000 -3.731621e-06  5.095943e-06 -2.874370e-06  2.063827e-03 -2
DEJ2000 -5.653971e-08 -5.963904e-08 -1.332187e-07 -1.564028e-05 -6
Xpos    -8.815515e-01  4.720855e-01 -1.447687e-03 -2.380232e-04 -4
Ypos     4.720765e-01  8.815475e-01  4.217340e-03  8.608607e-05  4
Imag     1.833989e-05  5.540923e-05 -1.171284e-02  5.049940e-02 -8
U_B      2.443252e-04 -2.410775e-05 -6.546695e-04 -9.561563e-01 -2
```


Plot for Principal Component Analysis
Variances
0
50000
100000
150000
200000
250000
300000

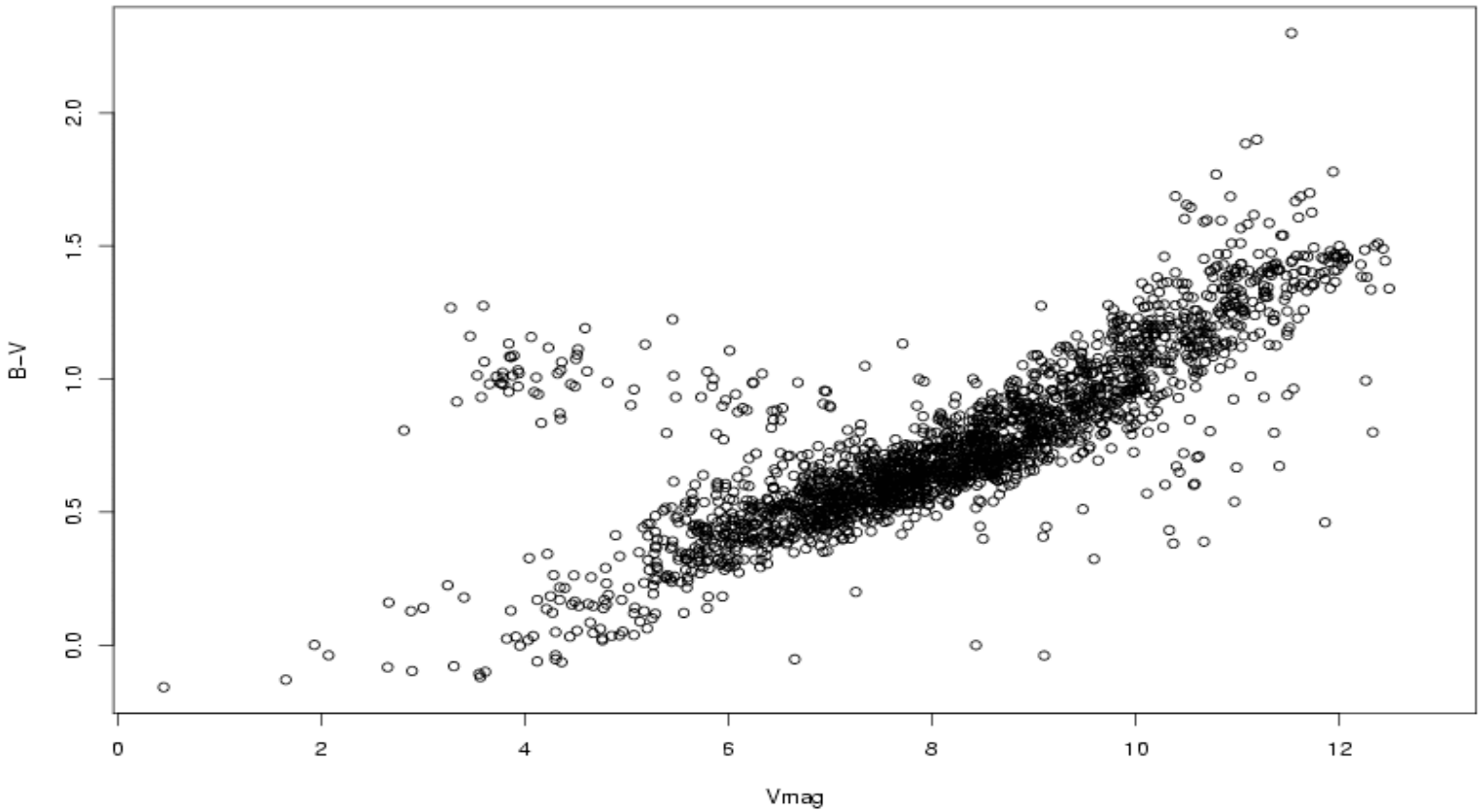
RAJ2000
DEJ2000
Xpos
Ypos
Imag
U_B
B_V
V_I
recno
1.0366
1.0374
600
1600
-1
1
0.6
1.6
189.16
189.24
21
23
25
0
1
2
3
0
100
200


RAJ2000
DEJ2000
Xpos
Ypos
Imag
U_B
B_V
V_I
1.0366
1.0372
600
1600
21
23
26
-1
1
2
0
1
2
3
0.6
1.6
189.16
189.22

Rediscovering HR diagram

- Hyades stars (Hipparcus main catalog)
- Mean/median/boxplot
- Density estimation (Histogram)
- Kernel smoothing
- Correlation matrix
- X-Y plot
- Multivariate clustering

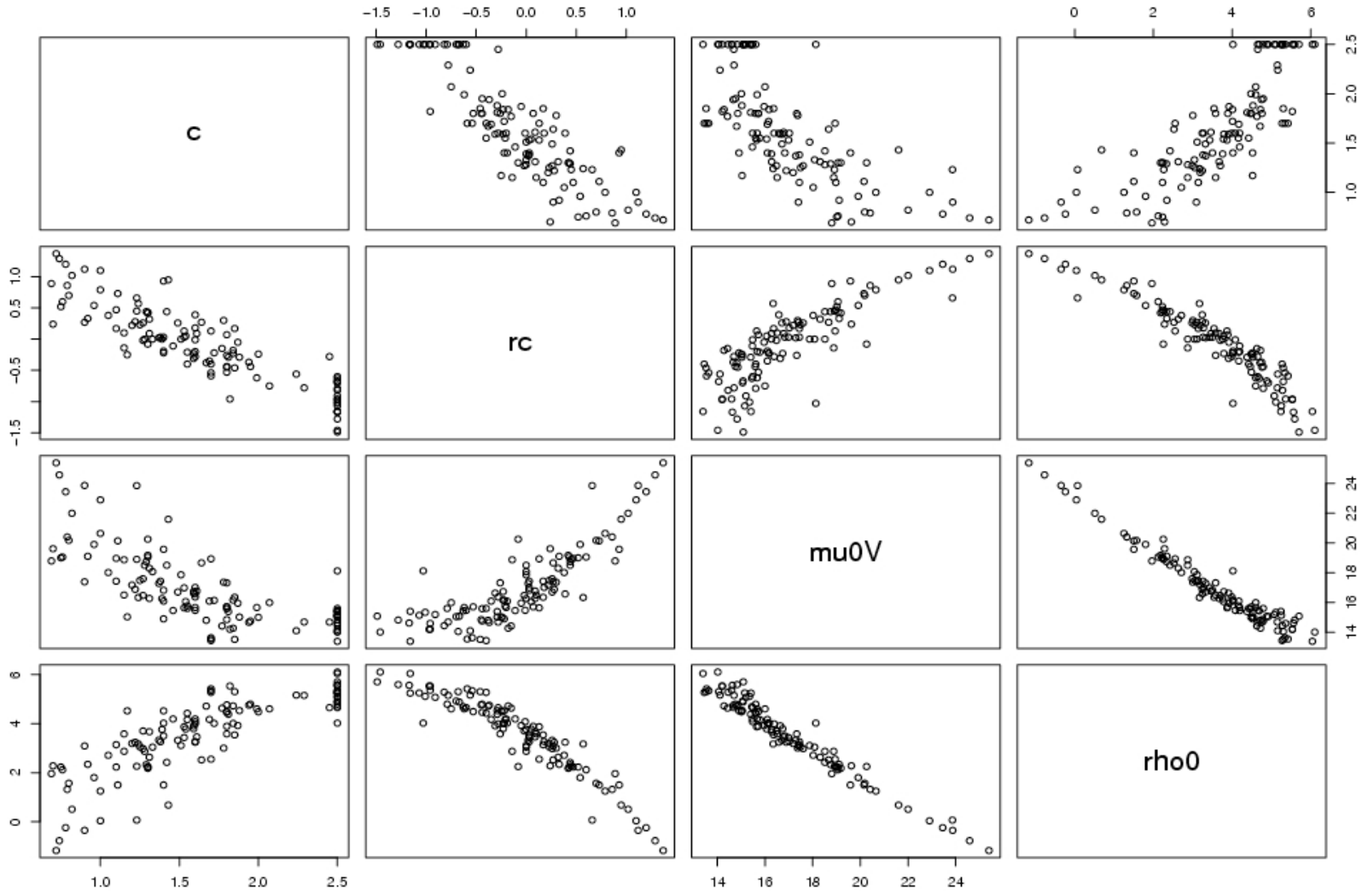
- X-Y plot between Vmag and B-V reveals the famous structure in the dataset: the color-magnitude of bright stars showing the main sequence, giant branch (with red clump stars), and a few Hyades white dwarfs.



FP of Globular clusters

- Matrix of pairwise correlation coefficients
- Pairwise plots
- Principal Component Analysis

- Core parameters as a group tend to be highly correlated, unlike the half-light parameters. This is indicative of the dynamical evolution driven by the core collapse.



- R GUI (R commander http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/)
- http://vostat.org
- http://scc.stat.ucla.edu/mini-courses
- http://cran.r-project.org
- http://www.r-project.org
- http://en.wikipedia.org/wiki/R_(programming_language)