How to Use the R Programming Language for Statistical Analyses

An Introduction to R

(diperkaya dari beberapa sumber tutorial R)

What Is R?

- a programming “environment”
- object-oriented
- similar to S-Plus
- freeware
- provides calculations on matrices
- excellent graphics capabilities
- supported by a large user network

What is R Not?

- a statistics software package
- menu-driven
- quick to learn
- a program with a complex graphical interface

Installing R

- www.r-project.org/
- download from CRAN
- select a download site
- download the base package at a minimum
- download contributed packages as needed

The R Project for Statistical Computing

Important News:

The R Development Core Team would like to formally announce the creation of the

R Foundation for Statistical Computing

There are many reasons for this decision on our part, largely it is based on the belief that R has become a mature and valuable tool and we would like to ensure its continued development and the development of future innovations in software for statistical and computational research.

The R Foundation is a not for profit foundation whose general goals are to provide support for the R project and other innovations in statistical computing. The R Foundation will provide a reference point for individuals, instititutions or commercial enterprises that want to support or interact with the R development community.

We would like to solicit memberships from interested parties (individual and institutional) in the R Foundation. Details regarding fees and membership categories can be obtained from the web site and email enquiries can be sent to R-foundation@R-project.org.

Among the goals of the Foundation are the support of continued development of R, the exploration of new methodology, teaching and training of statistical computing and the organization of meetings and conferences with a statistical computing orientation. We hope to attract sufficient funding to make these goals realities.

For the R Development Core Team:

Robert Gentleman & Ross Ihaka
Presidents, R Foundation

Friedrich Leisch
Secretary General, R Foundation

About R
What is R?
Contributors
Screenshots
What's new?

Download
CRAN

R Project
Foundation
Mailing Lists
Bug Tracking
Developer Page
Search

Documentation
Manuals
FAQs
Contributed
Newsletter
Help Pages
Publications

Related Projects
Bioconductor

About R
What is R?
Contributors
Screenshots
What's new?

Download
CRAN

R Project
Foundation
Mailing Lists
Bug Tracking
Developer Page
Search

Documentation
Manuals
FAQs
Contributed
Newsletter
Help Pages
Publications

Related Projects
Bioconductor

http://cran.za.r-project.org	Rhodes University
Switzerland	
http://cran.ch.r-project.org	ETH Zürich
United Kingdom	
http://cran.uk.r-project.org	University of Bristol
United States of America	
http://cran.us.r-project.org	University of Wisconsin, Madison, WI
http://cran.stat.ucla.edu/	University of California, Los Angeles, CA
http://www.bioconductor.org/CRAN/	Dana Farber Cancer Institute, Boston, MA
http://cran.get-software.com	Get-Software.com, Augusta, ME
http://www.ibiblio.org/pub/languages/R/CRAN/	University of North Carolina, Chapel Hill, NC
http://lib.stat.cmu.edu/R/CRAN/	Statlib, Carnegie Mellon University, Pittsburgh, PA
http://cran.mirrors.pair.com/	Pair Networks, Pittsburgh, PA
http://www.binarycode.org/cran	BinaryCode.org, Austin, TX
http://mirrors.theonlinerecordstore.com/CRAN/	The Online Record Store, Houston, TX

Many of these sites can also be accessed using FTP. In addition, several StatLib mirrors around the world provide a complete CRAN mirror. Please let us know if you want your server being added to the list of mirrors.

The CRAN master site at TU Wien, Austria, can be found at the URLs

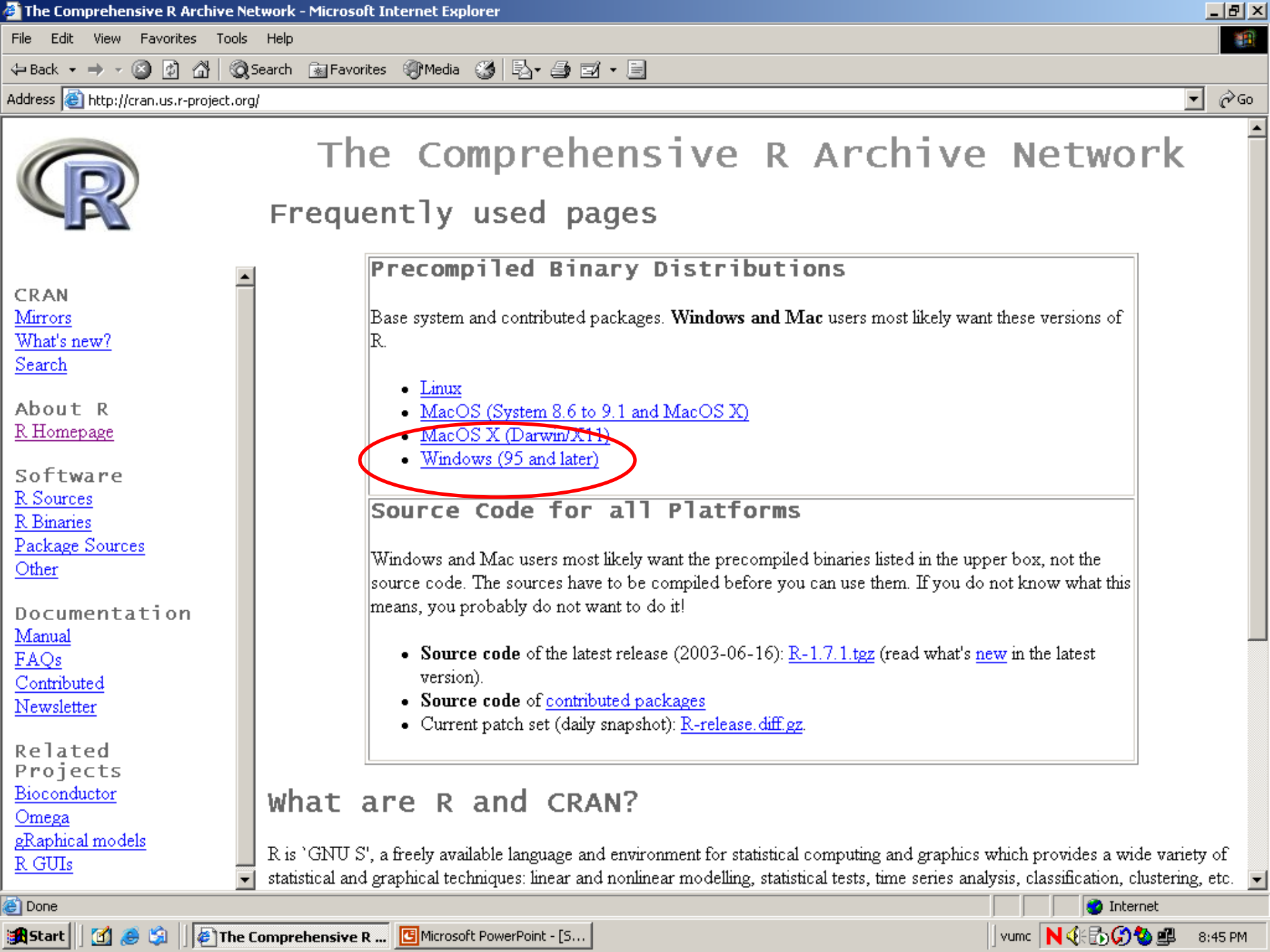
http://cran.r-project.org

ftp://cran.r-project.org/pub/R/

rsync: cran.r-project.org::CRAN

To "submit" to CRAN, simply upload to ftp://cran.r-project.org/incoming and send email to cran@r-project.org. Please indicate the copyright situation (GPL, ...) in your submission.

Last modified: July 4, 2003 by Friedrich Leisch


The Comprehensive R Archive Network - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://cran.us.r-project.org/
Go
The Comprehensive R Archive Network
Frequently used pages
CRAN
Mirrors
What's new?
Search
About R
R Homepage
Software
R Sources
R Binaries
Package Sources
Other
Documentation
Manual
FAQs
Contributed
Newsletter
Related Projects
Bioconductor
Omega
gRaphical models
R GUIs
Precompiled Binary Distributions
Base system and contributed packages. Windows and Mac users most likely want these versions of R.
Linux
MacOS (System 8.6 to 9.1 and MacOS X)
MacOS X (Darwin/X11)
Windows (95 and later)
Source Code for all Platforms
Windows and Mac users most likely want the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
Source code of the latest release (2003-06-16): R-1.7.1.tgz (read what's new in the latest version).
Source code of contributed packages
Current patch set (daily snapshot): R-release.diff.gz.
What are R and CRAN?
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.
Done
Internet
Start
The Comprehensive R ...
Microsoft PowerPoint - [S...
vumc
8:45 PM

R for Windows

This directory contains binaries for a base distribution and packages to run on Windows (NT, 95 and later) on Intel and clones (but not NT on Alpha and other platforms).

CRAN
Mirrors
What's new?
Search

About R
R Homepage

Software
R Sources
R Binaries
Package Sources
Other

Documentation
Manual
FAQs
Contributed
Newsletter

Related Projects
Bioconductor
Omega
gRaphical models
R GUIs

Note: CRAN does not have Windows systems and cannot check these binaries for viruses. Use the normal precautions with downloaded executables.

Subdirectories:

base	Binaries for base distribution (managed by Duncan Murdoch)
contrib	Binaries of contributed packages (managed by Uwe Ligges)
unsupported	Unsupported or obsolete packages

Please send contributions to Duncan Murdoch or Uwe Ligges, not to CRAN.

You may also want to read the R FAQ and R for Windows FAQ.

Last modified: June 3, 2003, by Friedrich Leisch

Tutorials

- From R website under “Documentation”
  - “Manual” is the listing of official R documentation
    - An Introduction to R
    - R Language Definition
    - Writing R Extensions
    - R Data Import/Export
    - R Installation and Administration
    - The R Reference Index

Tutorials cont.

- “Contributed” documentation are tutorials and manuals created by R users
  - Simple R
  - R for Beginners
  - Practical Regression and ANOVA Using R
- R FAQ
- Mailing Lists (listserv)
  - r-help

Tutorials cont.

- Textbooks
  - Venables & Ripley (2002) *Modern Applied Statistics with S.* New York: Springer-Verlag.
  - Chambers (1998). *Programming With Data: A guide to the S language.* New York: Springer-Verlag.

R Basics

- objects
- naming convention
- assignment
- functions
- workspace
- history

Objects

- names
- types of objects: vector, factor, array, matrix, data.frame, ts, list
- attributes
  - mode: numeric, character, complex, logical
  - length: number of elements in object
- creation
  - assign a value
  - create a blank object

Naming Convention

- must start with a letter (A-Z or a-z)
- can contain letters, digits (0-9), and/or periods “.”
- case-sensitive
  - `mydata` different from `MyData`
- do not use underscore “_”

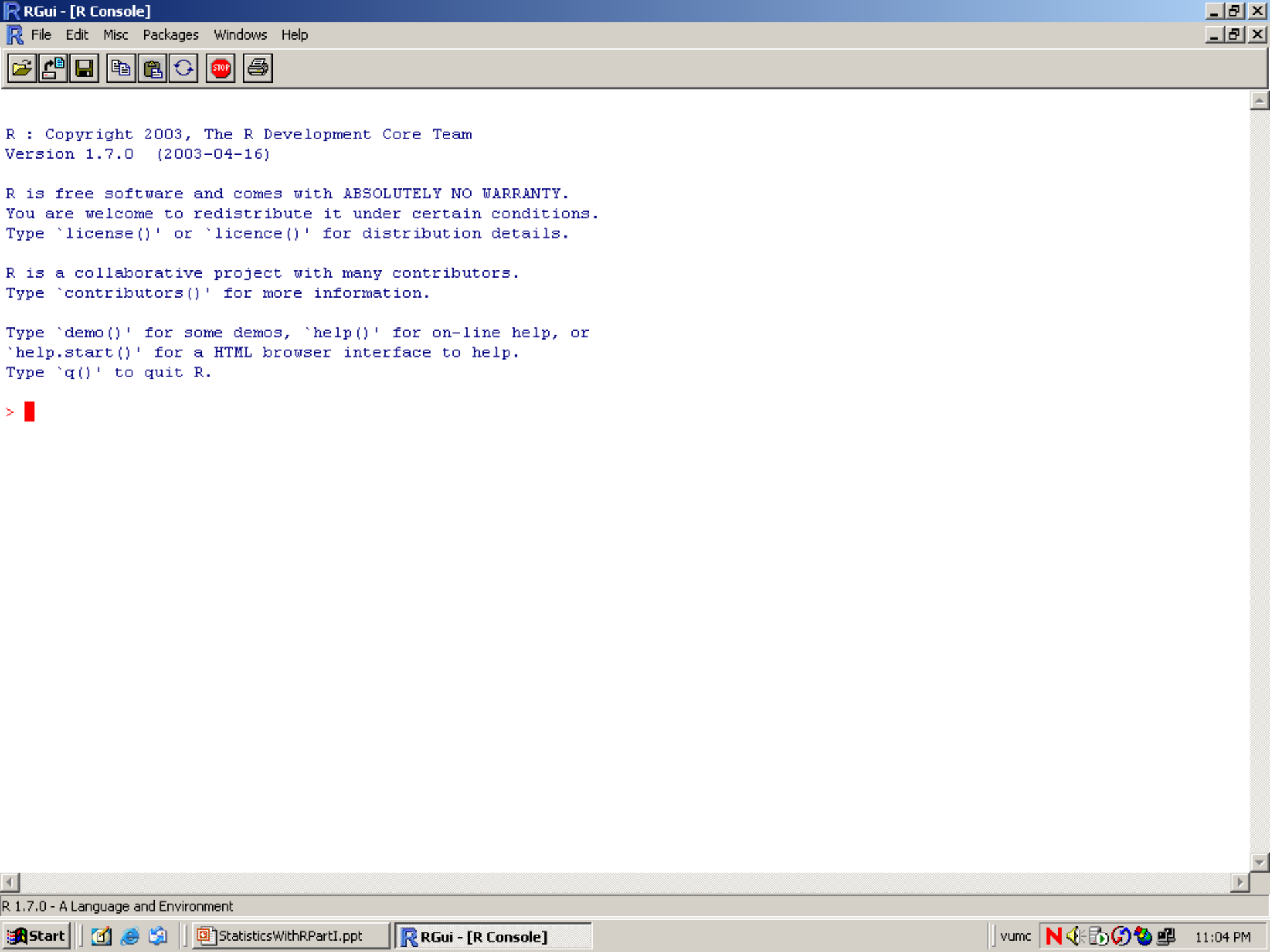
Assignment

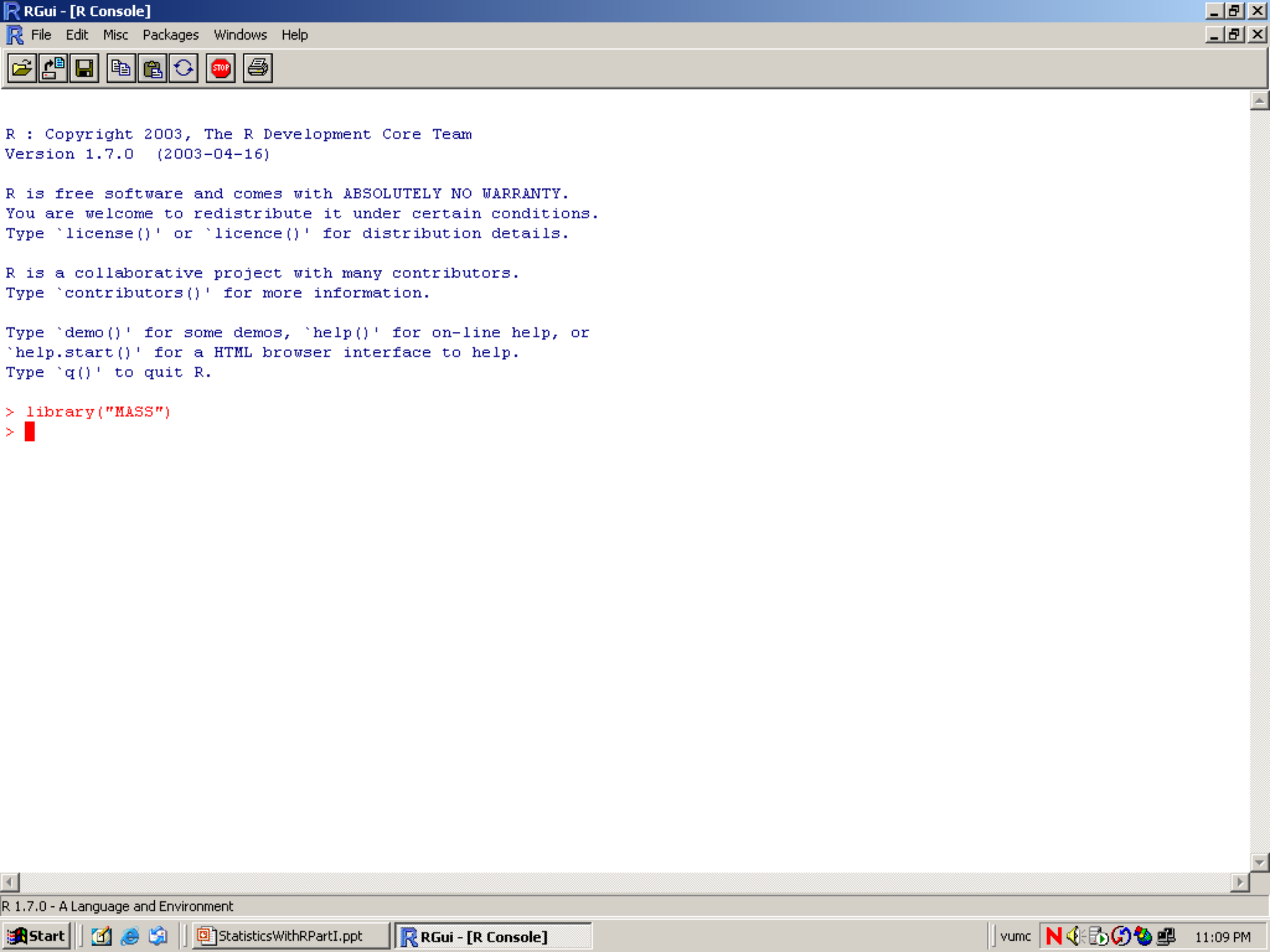
- “<-” used to indicate assignment
  - `x<-c(1,2,3,4,5,6,7)`
  - `x<-c(1:7)`
  - `x<-1:4`

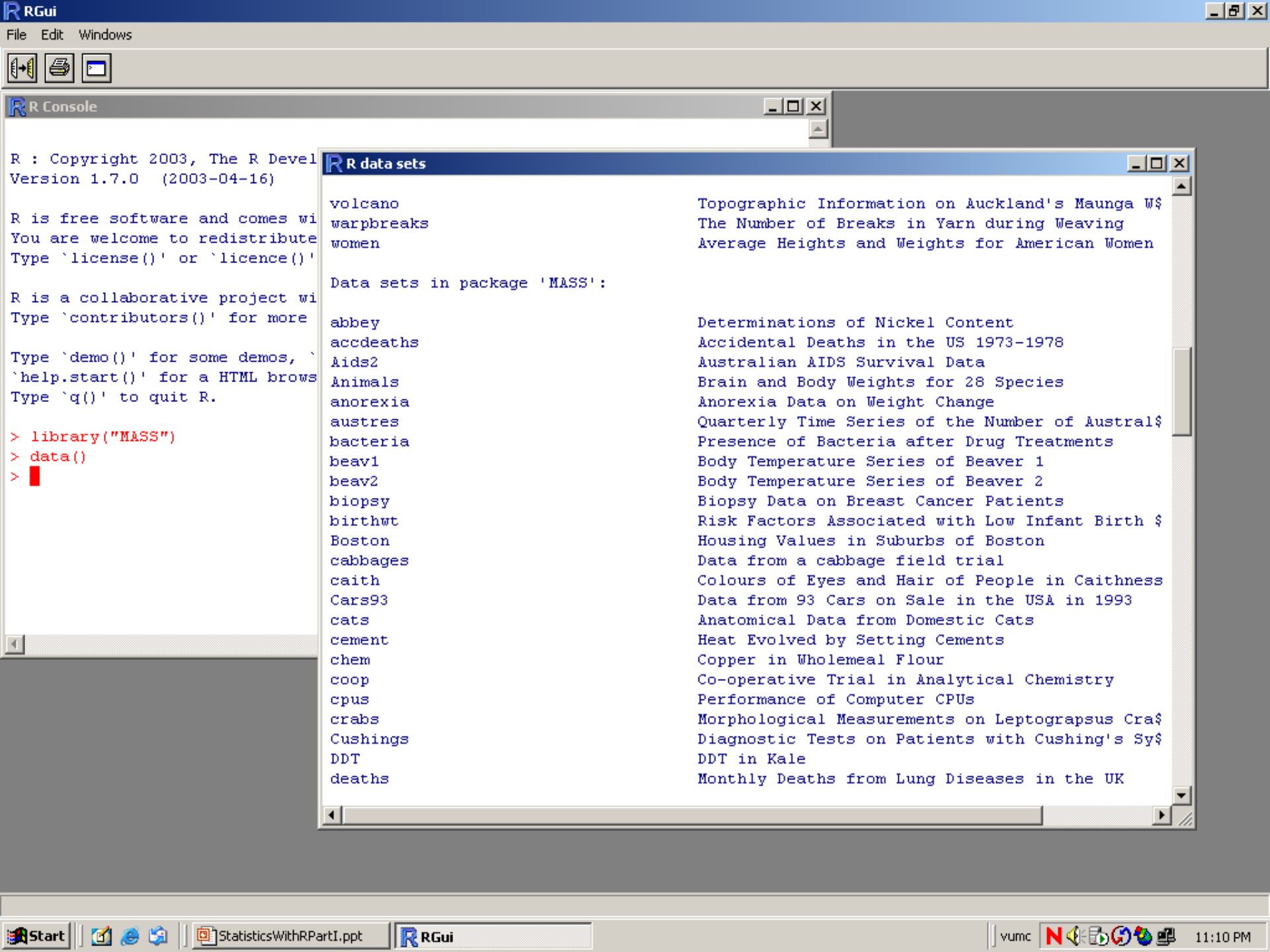
Functions

- actions can be performed on objects using functions (note: a function is itself an object)
- have arguments and options, often there are defaults
- provide a result
- parentheses () are used to specify that a function is being called

Let’s look at R


RGui - [R Console]
File Edit Misc Packages Windows Help
R : Copyright 2003, The R Development Core Team
Version 1.7.0 (2003-04-16)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.
R is a collaborative project with many contributors.
Type `contributors()' for more information.
Type `demo()' for some demos, `help()' for on-line help, or
`help.start()' for a HTML browser interface to help.
Type `q()' to quit R.
>
R 1.7.0 - A Language and Environment
Start
StatisticsWithRPartI.ppt
RGui - [R Console]
vumc
11:04 PM


RGui - [R Console]
File Edit Misc Packages Windows Help
R : Copyright 2003, The R Development Core Team
Version 1.7.0 (2003-04-16)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.
R is a collaborative project with many contributors.
Type `contributors()' for more information.
Type `demo()' for some demos, `help()' for on-line help, or
`help.start()' for a HTML browser interface to help.
Type `q()' to quit R.
> library("MASS")
>
R 1.7.0 - A Language and Environment
Start
StatisticsWithRPartI.ppt
RGui - [R Console]
vumc
11:09 PM


RGui
File Edit Windows
R Console
R : Copyright 2003, The R Devel
Version 1.7.0 (2003-04-16)
R is free software and comes wi
You are welcome to redistribute
Type `license()' or `licence()'
R is a collaborative project wi
Type `contributors()' for more
Type `demo()' for some demos, `
`help.start()' for a HTML brows
Type `q()' to quit R.
> library("MASS")
> data()
R data sets
volcano Topographic Information on Auckland's Maunga W$
warpbreaks The Number of Breaks in Yarn during Weaving
women Average Heights and Weights for American Women
Data sets in package 'MASS':
abbey Determinations of Nickel Content
accdeaths Accidental Deaths in the US 1973-1978
Aids2 Australian AIDS Survival Data
Animals Brain and Body Weights for 28 Species
anorexia Anorexia Data on Weight Change
austres Quarterly Time Series of the Number of Austral$
bacteria Presence of Bacteria after Drug Treatments
beav1 Body Temperature Series of Beaver 1
beav2 Body Temperature Series of Beaver 2
biopsy Biopsy Data on Breast Cancer Patients
birthwt Risk Factors Associated with Low Infant Birth $
Boston Housing Values in Suburbs of Boston
cabbages Data from a cabbage field trial
caith Colours of Eyes and Hair of People in Caithness
Cars93 Data from 93 Cars on Sale in the USA in 1993
cats Anatomical Data from Domestic Cats
cement Heat Evolved by Setting Cements
chem Copper in Wholemeal Flour
coop Co-operative Trial in Analytical Chemistry
cpus Performance of Computer CPUs
crabs Morphological Measurements on Leptograpsus Cra$
Cushings Diagnostic Tests on Patients with Cushing's Sy$
DDT DDT in Kale
deaths Monthly Deaths from Lung Diseases in the UK
Start
StatisticsWithRPartI.ppt
RGui
vumc
11:10 PM

RGui - [R Console]

File Edit Misc Packages Windows Help

```
R : Copyright 2003, The R Development Core Team
Version 1.7.0  (2003-04-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.

R is a collaborative project with many contributors.
Type `contributors()' for more information.

Type `demo()' for some demos, `help()' for on-line help, or
`help.start()' for a HTML browser interface to help.
Type `q()' to quit R.

> library("MASS")
> data()
> data(Cars93)
> Cars93
    Manufacturer          Model    Type Min.Price Price Max.Price MPG.city
1          Acura        Integra   Small      12.9  15.9      18.8       25
2          Acura         Legend Midsize      29.2  33.9      38.7       18
3           Audi             90 Compact      25.9  29.1      32.3       20
4           Audi            100 Midsize      30.8  37.7      44.6       19
5            BMW           535i Midsize      23.7  30.0      36.2       22
6          Buick        Century Midsize      14.2  15.7      17.3       22
7          Buick        LeSabre   Large      19.9  20.8      21.7       19
8          Buick     Roadmaster   Large      22.6  23.7      24.9       16
9          Buick        Riviera Midsize      26.3  26.3      26.3       19
10      Cadillac        DeVille   Large      33.0  34.7      36.3       16
11      Cadillac        Seville Midsize      37.5  40.1      42.7       16
12     Chevrolet       Cavalier Compact       8.5  13.4      18.3       25
13     Chevrolet        Corsica Compact      11.4  11.4      11.4       25
14     Chevrolet         Camaro  Sporty      13.4  15.1      16.8       19
15     Chevrolet         Lumina Midsize      13.4  15.9      18.4       21
16     Chevrolet     Lumina_APV     Van      14.7  16.3      18.0       18
17     Chevrolet          Astro     Van      14.7  16.6      18.6       15
18     Chevrolet        Caprice   Large      18.0  18.8      19.6       17
```

R 1.7.0 - A Language and Environment

Start | StatisticsWithRPartI.ppt | RGui - [R Console] | vumc | 11:11 PM

R Workspace & History

Workspace

- during an R session, all objects are stored in a temporary, working memory
- list objects
  - `ls()`
- remove objects
  - `rm()`
- objects that you want to access later must be saved in a “workspace”
  - from the menu bar: File->save workspace
  - from the command line: `save(x,file="MyData.Rdata")`

RGui - [R Console]

File Edit Misc Packages Windows Help

```
59       Mercedes-Benz 300E
60            Mercury Capri
61          Mercury Cougar
62         Mitsubishi Mirage
63       Mitsubishi Diamante
64            Nissan Sentra
65            Nissan Altima
66             Nissan Quest
67            Nissan Maxima
68        Oldsmobile Achieva
69 Oldsmobile Cutlass_Ciera
70     Oldsmobile Silhouette
71   Oldsmobile Eighty-Eight
72            Plymouth Laser
73            Pontiac LeMans
74           Pontiac Sunbird
75          Pontiac Firebird
76        Pontiac Grand_Prix
77        Pontiac Bonneville
78                  Saab 900
79                Saturn SL
80              Subaru Justy
81            Subaru Loyale
82            Subaru Legacy
83             Suzuki Swift
84            Toyota Tercel
85            Toyota Celica
86             Toyota Camry
87            Toyota Previa
88          Volkswagen Fox
89        Volkswagen Eurovan
90         Volkswagen Passat
91        Volkswagen Corrado
92                  Volvo 240
93                  Volvo 850
> ls()
[1] "Cars93"
>
```

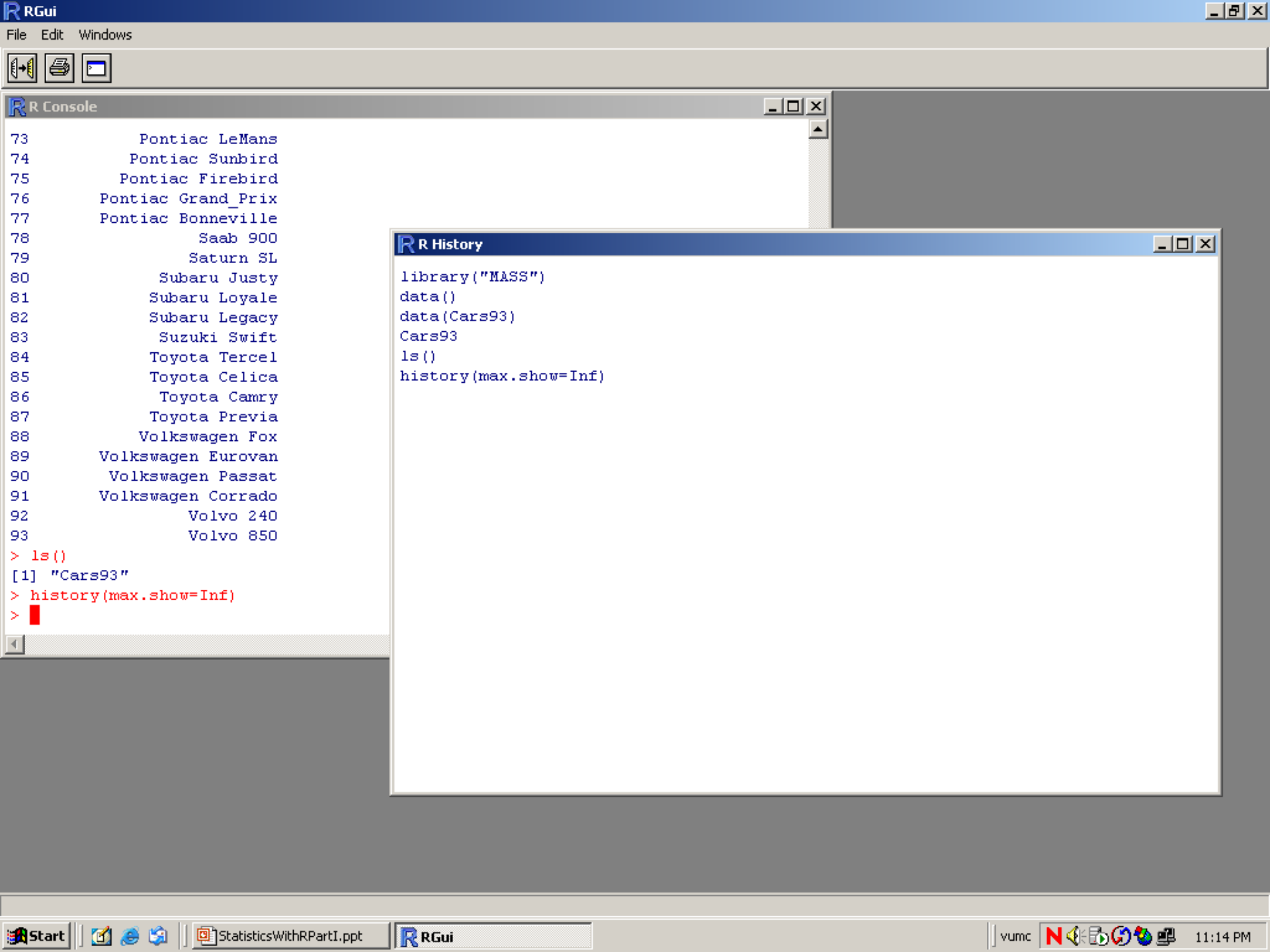
R 1.7.0 - A Language and Environment

Start | StatisticsWithRPartI.ppt | RGui - [R Console] | vumc | 11:12 PM

History

- command line history
- can be saved, loaded, or displayed
  - `savehistory(file="MyData.Rhistory)`
  - `loadhistory(file="MyData.Rhistory)`
  - `history(max.show=Inf)`

- during a session you can use the arrow keys to review the command history


RGui
File Edit Windows
R Console
73 Pontiac LeMans
74 Pontiac Sunbird
75 Pontiac Firebird
76 Pontiac Grand_Prix
77 Pontiac Bonneville
78 Saab 900
79 Saturn SL
80 Subaru Justy
81 Subaru Loyale
82 Subaru Legacy
83 Suzuki Swift
84 Toyota Tercel
85 Toyota Celica
86 Toyota Camry
87 Toyota Previa
88 Volkswagen Fox
89 Volkswagen Eurovan
90 Volkswagen Passat
91 Volkswagen Corrado
92 Volvo 240
93 Volvo 850
> ls()
[1] "Cars93"
> history(max.show=Inf)
>
R History
library("MASS")
data()
data(Cars93)
Cars93
ls()
history(max.show=Inf)
Start
StatisticsWithRPartI.ppt
RGui
vumc
11:14 PM

Two most common object types for statistics:

matrix

data frame

Matrix

- a matrix is a vector with an additional attribute (dim) that defines the number of columns and rows
- only one mode (numeric, character, complex, or logical) allowed
- can be created using `matrix()`

```
x<-matrix(data=0,nr=2,nc=2)
        or
x<-matrix(0,2,2)
x=matrix(c(1,2,3,4,5,6,7,8,9),nr=3,nc=3)
```

Data Elements

- select only one element
  - `x[2]`
- select range of elements
  - `x[1:3]`
- select all but one element
  - `x[-3]`
- slicing: including only part of the object
  - `x[c(1,2,5)]`
- select elements based on logical operator
  - `x(x>3)`

Basic Statistics

1. Qualitative Data
2. Quantitative Data

    a. Descriptive
    b. Classification
    c. Plotting
    d. Variance
    e. Analysis of Variance

Exercise

- Open library : `data(airquality)`
- Select coloumn Ozone : `airquality$Temp`
- Several descriptive calculation:
  - `mean()`
  - `median()`
  - Range : `max() and min()` or `range()`
  - `quantile()`
  - Percentile : `quantile(data,c(P1,P2,P3,…))`
  - Standard Deviation : `sd()`
  - Correlation : `cor(x,y)`
  - `summary()`

Exercise

- Open library : `data(airquality)`
- Select coloumn Ozone : `airquality$Temp`
  - Range : `max() and min()` or `range()`
  - `class<-seq(a,b,by=..)`
  - `A.cut<-cut(A,class,right=FALSE)`

Exercise

- Plotting several information:
  - Histogram : `hist ()`
  - Barplot : `barplot()`
    - `color=c("red","yellow","green","blue","cyan")`
    - `barplot(..,col=color)`
  - Pie Chart : `pie()`
  - Scatter plot : `plot(x,y)`
  - Boxplot: `boxplot(x,y)`
  - Cummulative `Plot : plot(x,y)`; x = class, `y=c(0,cumsum())`

STATISTIK INFERENSIAL

Estimasi -1

We take a sample from a population to learn about (i.e., estimate) population parameters such as:

Population Parameter:	Mean μ	Prop. π	Std. Dev. σ	Corr. ρ	Slope β
Estimate or Statistic:	$\hat{\mu} = \bar{x}$	$\hat{\pi} = p$	$\hat{\sigma} = s$	$\hat{\rho} = r$	$\hat{\beta} = b$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (average)

p is the sample proportion

$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation

r is the sample correlation coefficient

b is the estimated slope in a linear regression model

Estimasi -2

$$\text{var}\left(\bar{X}\right) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

- Hence, the standard deviation of the sample mean is

$$\sqrt{\text{var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- The larger the sample size n, the smaller the spread of the sample mean $\bar{X}$.
- But not proportionally: taking 4 times as large a sample only reduces the spread (=precision) of $\bar{X}$ by a factor of 2!

Inference: Hypothesis Test for π

- **Step 1:** Specify null and alternative hypotheses (always about a population parameter):

two-sided:	$H_0 : \pi = \pi_0$	$H_A : \pi \neq \pi_0$
one-sided:	$H_0 : \pi \leq \pi_0$	$H_A : \pi > \pi_0$
one-sided:	$H_0 : \pi \geq \pi_0$	$H_A : \pi < \pi_0,$

  where π **is the true (unknown) proportion** and π_0 is some specific value. Also, choose α-level (controls type I error, see later).

- **Step 2:** Specify the (asymptotic) distribution for the estimator of the unknown parameter. In almost all cases: apply the CLT assuming H_0 is true:

$$P \sim N(\pi_0, \pi_0(1-\pi_0)/n)$$

Inference: Hypothesis Test for π

- **Step 3 (P-value):** Assuming that H_0 is true, find the probability of observing an even more extreme (as specified by the alternative hypothesis) sample proportion as the one observed:

  I.e., in the one sided case with H_A: $\pi > \pi_0$, find $\Pr(P > p)$, where p is the observed proportion.

  In the two-sided case, find $2 \times \Pr(P > |p|)$.

  To find this, calculate the **Test Statistic:** under H_0

$$Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

Inference: Hypothesis Test for π

- Calculate the **Test Statistic:** under H_0

$$Z = \frac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

The probability $\Pr(P > p)$ is the same as $\Pr(Z > z)$, and $2 \times \Pr(P > |p|)$ is the same as $2 \times \Pr(Z > |z|)$, where Z is a standard normal random variable (i.e., $Z \sim N(0,1)$).

The probability under the $N(0,1)$ model is easy to calculate (Tables, R). The resulting probability is known as the P-value.

Inference: Hypothesis Test for π

- **Step 4 (Conclusion):**
  If P-value $< \alpha$: **Sufficient evidence** for H_A.
  - H_0 is no more tenable, reject it. The likelihood of observing such a sample proportion when the null hypothesis is true is so small, so that the null hypothesis must be wrong.

  If P-value $\geq \alpha$: **Insufficient evidence**. Cannot reject the claim H_0, therefore retain it.
  - The sample did not provide overwhelming evidence to reject the null hypothesis. The likelihood of observing such a sample proportion is not so small when the null hypothesis is correct. Therefore, no reason to reject it.
- How to choose α?

Exercise – One tailed Test (Known variance)

```
> xbar = 9900              # sample mean
> mu0 = 10000              # hypothesized value
> sigma = 120              # population standard deviation
> n = 30                   # sample size
> z = (xbar−mu0)/(sigma/sqrt(n))
> z                        # test statistic
[1] −4.5644

> alpha = .05
> z.alpha = qnorm(1−alpha)
> -z.alpha                 # critical value
[1] −1.6449
```

Exercise

- Buka data airquality, pilih Temp
- Hitung rerata Temp dan standar deviasinya.
- Buat sample secara random dari data Temp sebanyak 100 data
  - Y<- sample(X,100)
- Hitung rerata dari sample Y
- Hitung Nilai Z sample dan ujilah null hypothesis

Exercise – Two tailed Test (Known variance)

- > xbar = 14.6 # sample mean
  > mu0 = 15.4 # hypothesized value
  > sigma = 2.5 # population standard deviation
  > n = 35 # sample size
  > z = (xbar−mu0)/(sigma/sqrt(n))
  > z # test statistic
  [1] −1.8931
- > alpha = .05
  > z.half.alpha = qnorm(1−alpha/2)
  > c(−z.half.alpha, z.half.alpha)
  [1] −1.9600 1.9600

Exercise – One tailed Test (Unknown variance)

- > xbar = 9900 # sample mean
  > mu0 = 10000 # hypothesized value
  > s = 125 # sample standard deviation
  > n = 30 # sample size
  > t = (xbar−mu0)/(s/sqrt(n))
  > t # test statistic
  [1] −4.3818
- > alpha = .05
  > t.alpha = qt(1−alpha, df=n−1)
  > −t.alpha # critical value
  [1] −1.6991

Exercise

- Buka data airquality, pilih Temp
- Hitung rerata Temp dan standar deviasinya.
- Buat sample secara random dari data Temp sebanyak 100 data
  - Y<- sample(X,100)
- Hitung rerata dari sample Y
- Hitung Nilai Z sample dan ujilah null hypothesis

Exercise – Two tailed Test (Unknown variance)

- > xbar = 14.6 # sample mean
  > mu0 = 15.4 # hypothesized value
  > s = 2.5 # sample standard deviation
  > n = 35 # sample size
  > t = (xbar−mu0)/(s/sqrt(n))
  > t # test statistic
  [1] −1.8931
- > alpha = .05
  > t.half.alpha = qt(1−alpha/2, df=n−1)
  > c(−t.half.alpha, t.half.alpha)
  [1] −2.0322 2.0322

Alternative – Two tailed Test (Unknown variance)

- > pval = 2 ∗ pt(t, df=n−1) # lower tail
  > pval # two−tailed p−value
  [1] 0.066876

- Uji pval> 0.05 → Do not Reject
- Uji pval <=0.05 → Reject

- t.test(x,y) #uji t untuk variable x dan y