Introduction to R Statistical Software

Anthony (Tony) R. Olsen
USEPA ORD NHEERL
Western Ecology Division
Corvallis, OR 97333
(541) 754-4790
Olsen.Tony@epa.gov

What is R?

- A language and environment for statistical computing and graphics
- Based on the “S” system developed by Bell Laboratories
- From R Project for Statistical Computing
- R is available as Free Software
- S-Plus is a commercial implementation of the “S” system

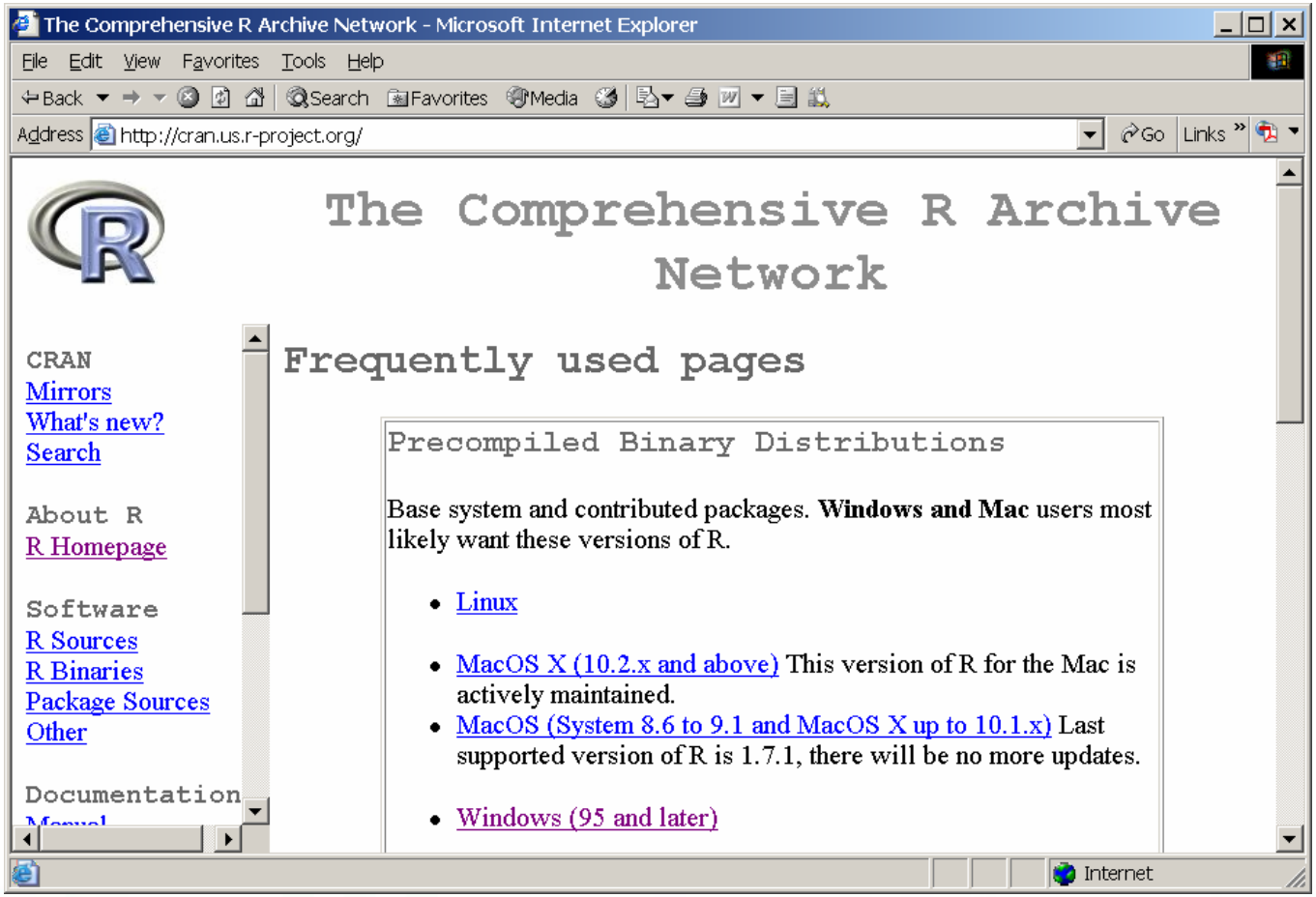
The R Environment is an integrated suite of software

- effective data handling and storage
- calculations on spreadsheets and arrays
- integrated collection of statistical analysis tools
- graphical display on-screen or hardcopy
- a programming language
- add capability using packages (libraries)

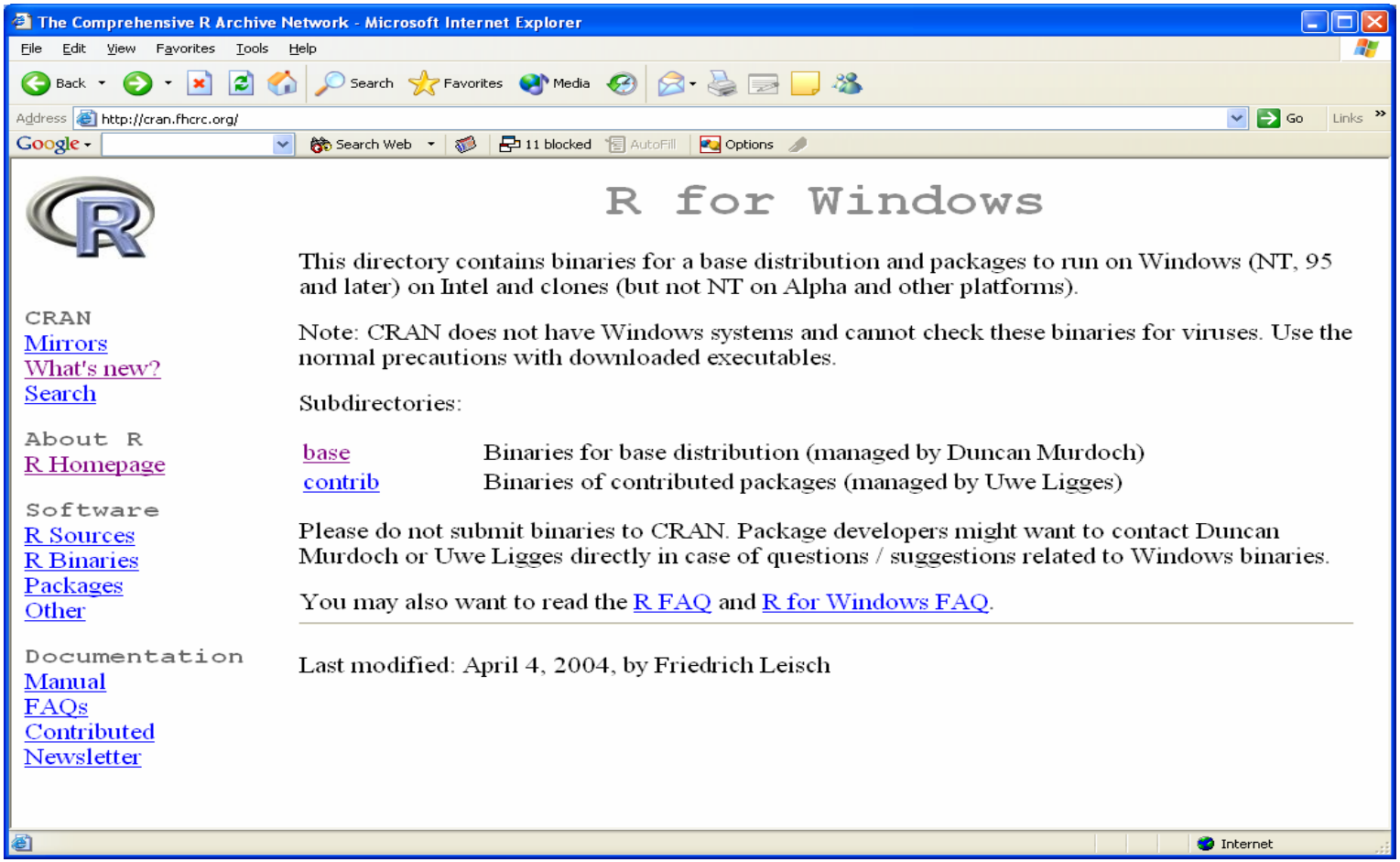
Why use R for Aquatic Monitoring Survey Analysis

- Needed way to provide statistical estimation algorithms to states and tribal nations
- No existing statistical software has the algorithms (at least not all) required
- States & tribal nations limited ability to purchase any recommended commercial statistical software
- R is free, easily installed, and usable with training

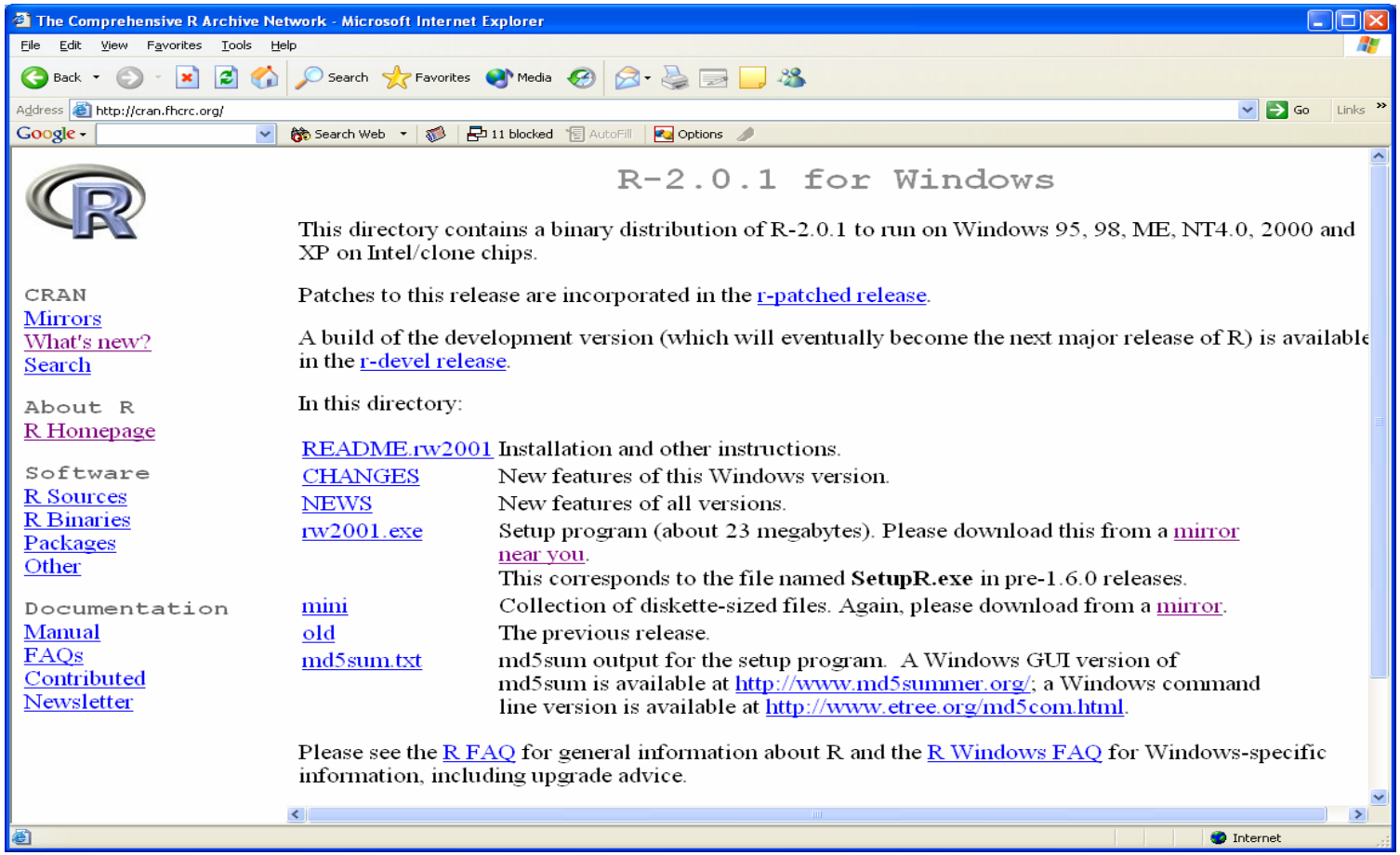
Acquiring and Installing R
Use web browser to access http://cran.us.r-project.org/



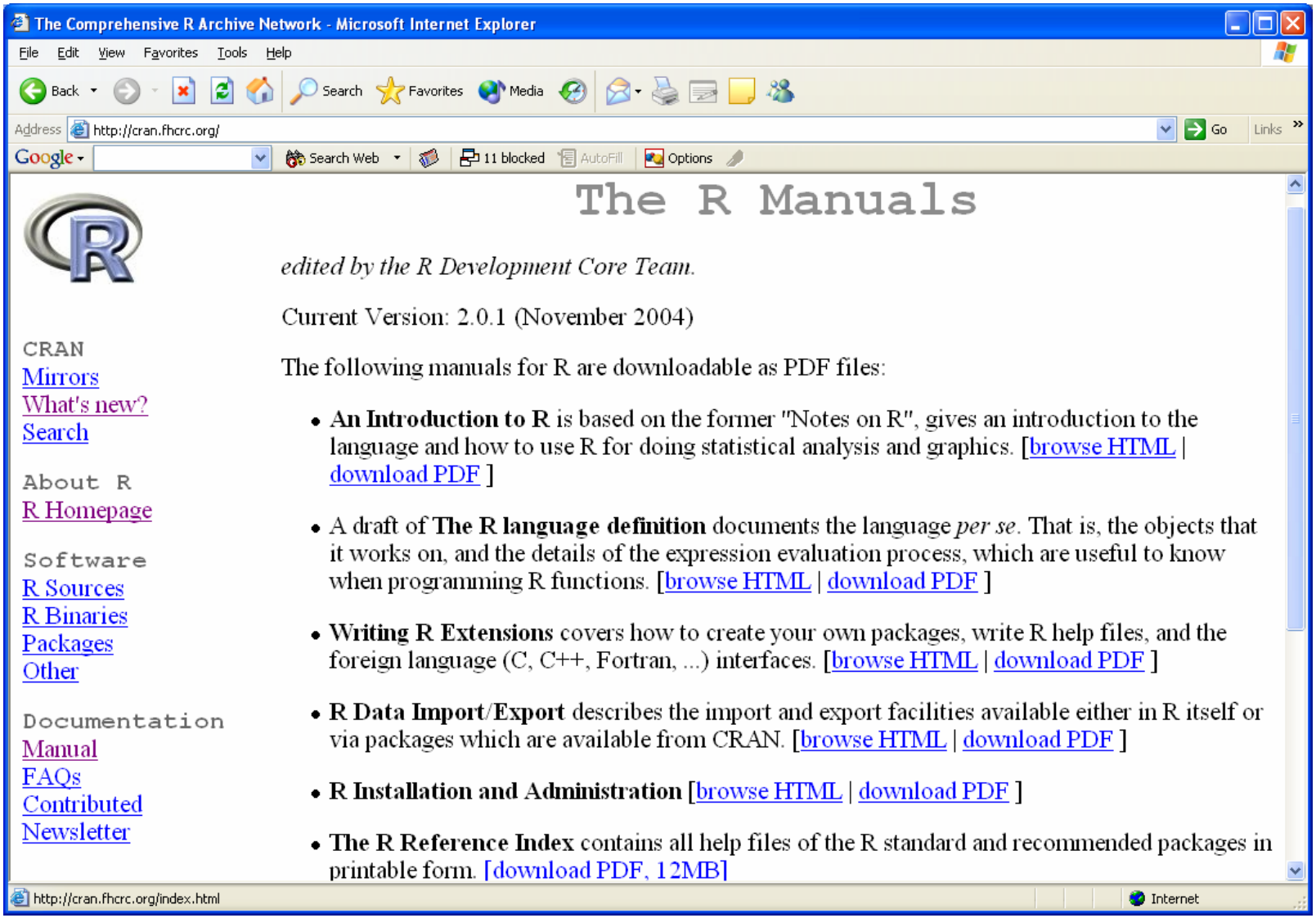
Click on base



Click on CHANGES to see changes
Click on README.rw2001 for instructions
Click on rw2001.exe to download



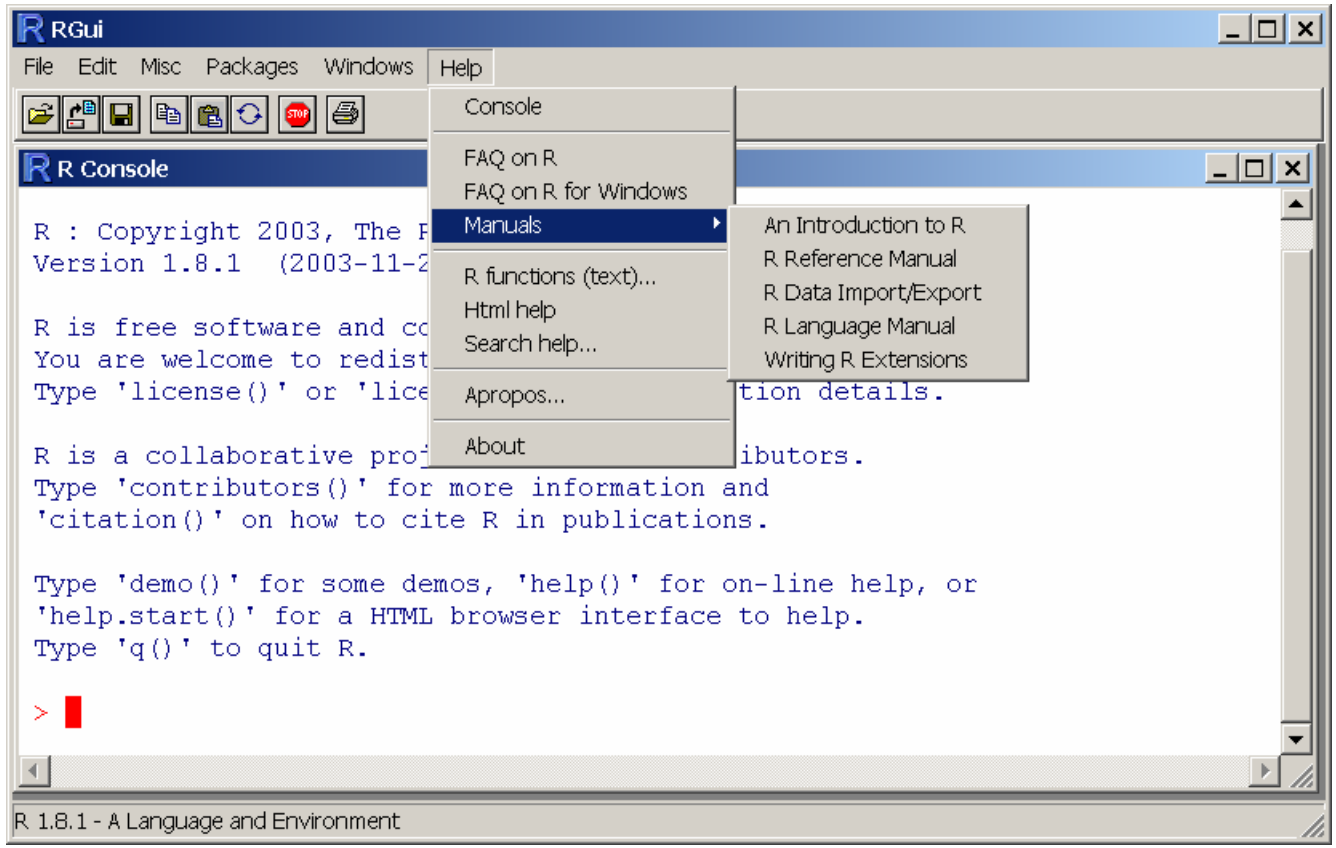
Documentation on Using R



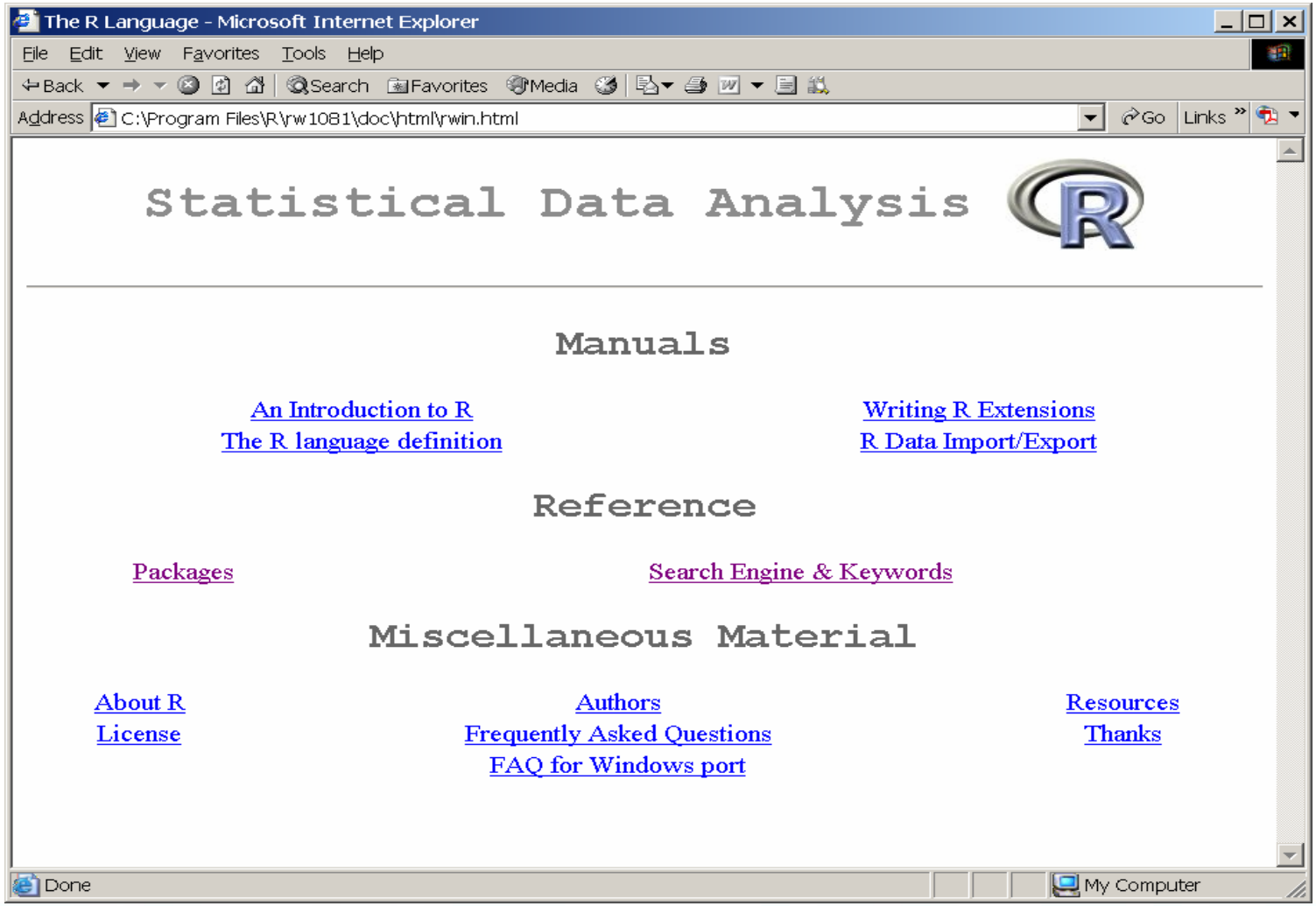
Installing R

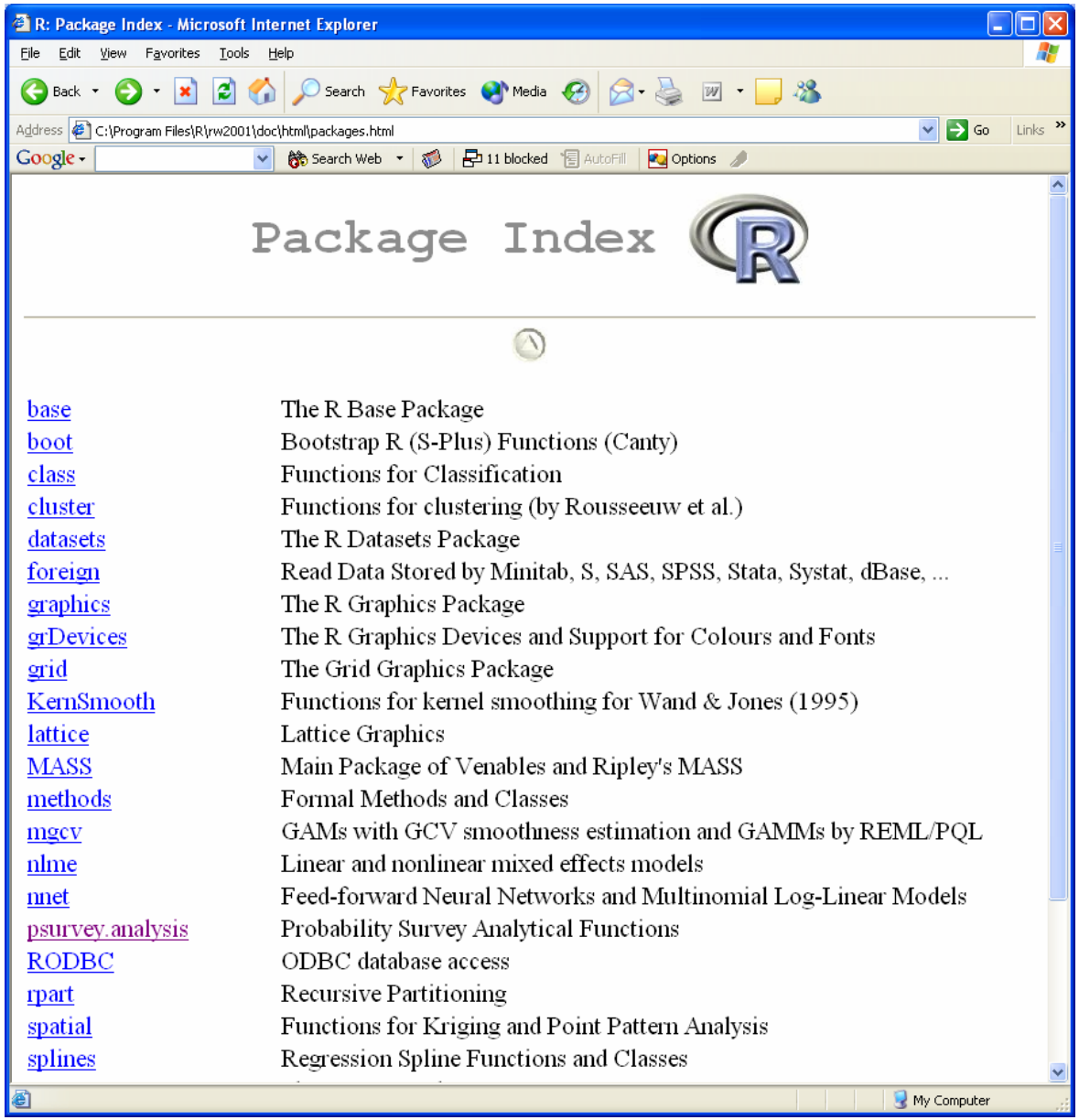
- Save the downloaded file to the desktop
- Uninstall any previous versions of R, if any
- Close all other programs
- Install the downloaded file by double clicking it
  - Standard Windows installer
  - Puts R in your Start Menu
  - Places an R icon on desktop
  - May take 5-10 minutes

- Start R: Double click R desktop icon
- Note documentation is available from Help Menu
  - Manuals
  - Html help



Accessing Help on commands: base and packages

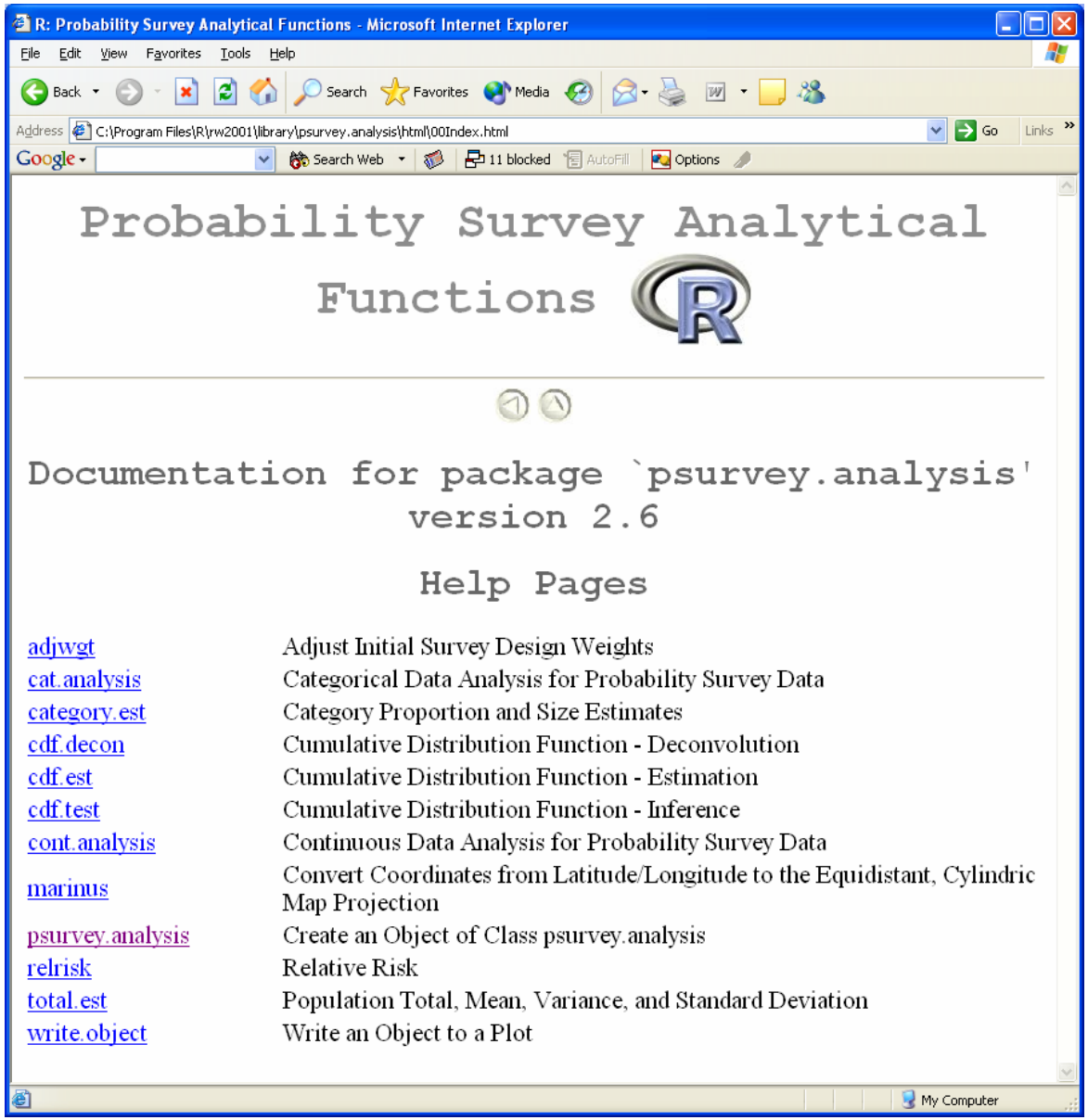


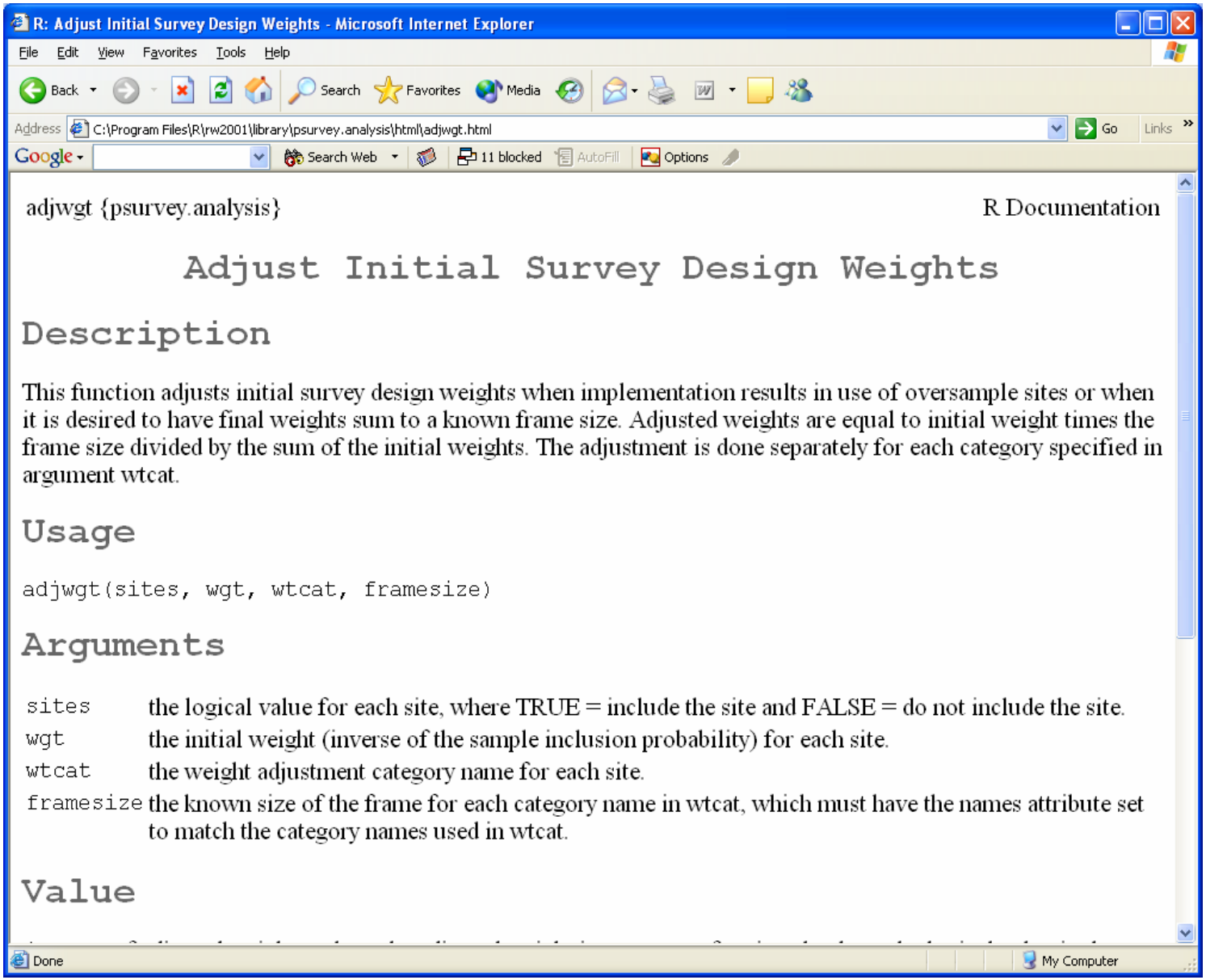


List of packages
Key ones:

base

psurvey.analysis

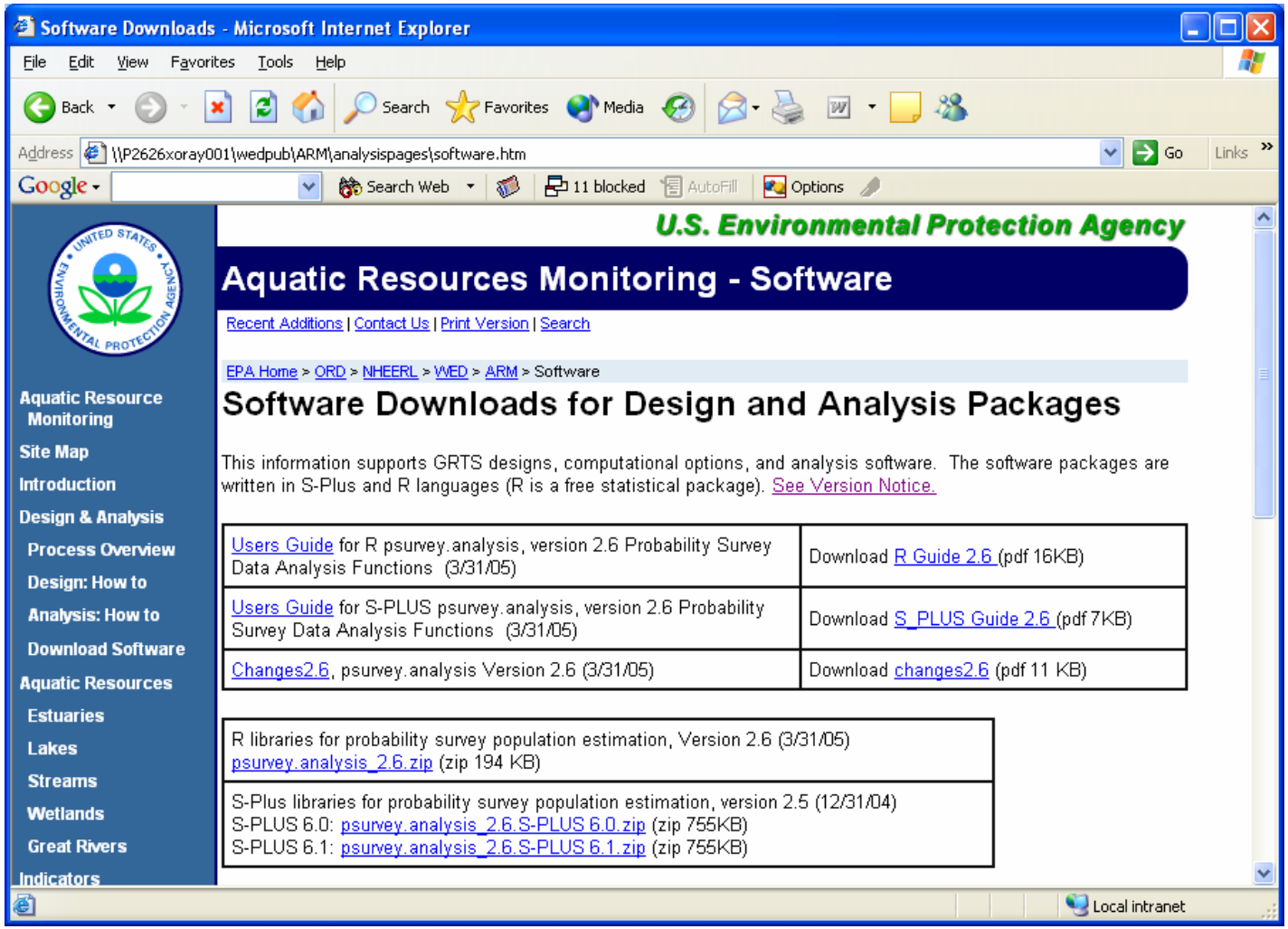

R: Probability Survey Analytical Functions - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Program Files\R\rw2001\library\psurvey.analysis\html\00Index.html
Go Links
Google Search Web 11 blocked AutoFill Options
Probability Survey Analytical Functions
Documentation for package `psurvey.analysis' version 2.6
Help Pages
adjwgt Adjust Initial Survey Design Weights
cat.analysis Categorical Data Analysis for Probability Survey Data
category.est Category Proportion and Size Estimates
cdf.decon Cumulative Distribution Function - Deconvolution
cdf.est Cumulative Distribution Function - Estimation
cdf.test Cumulative Distribution Function - Inference
cont.analysis Continuous Data Analysis for Probability Survey Data
marinus Convert Coordinates from Latitude/Longitude to the Equidistant, Cylindric Map Projection
psurvey.analysis Create an Object of Class psurvey.analysis
relrisk Relative Risk
total.est Population Total, Mean, Variance, and Standard Deviation
write.object Write an Object to a Plot
My Computer


R: Adjust Initial Survey Design Weights - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Program Files\R\rw2001\library\psurvey.analysis\html\adjwgt.html
Go Links
Google Search Web 11 blocked AutoFill Options
adjwgt {psurvey.analysis} R Documentation
Adjust Initial Survey Design Weights
Description
This function adjusts initial survey design weights when implementation results in use of oversample sites or when it is desired to have final weights sum to a known frame size. Adjusted weights are equal to initial weight times the frame size divided by the sum of the initial weights. The adjustment is done separately for each category specified in argument wtcat.
Usage
adjwgt(sites, wgt, wtcat, framesize)
Arguments
sites the logical value for each site, where TRUE = include the site and FALSE = do not include the site.
wgt the initial weight (inverse of the sample inclusion probability) for each site.
wtcat the weight adjustment category name for each site.
framesize the known size of the frame for each category name in wtcat, which must have the names attribute set to match the category names used in wtcat.
Value
Done
My Computer

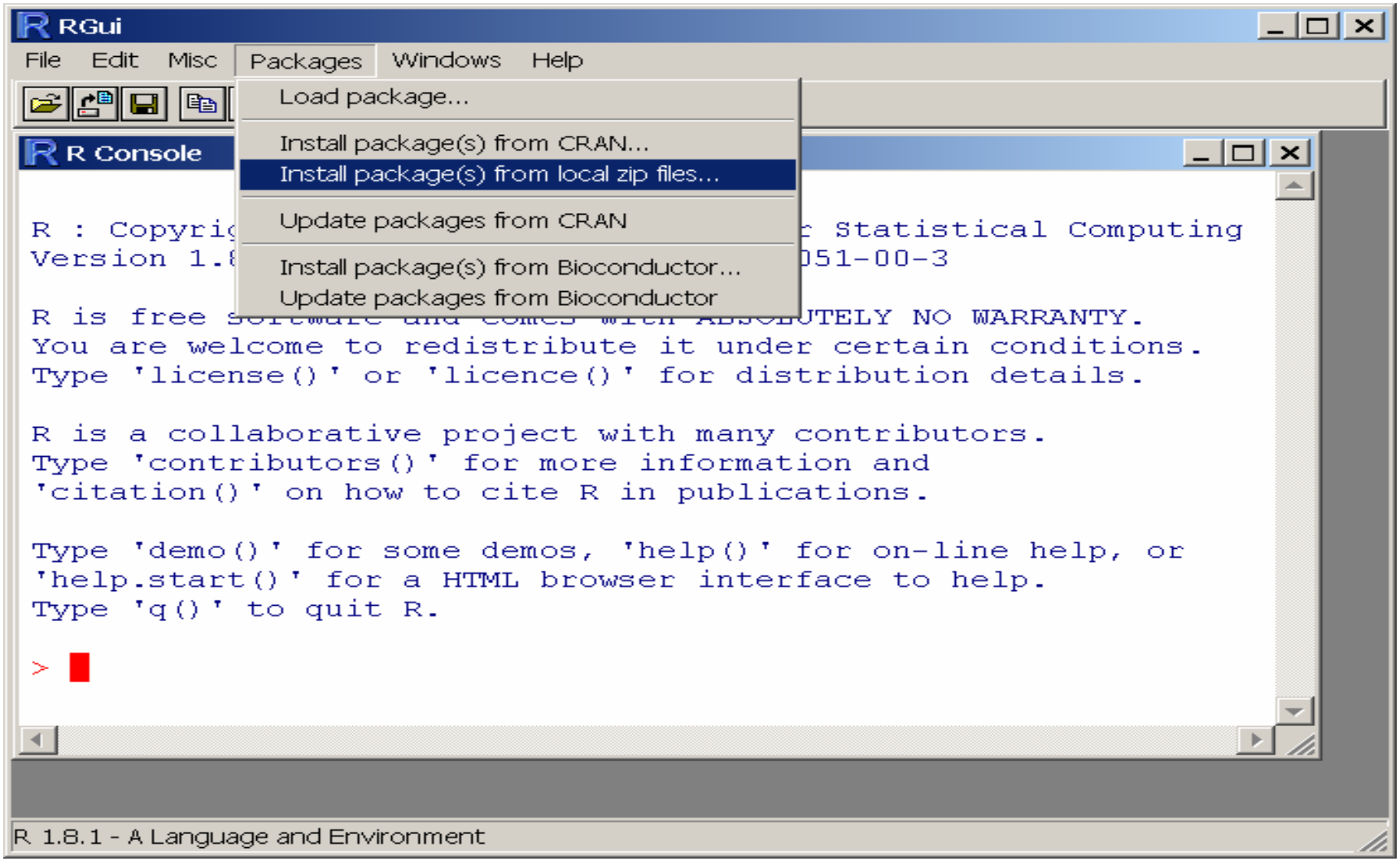
Installing psurvey.analysis Library

- Download library from ARM web site
  - http://www.epa.gov/nheerl/arm/
  - Library on page: http://www.epa.gov/nheerl/arm/analysispages/techinfoanalysis.htm
  - Put zip file on desktop or in folder
- Install library in R using menu

- Use latest version (2.6 currently on web)
- Key Files: Users Guide, Changes, psurvey.analysis_2.6.zip



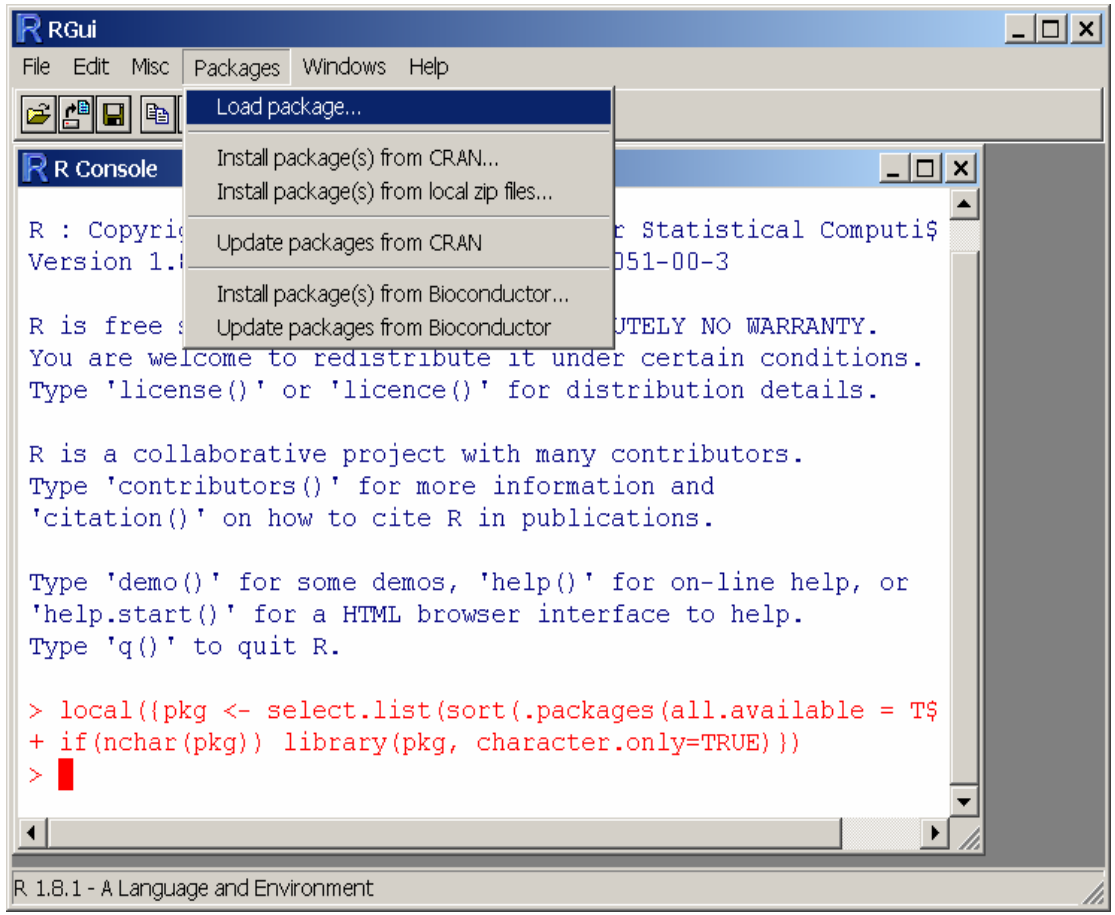
Install from local zip file

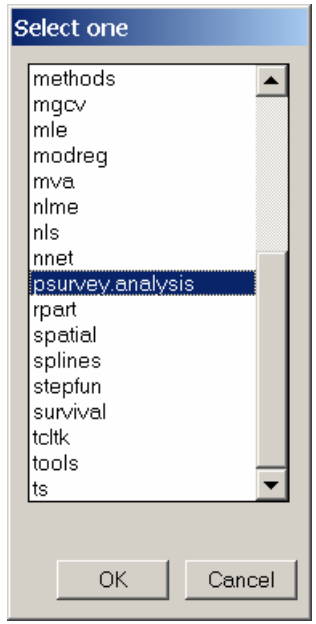


Begin Statistical Analysis for a Project using R

- Create a new folder specific for the statistical analysis
- Recommend create a sub folder named “Original Data”
  - Place any original data files in this folder
  - Never change these files
- Double click R desktop icon to start R
- Under R File menu,
  - go to Change Dir
  - Browse to find your project directory
- When done using R, exit and respond YES when asked if want to save workspace.
- When want to start R again to work on this project, go to project folder and open “.RData” file. This will automatically set R to use your project directory

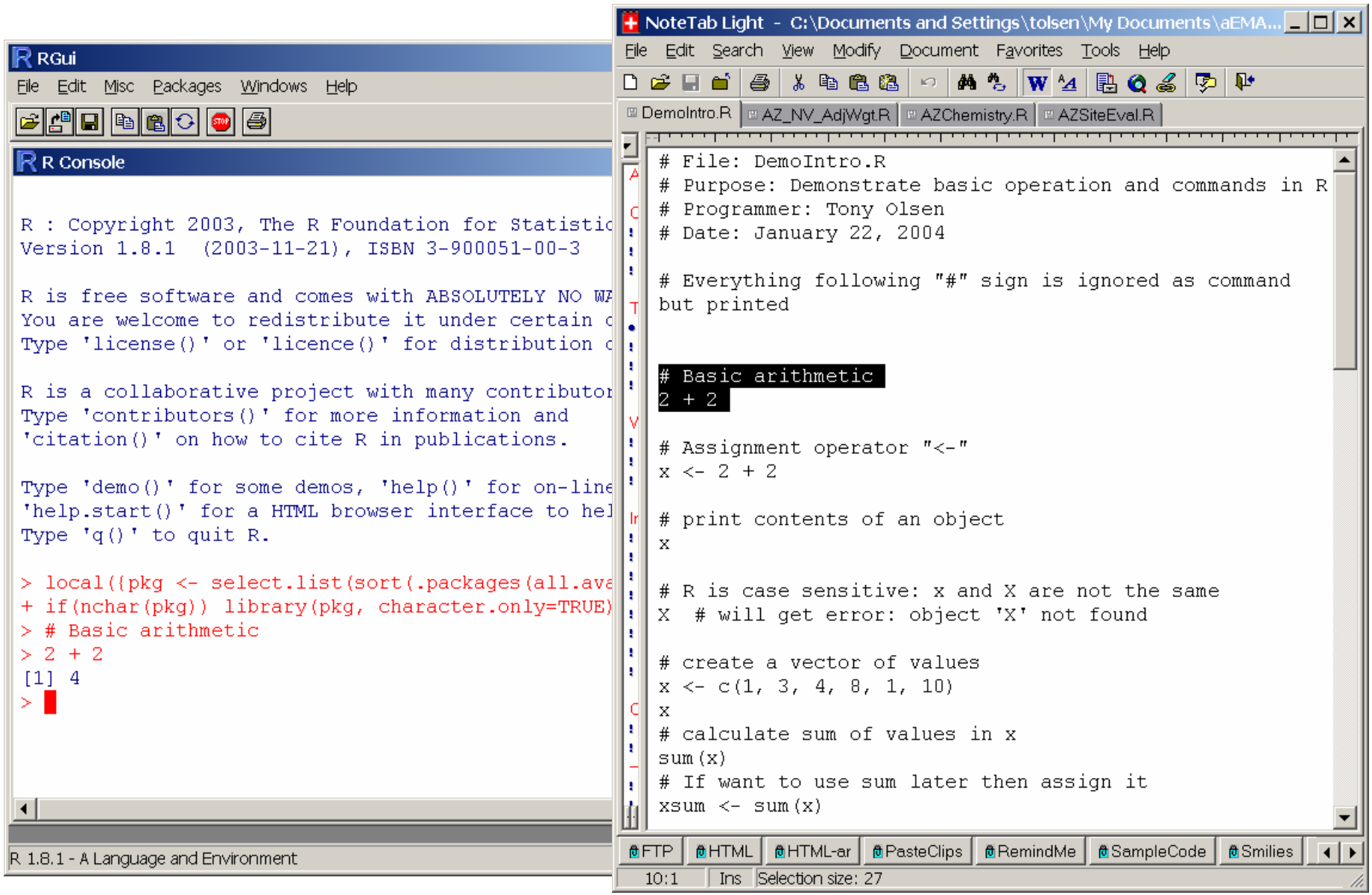
Load psurvey.analysis everytime start R





Executing Commands in R

- Three options to execute commands in R
  - Type commands in console window
    - window automatically opens when start R
    - Commands execute when hit “Enter” key
  - Use script window in R
    - Under File menu either create New or Open existing script file
    - Type commands in script window
    - To execute commands, copy and paste into console window
    - Close window and save to project directory
    - Use saved script file when want to continue work on same project
  - Use another text editor program
    - Type commands in editor
    - Copy and Paste into R console window
- Advantages of R script window or text editor approach
  - Can save all your commands in both cases
  - Some text editors recognize R language and structure making it easy to locate errors (e.g. UltraEdit)
  - Redo the analysis if change data with little extra work
  - Can set up an example analysis and use as template for other analyses
  - Aid in QA for analysis process


RGui
File Edit Misc Packages Windows Help
R Console
R : Copyright 2003, The R Foundation for Statistic
Version 1.8.1 (2003-11-21), ISBN 3-900051-00-3
R is free software and comes with ABSOLUTELY NO WA
You are welcome to redistribute it under certain c
Type 'license()' or 'licence()' for distribution c
R is a collaborative project with many contributor
Type 'contributors()' for more information and
'citation()' on how to cite R in publications.
Type 'demo()' for some demos, 'help()' for on-line
'help.start()' for a HTML browser interface to hel
Type 'q()' to quit R.
> local({pkg <- select.list(sort(.packages(all.ava
+ if(nchar(pkg)) library(pkg, character.only=TRUE)
> # Basic arithmetic
> 2 + 2
[1] 4
>
R 1.8.1 - A Language and Environment
NoteTab Light - C:\Documents and Settings\tolsen\My Documents\aEMA...
File Edit Search View Modify Document Favorites Tools Help
DemoIntro.R AZ_NV_AdjWgt.R AZChemistry.R AZSiteEval.R
File: DemoIntro.R
Purpose: Demonstrate basic operation and commands in R
Programmer: Tony Olsen
Date: January 22, 2004
Everything following "#" sign is ignored as command
but printed
Basic arithmetic
2 + 2
# Assignment operator "<-"
x <- 2 + 2
# print contents of an object
x
R is case sensitive: x and X are not the same
X # will get error: object 'X' not found
# create a vector of values
x <- c(1, 3, 4, 8, 1, 10)
x
# calculate sum of values in x
sum(x)
# If want to use sum later then assign it
xsum <- sum(x)
FTP HTML HTML-ar PasteClips RemindMe SampleCode Smilies
10:1 Ins Selection size: 27

Things to be aware of

- Column names may be modified in R
  - "blanks" will become "."
  - Unusual characters will become "."
  - Unusual character at beginning (e.g. %), will change to "X."
- Blanks in spreadsheet will be turned into missing data
- "NA" in spreadsheet will be interpreted as missing data, except when column is character
- Everything after "#" is ignored. Can be overridden.
- Reading a "csv" file may result in unanticipated results if columns contain "," in character fields (use "tab" delimited instead)