Advanced Data Analysis

Spring 2012

STAT W4201 ADVANCED DATA ANALYSIS

Instructor: E-mail: alem@stat.columbia.edu

Office Hours: Friday: 5:00 PM - 6 PM,

and by appointment

TAs:

- Chien Hsun Huang. (ch2526@columbia.edu)
- Yunxiao Chen (yc2710@columbia.edu)
  TA Office Hours: Wed and Thursday 5 PM – 7 PM

Prerequisites:

- At least two of the following courses are prerequisites: W4315/4200, W4220/W4325, W4437, W4413, W4543, W4290, W4240 and W4330

Course Objectives: Emphasis will be on hands-on experience with data analysis, involving case studies and using common statistical packages.

Method of Evaluation:

- Homework 30%.
  - Assigned weekly, on Fridays after class and is due the following Friday. Homework should be turned in at designated boxes in Room 904, before 5 PM each Friday.
- Midterm/Test 30%.
  - Date TBD
- Project 40%.
  - Due on Monday April 30, 2012; 5 PM.

Homework must be placed in designated boxes in Room 904 each Friday, before class.

Suggested Reference Books:

Because of the nature of the course, no single text book is required.

However, the following text is suggested:

- The Statistical Sleuth: A Course in Methods of Data Analysis. Ramsey & Schafer 1st Duxbury 534386709
- RECOMMENDED:
  - Introduction Time Series and Forecasting Brockwell & Davis 2nd Springer-Verlag 0387953515
  - Generalized Linear Models Peter Mccullagh, John A. Nelder CRC Press 412317605
  - An Introduction to S and S-Plus Spector 1st Duxbury 053419866X

The following are also useful references

1. Miller, R. *Survival Analysis*. 1981, Wiley
2. McCullagh and Nedler. *Generalized Linear Models.* Chapman/Hall.
3. Selvin. *Practical Biostatistical Methods*, Duxbury
4. Hoaglin, et al. *Fundamentals of Explor. Analysis of Variance*, Wiley
5. Hosmer and Lemshow. *Applied Logistic Regression*, Wiley
6. Neter, Wasserman and Kutner. *Applied Linear Statistical Models*. Wiley.
7. Chambers and Hastie. *Statistical Models in S*. Wadsworth.

GSAS Statement on Academic Honesty

Students should be aware that academic dishonesty (for example, plagiarism, cheating on an examination, or dishonesty in dealing with a faculty member or other University official) or the threat of violence or harassment are particularly serious offenses and will be dealt with severely under Dean's Discipline. Graduate students are expected to exhibit the high level of personal and academic integrity and honesty required of all members of an academic community as they engage in scholarly discourse and research.

Scholars draw inspiration from the work done by other scholars; they argue their claims with reference to others' work; they extract evidence from the world or from earlier scholarly works. When a student engages in these activities, it is vital to credit properly the source of his or her claims or evidence. To fail to do so would violate one's scholarly responsibility.

In practical terms, students must not cheat on examinations, and deliberate plagiarism is of course prohibited. Plagiarism includes buying, stealing, borrowing, or otherwise obtaining all or part of a paper (including obtaining or posting a paper online); hiring someone to write a paper; copying from or paraphrasing another source without proper citation or falsification of citations; and building on the ideas of another without citation. Students also should not submit the same paper to more than one class.

Graduate students are responsible for proper citation and paraphrasing, and must also take special care to avoid even accidental plagiarism. The best strategy is to use great caution in the handling of ideas and prose passages: take notes carefully and clearly mark words and ideas not one's own. Failure to observe these rules of conduct will result in serious academic consequences, which can include dismissal from the university.

Students engaging in research must be aware of and follow university policies regarding intellectual and financial conflicts of interest, integrity and security in data collection and management, intellectual property rights and data ownership, and necessary institutional approval for research with human subjects and animals.

http://www.columbia.edu/cu/gsas/rules/chapter-9/pages/honesty/index.html

Ground Rules

- Attendance mandatory
  - surprise quizzes may be given to encourage attendance
  - No cellphones, web surfing or chatting during lecture
- No late homework or make-up test or project
- All homework, tests and project reports should reflect individual effort
- Active participation in class expected
  - Questions and answers
  - Project presentations

Topics:

- Exploratory data analysis
- Model formulation, goodness of fit testing
- Standard and non-standard statistical procedures, including:
  - Linear regression,
  - Analysis of variance
  - Nonlinear regression
  - Generalized linear models
  - Survival analysis
  - Time series analysis
  - Bayesian methods
- Each student will be asked to propose a data set, with the approval of the instructor, define a research problem, develop an analysis plan, and submit a report at the end of the semester. Only original (and NO group) projects will be accepted

Generating and Summarizing Data

Experiments: Performed to generate data to help make decisions.

1. *Clinical Research*: Is a new therapy superior to the standard?
2. *Genetics*: Is there any association between genetic make-up and occurrence of a certain type of disease?
3. *Agriculture*: What is the effect of soil types on crop yield?
4. *Finance*: What factors affect the performance of a company's stock?
5. *Weather*: What is the forecast in the next quarter?

Generating and Summarizing Data (cont’d)

6. *Political Science*: How do the polls predict election results?
7. *Game Theory*: Why does the casino make a profit at a roulette?
8. *Manufacturing*: What is the reliability of a certain manufacturing process?
9. *Demography*: What is the growth rate of a population in a given region?

Types of Studies

- Controlled
- Observational studies

Types of Studies

- Randomized, controlled, double-blind
  - Randomization guards against selection bias
  - Ensures that groups are comparable.
  - Double-blind: Minimizes bias, either in the response or in the evaluation of the experimental outcomes.
- Observational studies:
  - Assignment of experimental subjects to study groups not done by the investigator.
  - May lack advantages of controlled trials
  - May help establish association when RCTs not feasible

Questions to Ask in Data Analysis

- What is the objective of the analysis and/or the original experiment?
- What was the design of the study?
  - Randomized controlled or observational?
  - If a controlled trial, how were subjects assigned to the different groups?
  - Was the assignment process controlled by the investigator?
  - If an observational study: Are the groups comparable? What factors are confounded with treatment?
- What procedure would be appropriate for the data?
  - Exploratory data analysis techniques?
  - Inferential statistical techniques?
  - Model building?
- Implementation of analysis plan?
- Interpretation of Results?
  - Are the results relevant?

Analysis Plan

- Was the planned analysis followed?
  - If a large number of analyses are performed, some of them will be sure to show structure.

``If you torture the data long enough, they'll admit to anything"

- Were assumptions validated?
- Were there any confounding factors. If so, were appropriate measures taken?
- Were multiple procedures/subgroup analyses performed?
  - If so, what adjustments were made for multiplicity?

Exploratory Data Analysis (EDA)

Preliminary look at data:

- Evaluating data quality
  - Missing values
  - Outliers/Influential points
- Checking assumptions: Distributions, relationships, etc.
- Measures of location & dispersion

EDA cont.

Approaches:

- Descriptive Statistics
  - Measures of location and dispersion
- Graphical
  - Histograms, box-plots, Q-Q plots, etc.

EDA cont.

Measures of Location

- Properties of the sample mean
  - It is easy to compute
  - It is easy to interpret/understand
  - Its variance has a simple expression
  - It is susceptible to outliers

- Properties of the sample median:
  - Relatively more complex to compute and understand,
  - Strongly resistant to outliers.
  - Variability does not have a simple expression.

EDA cont.

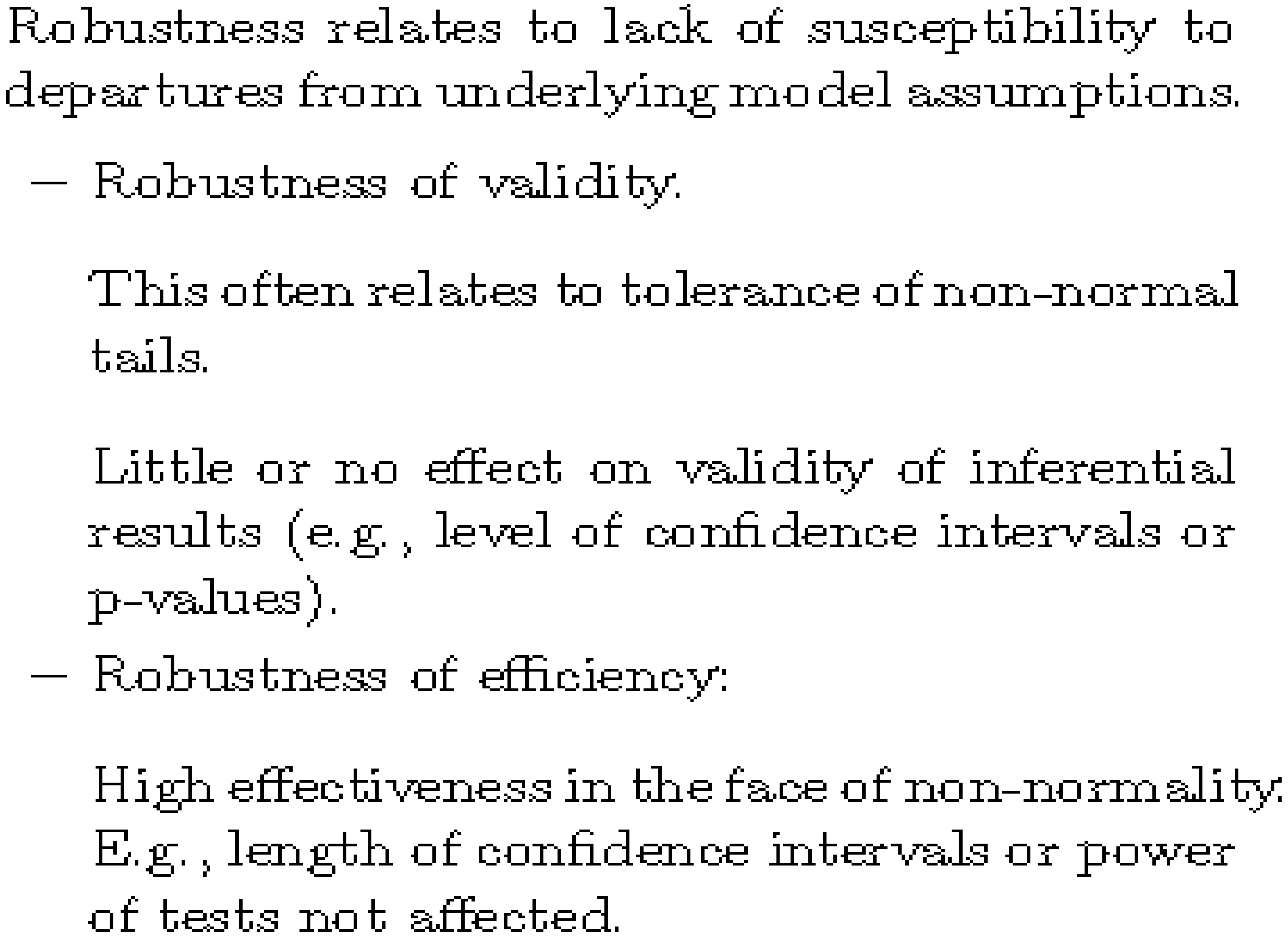
Measures of Dispersion

- Properties of the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Properties of the IQR, Range
  - Sampling distribution?

Robustness



Robustness relates to lack of susceptibility to departures from underlying model assumptions.

- Robustness of validity.

  This often relates to tolerance of non-normal tails.

  Little or no effect on validity of inferential results (e.g., level of confidence intervals or p-values).

- Robustness of efficiency:

  High effectiveness in the face of non-normality. E.g., length of confidence intervals or power of tests not affected.

Robustness (cont’d)

- Breakdown Point
  - Fraction of data that could be made arbitrarily large, without making the estimator useless
  - Breakdown point for sample mean vs. median

Some examples of robust procedures

– *Sample trimmed mean*

$$\bar{X}_t = \frac{X_{(g+1)} + \cdots + X_{(n-g)}}{n-2g}$$

where $g = [\gamma n]$, $0 \leq \gamma \leq 0.5$. Usually, $\gamma = 0.2$ is used.

$$Var(\bar{X}_t) \approx \frac{1}{n^2(1-2\gamma)^2} \Sigma(W_I - \bar{W})^2$$

where

$$W_i = \begin{cases} X_{(g+1)}, & X_i \leq X_{(g+1)} \\ X_i, & X_{(g+1)} < X_i < X_{(n-g)} \\ X_{(n-g)}, & otherwise \end{cases}$$

$\bar{W} = \frac{\Sigma W_i}{n}$ is called the winsorized sample mean.

– *M Estimates*

M estimates are obtained as minimizers of the quantity

$$\sum_{j=1}^{n} \rho\left(\frac{X_i - \mu}{\sigma}\right)$$

A common choice for the weight function ρ is the Huber weight functions,

$$\rho(u) = \begin{cases} \frac{u^2}{2} & \mid u \mid \leq k \\ k \mid u \mid & \mid u \mid > k \end{cases}$$

which are quadratic near zero, and linear beyond a prespecified cutoff point k. When k=∞ we get the sample mean, while a value of k=0 gives the median.

- *Median absolute deviation*

$$median\{\mid X_j - \tilde{X} \mid, j = 1, \cdots, n\}$$

The Jackknife Method

Let θ be an unknown population parameter of interest.

Let $\hat{\theta}$ be a statistic or estimator of θ.

Then the bias of $\hat{\theta}$ is given by $E[\hat{\theta}] - \theta$.

Generally, the bias and variance of an estimator may not be readily computable.

A method, due to Quenouille (1956), that may be used to compute bias and variance is the *jackknife* procedure.

Let $\hat{\theta}_{(j)}$ be an estimator computed based on all but X_j, i.e., leaving out the j'th observation.

Then the jackknife estimator of bias is given by

$$B_{JACK} = (n-1)\left[\frac{\Sigma_j^n \hat{\theta}_{(j)}}{n} - \hat{\theta}\right]$$

The bias reduced jackknife estimator is given by

$$\hat{\theta}_{JACK} = \hat{\theta} - B_{JACK}$$

and the variance

$$V_{JACK} = \frac{n-1}{n}\sum_j(\hat{\theta}_{(j)} - \frac{\Sigma_j \hat{\theta}_{(j)}}{n})^2$$

Caution in the use of the jackknife:

- Jackknife may not be appropriate in the presence of outliers or for markedly skewed distributions
- The jackknife may not be appropriate when θ has restricted values, e.g., $\theta \epsilon [0, 1]$.

Computer-Intensive Statistical Methods

- Monte Carlo Methods
- Bootstrap Methods
- Randomization Tests

Motivations

- Classical statistics mostly based on idealized assumptions
- Advance in computation helps replace complex theoretical analysis by computationally intensive methods

Monte Carlo Methods

- Requirement:
  - Knowledge of the distribution to easily generate new samples

Example:

- Simulate the sampling distribution of the interquartile range (IQR) of scores
  - Sampling distribution unknown

Procedure Monte Carlo Sampling: IQR

1. Let F be Normal($\mu = 50$, $\sigma = 5$).

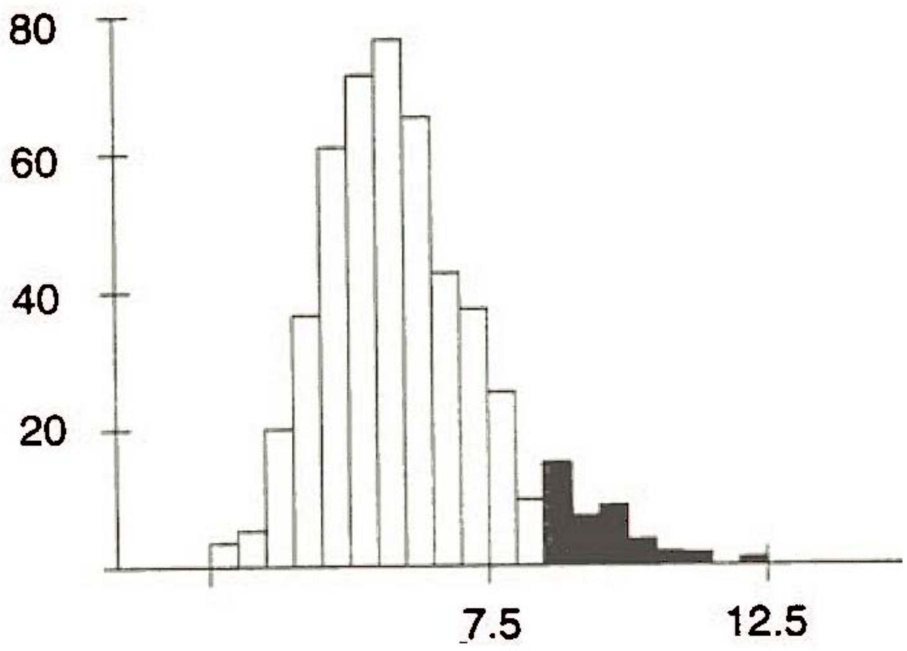
Let $\mathfrak{S}$ be the original sample of size $N = 20$ students, and let IQR = IQR($\mathfrak{S}$) = 8.1 be our sample statistic.

2. Repeat $i = 1 .. K$ times:

   a. Draw a pseudosample $\mathfrak{S}_i^*$ of size n from F by random sampling.

   b. Calculate and record $\text{IQR}_i^* = \text{IQR}(\mathfrak{S}_i^*)$.

3. The distribution of IQR* is an empirical sampling distribution .



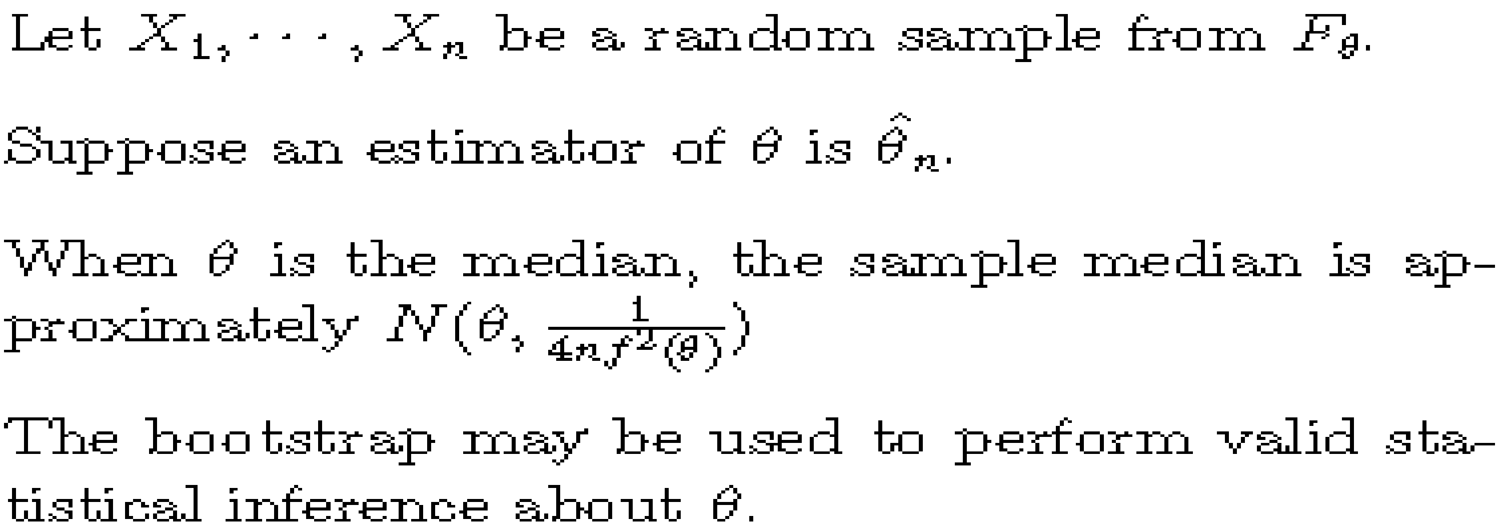
- MC sampling distribution of IQRs for *K* = 500.
- Dark: 40 scores greater than 8.1

Bootstrap Methods

Idea:

- Now suppose F is arbitrary, unknown distribution.
- Resample from the sample, treating the sample as the population.
  - The sample should be representative of the population.
- Approximate sampling distribution of statistics based on corresponding pseudo-sample quantities.

Bootstrap Methods (cont'd)

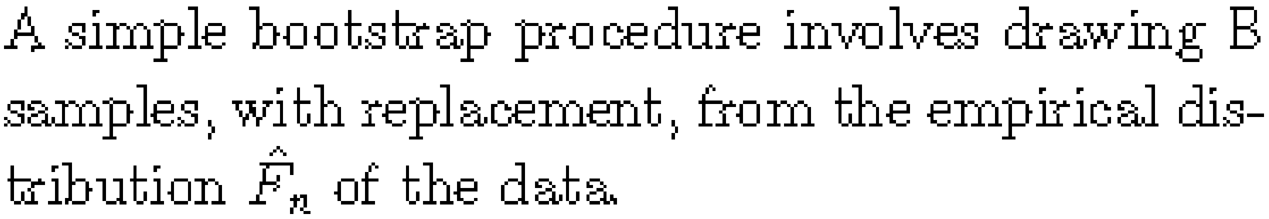


Let $X_1, \cdots, X_n$ be a random sample from F_θ.

Suppose an estimator of θ is $\hat{\theta}_n$.

When θ is the median, the sample median is approximately $N(\theta, \frac{1}{4nf^2(\theta)})$

The bootstrap may be used to perform valid statistical inference about θ.



A simple bootstrap procedure involves drawing B samples, with replacement, from the empirical distribution $\hat{F}_n$ of the data.

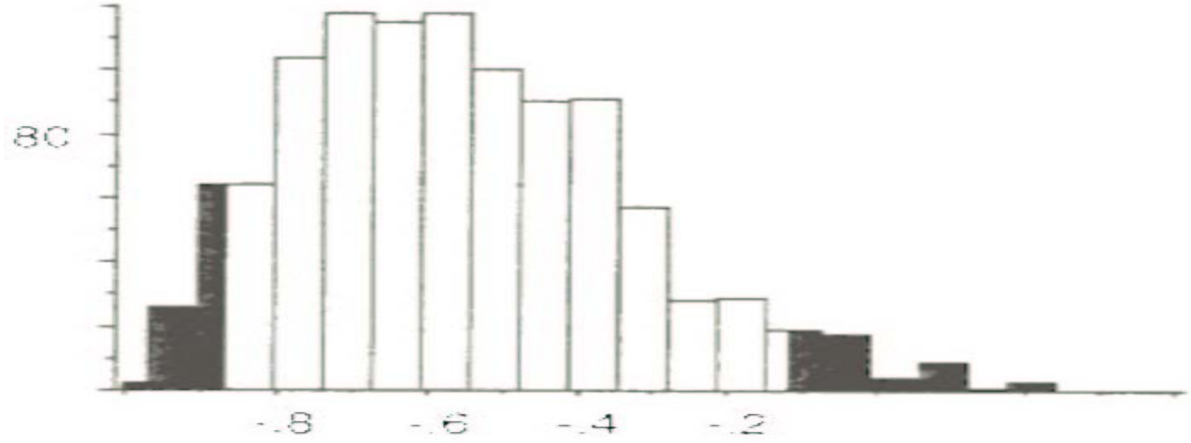
Bootstrap Methods (cont'd)

For each sample, compute a statistic of interest, say $\hat{\theta}_n^+$.

- Assess the variability of $\hat{\theta}_n$ about θ by that of $\hat{\theta}_n^+$ about $\hat{\theta}_n$
- Estimate the bias $\hat{\theta}_n - \theta$ by the mean of $\hat{\theta}^+ - \hat{\theta}$
- Estimate the distribution of $\hat{\theta}$ by the e.d.f. of $\hat{\theta}^+$.

Bootstrapping the sample correlation coefficient

x	5	1.75	0.8	5	1.75	5	1.75	1	5	1.75
y	27.8	20.82	44.12	29.41	31.19	28.68	29.53	34.62	20	41.54



Bootstrap sampling distribution of r (*K* = 1000).

Remarks

- In practice, the bootstrap works well in many situations, but not in all
- Assumes: The original sample is representative of the population.

Randomization Tests

- The bootstrap treats samples as “proxies” for populations.
- Sometimes may wish to determine whether two samples are related without any reference to population parameters.

Randomization Tests (cont'd)

Example: Is the performance of this year's students significantly more variable (IQR) than the performance of last year's students?

Sample 1: last year				
48.35	53.93	55.48	45.67	52.82
49.47	57.00	53.61	57.69	51.34
44.98	54.70	59.32	51.70	50.73
46.84	63.13	52.50	49.67	54.07
44.84	48.68	53.94	59.00	50.92

Sample 2: this year				
64.82	51.69	57.00	58.17	40.63
50.90	48.77	40.33	50.76	49.64
56.25	65.68	57.50	47.45	46.78
61.34	53.66	49.10	54.49	54.15

- Let d_{IQR} denote the difference between the IQRs of the samples, $d_{IQR} = 6 - 8.5 = -2.5$.
  - What is the probability that this difference occurs by chance?

- Claim: students’ performance is no more variable this year than last.
- If claim is really true, then randomly swapping scores between the samples will not influence d_{IQR} .

Approximate Randomization to Test Whether Two Samples Are Drawn from the Same Population

1. Let S_A and S_B be two samples of sizes n_A and n_B, respectively. Let $\theta = f(S_A, S_B)$ a statistic calculated from the two samples. Let S_{A+B} be the merge of S_A and S_B.

2. Do $i = 1 \ldots K$ times:

   a. Shuffle the elements of S_{A+B} thoroughly.

   b. Assign the first n_A elements of S_{A+B} to a randomized pseudosample A_i* and the remaining n_B elements to B_i*.

   c. Calculate $\theta_\mathrm{i}* = f(A_i*, B_i*)$ and record the result.

3. The distribution of θ_i* can now be used to find the probability of the sample result θ under the hypothesis that the samples are drawn from the same population.

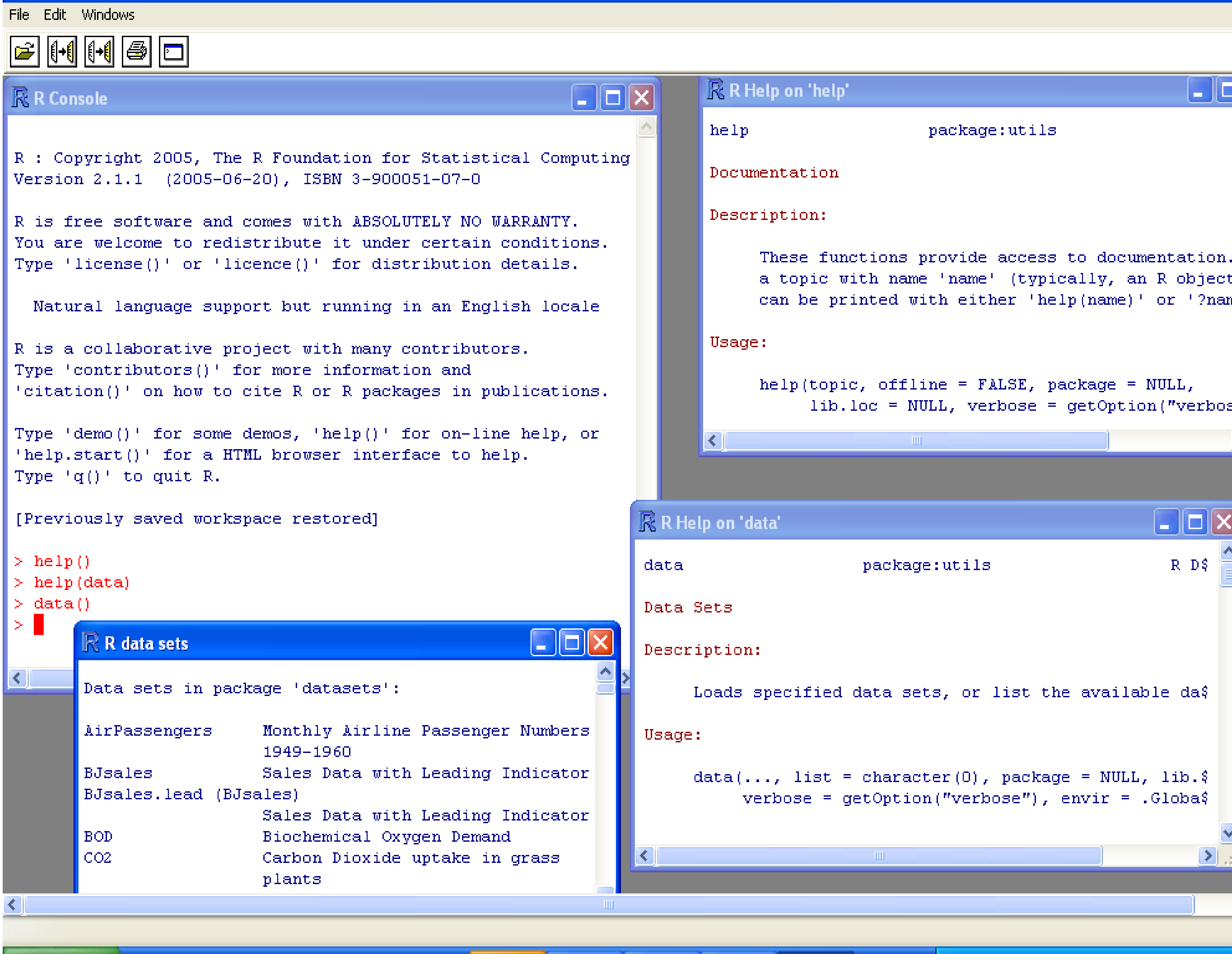
- A*pproximate randomization:* *K* iterations do not exhaust the space of all possible assignments of elements of S_{A+B} to A_i* and B_i*.
- *Exact randomization:* If can find exact probability by generating all possible outcomes.

Comparing Bootstrap and Randomization Procedures

- Both generate the distribution of a statistic by resampling from the original sample.
  - Bootstrap resamples with replacement.
  - Randomization resamples without replacement.
- Bootstrap simulates the process of drawing samples from a population, while randomization does not.
- They produce different distributions!
- Randomization cannot be used to draw inferences about population parameters (e.g., confidence intervals).

Computer-Intensive vs. Parametric Procedures

- Computer-intensive methods most desirable when:
  - No parametric sampling distribution exists for a statistic.
  - Assumptions of underlying a parametric test are violated and procedure not robust.

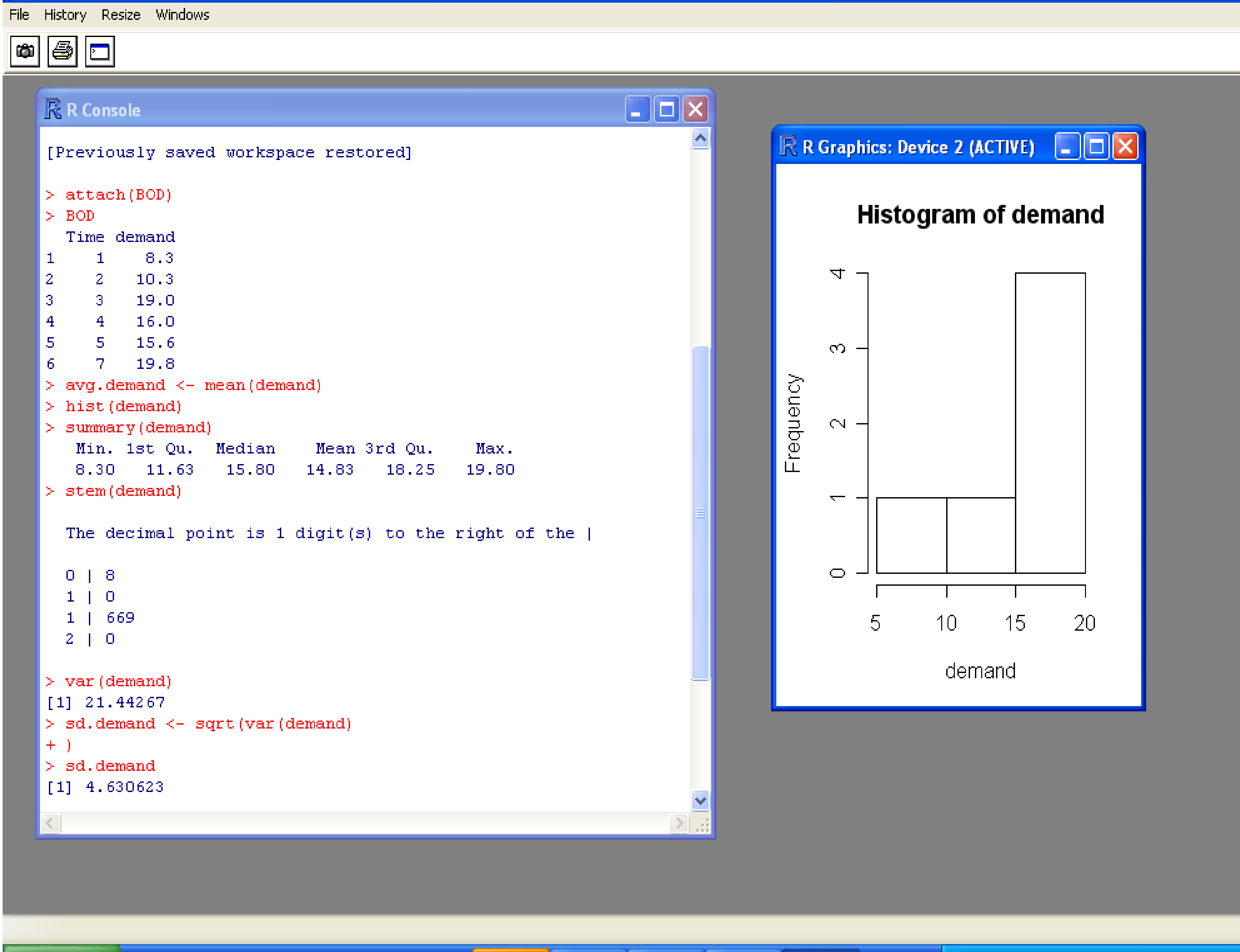


```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.1  (2005-06-20), ISBN 3-900051-07-0

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.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

[Previously saved workspace restored]

> help()
> help(data)
> data()
>
```

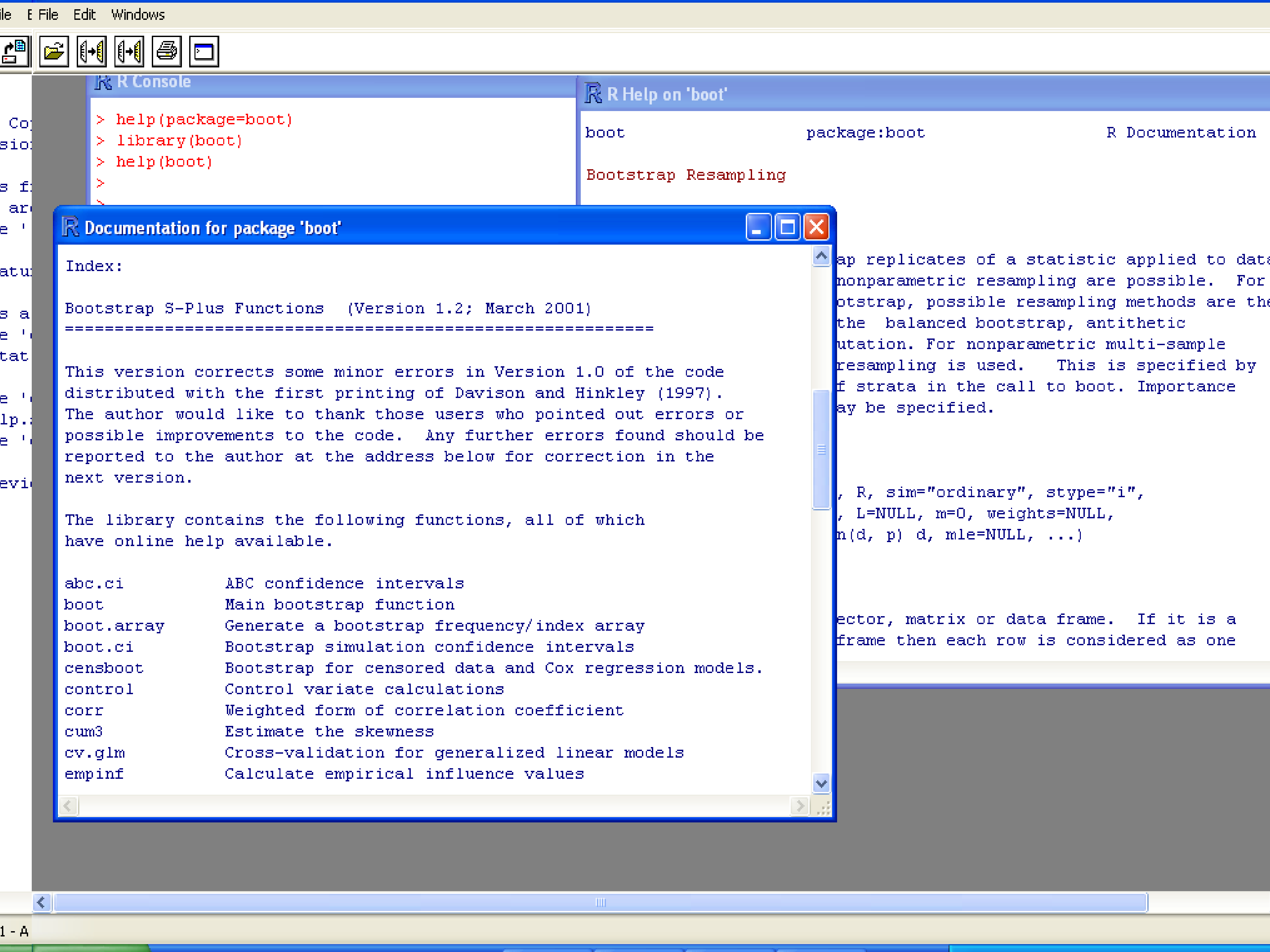
```
help                    package:utils

Documentation

Description:

     These functions provide access to documentation.
     a topic with name 'name' (typically, an R object
     can be printed with either 'help(name)' or '?nam

Usage:

     help(topic, offline = FALSE, package = NULL,
          lib.loc = NULL, verbose = getOption("verbos
```

```
data                    package:utils                    R D$

Data Sets

Description:

     Loads specified data sets, or list the available da$

Usage:

     data(..., list = character(0), package = NULL, lib.$
          verbose = getOption("verbose"), envir = .Globa$
```

```
Data sets in package 'datasets':

AirPassengers       Monthly Airline Passenger Numbers
                    1949-1960
BJsales             Sales Data with Leading Indicator
BJsales.lead (BJsales)
                    Sales Data with Leading Indicator
BOD                 Biochemical Oxygen Demand
CO2                 Carbon Dioxide uptake in grass
                    plants
```



```
[Previously saved workspace restored]

> attach(BOD)
> BOD
  Time demand
1    1    8.3
2    2   10.3
3    3   19.0
4    4   16.0
5    5   15.6
6    7   19.8
> avg.demand <- mean(demand)
> hist(demand)
> summary(demand)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.30   11.63   15.80   14.83   18.25   19.80
> stem(demand)

  The decimal point is 1 digit(s) to the right of the |

  0 | 8
  1 | 0
  1 | 669
  2 | 0

> var(demand)
[1] 21.44267
> sd.demand <- sqrt(var(demand)
+ )
> sd.demand
[1] 4.630623
```

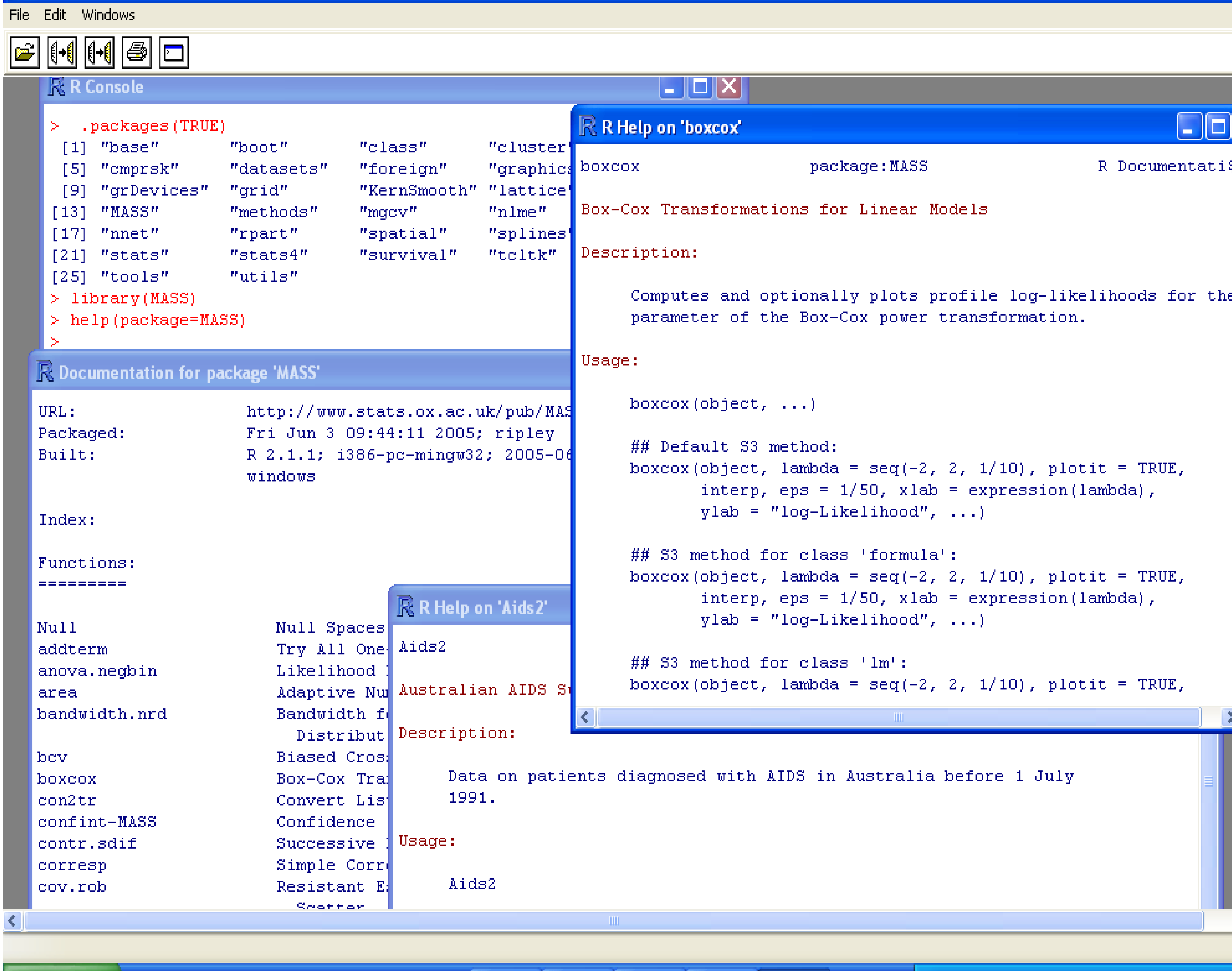


```
> help(package=boot)
> library(boot)
> help(boot)
>
```

```
boot                package:boot                R Documentation

Bootstrap Resampling
```

```
ap replicates of a statistic applied to data
nonparametric resampling are possible.  For
otstrap, possible resampling methods are the
the  balanced bootstrap, antithetic
utation. For nonparametric multi-sample
resampling is used.   This is specified by
f strata in the call to boot. Importance
ay be specified.

, R, sim="ordinary", stype="i",
, L=NULL, m=0, weights=NULL,
n(d, p) d, mle=NULL, ...)

ector, matrix or data frame.  If it is a
frame then each row is considered as one
```

```
Index:

Bootstrap S-Plus Functions  (Version 1.2; March 2001)
============================================================

This version corrects some minor errors in Version 1.0 of the code
distributed with the first printing of Davison and Hinkley (1997).
The author would like to thank those users who pointed out errors or
possible improvements to the code.  Any further errors found should be
reported to the author at the address below for correction in the
next version.

The library contains the following functions, all of which
have online help available.

abc.ci          ABC confidence intervals
boot            Main bootstrap function
boot.array      Generate a bootstrap frequency/index array
boot.ci         Bootstrap simulation confidence intervals
censboot        Bootstrap for censored data and Cox regression models.
control         Control variate calculations
corr            Weighted form of correlation coefficient
cum3            Estimate the skewness
cv.glm          Cross-validation for generalized linear models
empinf          Calculate empirical influence values
```

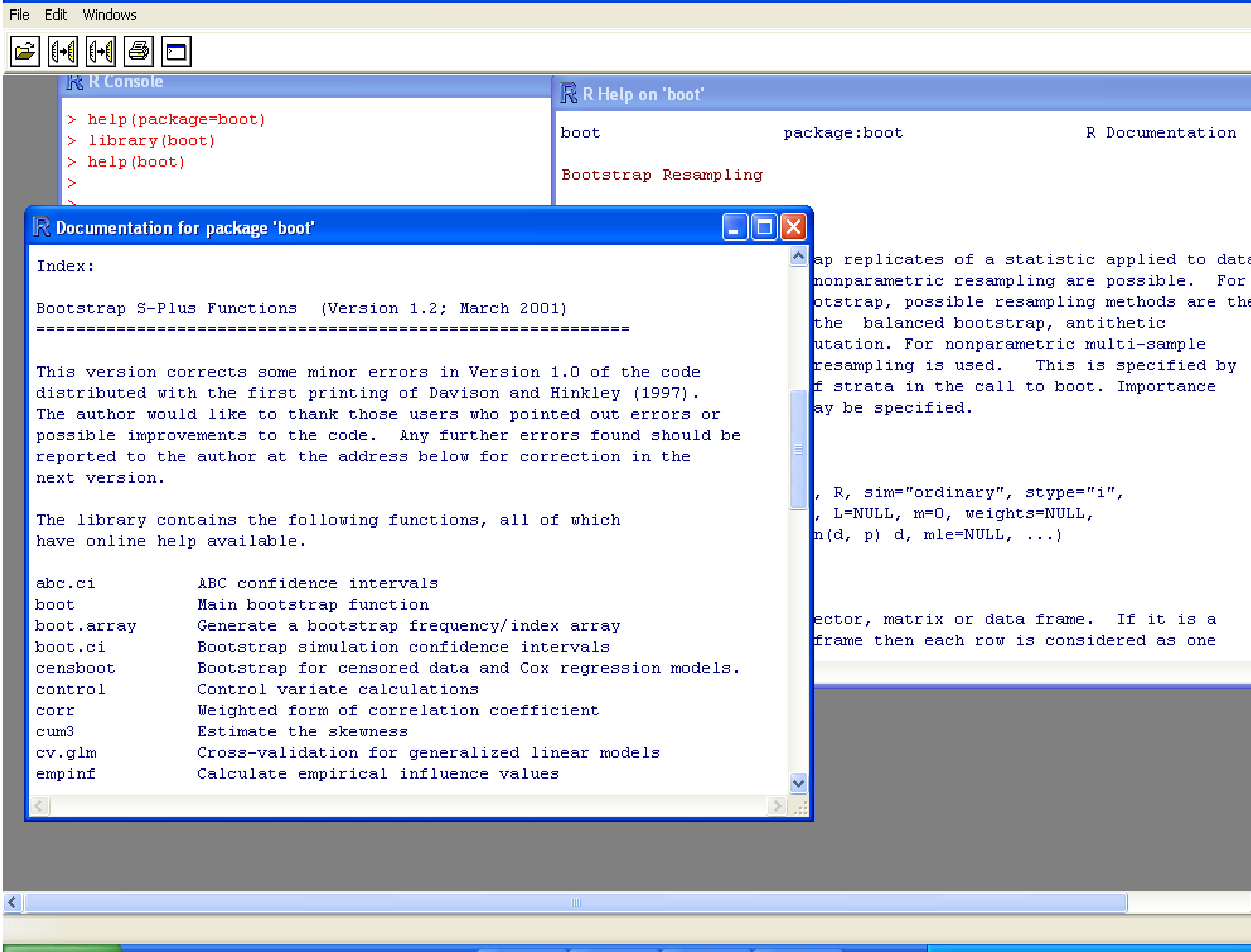


```
> .packages(TRUE)
 [1] "base"       "boot"       "class"      "cluster"
 [5] "cmprsk"     "datasets"   "foreign"    "graphics
 [9] "grDevices"  "grid"       "KernSmooth" "lattice"
[13] "MASS"       "methods"    "mgcv"       "nlme"
[17] "nnet"       "rpart"      "spatial"    "splines"
[21] "stats"      "stats4"     "survival"   "tcltk"
[25] "tools"      "utils"
> library(MASS)
> help(package=MASS)
>
```

```
URL:                 http://www.stats.ox.ac.uk/pub/MAS
Packaged:            Fri Jun 3 09:44:11 2005; ripley
Built:               R 2.1.1; i386-pc-mingw32; 2005-06
                     windows


Index:


Functions:
=========


Null                    Null Spaces
addterm                 Try All One-
anova.negbin            Likelihood R
area                    Adaptive Nu
bandwidth.nrd           Bandwidth fo
                          Distribut
bcv                     Biased Cros
boxcox                  Box-Cox Tra
con2tr                  Convert Lis
confint-MASS            Confidence
contr.sdif              Successive D
corresp                 Simple Corre
cov.rob                 Resistant E
                          Scatter
```

```
Aids2

Australian AIDS St

Description:

     Data on patients diagnosed with AIDS in Australia before 1 July
     1991.

Usage:

     Aids2
```

```
boxcox                 package:MASS                 R Documentati

Box-Cox Transformations for Linear Models

Description:

     Computes and optionally plots profile log-likelihoods for the
     parameter of the Box-Cox power transformation.

Usage:

     boxcox(object, ...)

     ## Default S3 method:
     boxcox(object, lambda = seq(-2, 2, 1/10), plotit = TRUE,
            interp, eps = 1/50, xlab = expression(lambda),
            ylab = "log-Likelihood", ...)

     ## S3 method for class 'formula':
     boxcox(object, lambda = seq(-2, 2, 1/10), plotit = TRUE,
            interp, eps = 1/50, xlab = expression(lambda),
            ylab = "log-Likelihood", ...)

     ## S3 method for class 'lm':
     boxcox(object, lambda = seq(-2, 2, 1/10), plotit = TRUE,
```

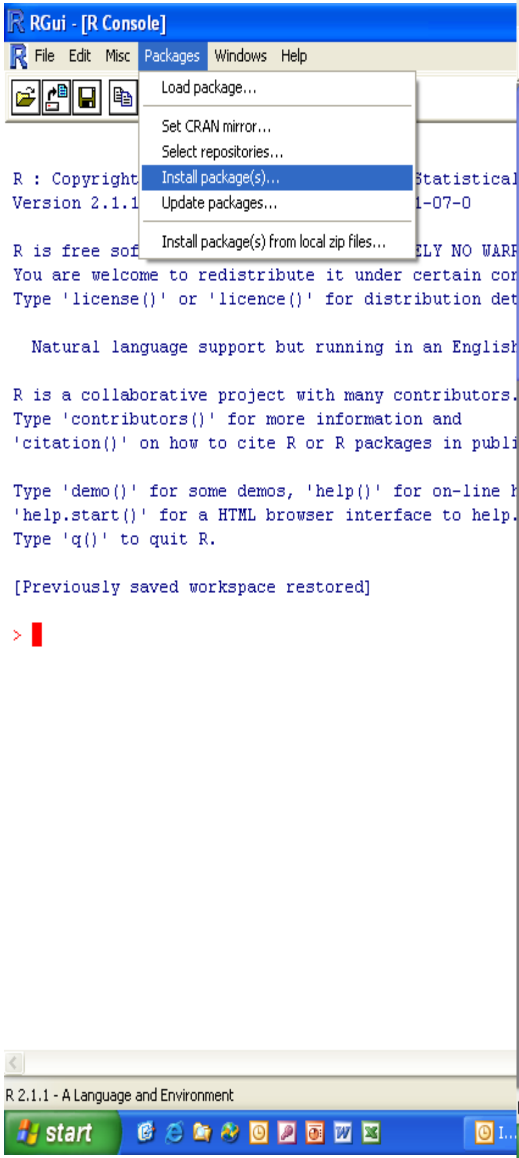


```
> help(package=boot)
> library(boot)
> help(boot)
>
```

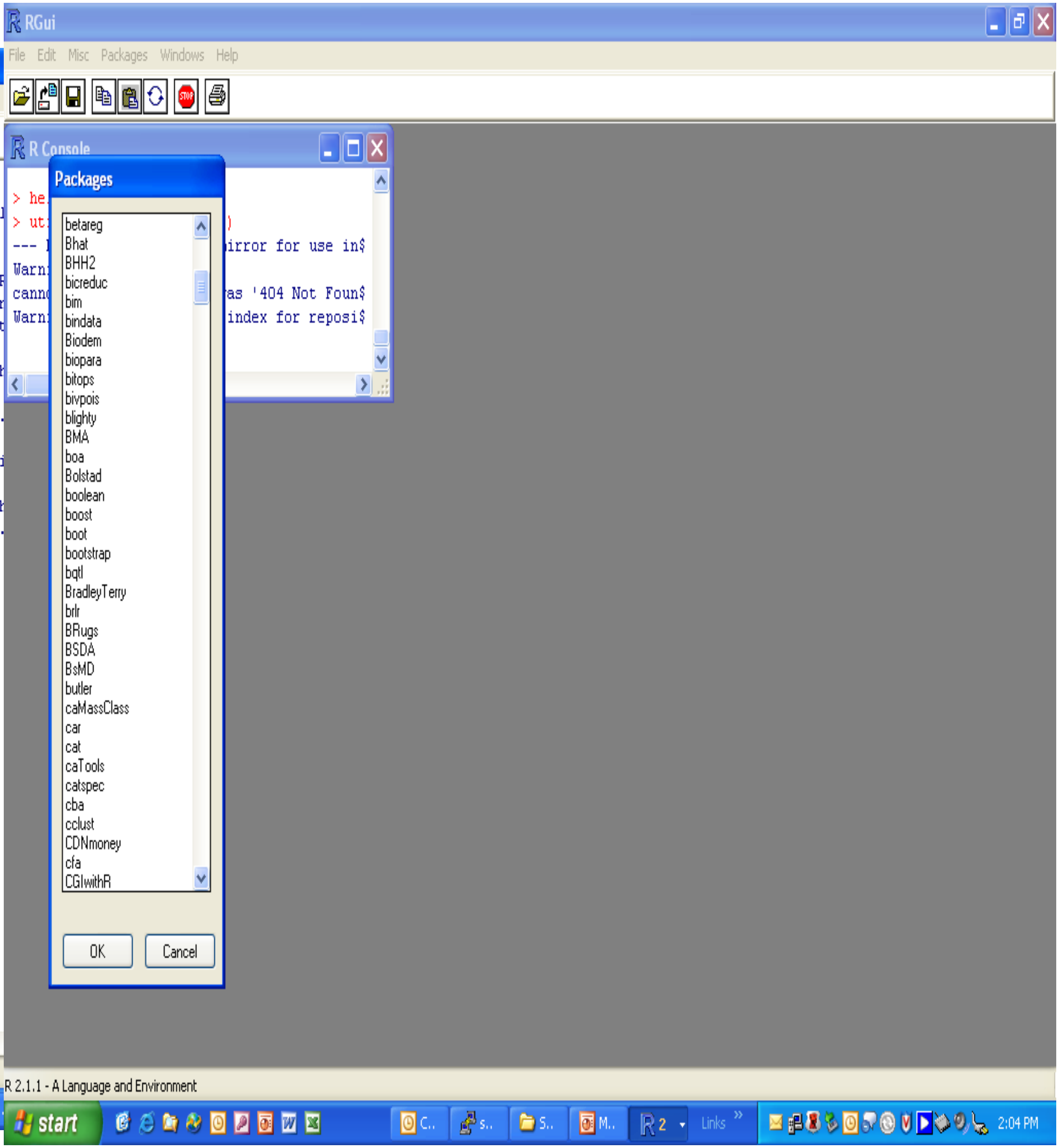
```
boot                  package:boot                  R Documentation

Bootstrap Resampling

ap replicates of a statistic applied to data
nonparametric resampling are possible.  For
otstrap, possible resampling methods are the
the  balanced bootstrap, antithetic
utation. For nonparametric multi-sample
resampling is used.   This is specified by
f strata in the call to boot. Importance
ay be specified.

, R, sim="ordinary", stype="i",
, L=NULL, m=0, weights=NULL,
n(d, p) d, mle=NULL, ...)

ector, matrix or data frame.  If it is a
frame then each row is considered as one
```

```
Index:

Bootstrap S-Plus Functions  (Version 1.2; March 2001)
===========================================================

This version corrects some minor errors in Version 1.0 of the code
distributed with the first printing of Davison and Hinkley (1997).
The author would like to thank those users who pointed out errors or
possible improvements to the code.  Any further errors found should be
reported to the author at the address below for correction in the
next version.

The library contains the following functions, all of which
have online help available.

abc.ci          ABC confidence intervals
boot            Main bootstrap function
boot.array      Generate a bootstrap frequency/index array
boot.ci         Bootstrap simulation confidence intervals
censboot        Bootstrap for censored data and Cox regression models.
control         Control variate calculations
corr            Weighted form of correlation coefficient
cum3            Estimate the skewness
cv.glm          Cross-validation for generalized linear models
empinf          Calculate empirical influence values
```

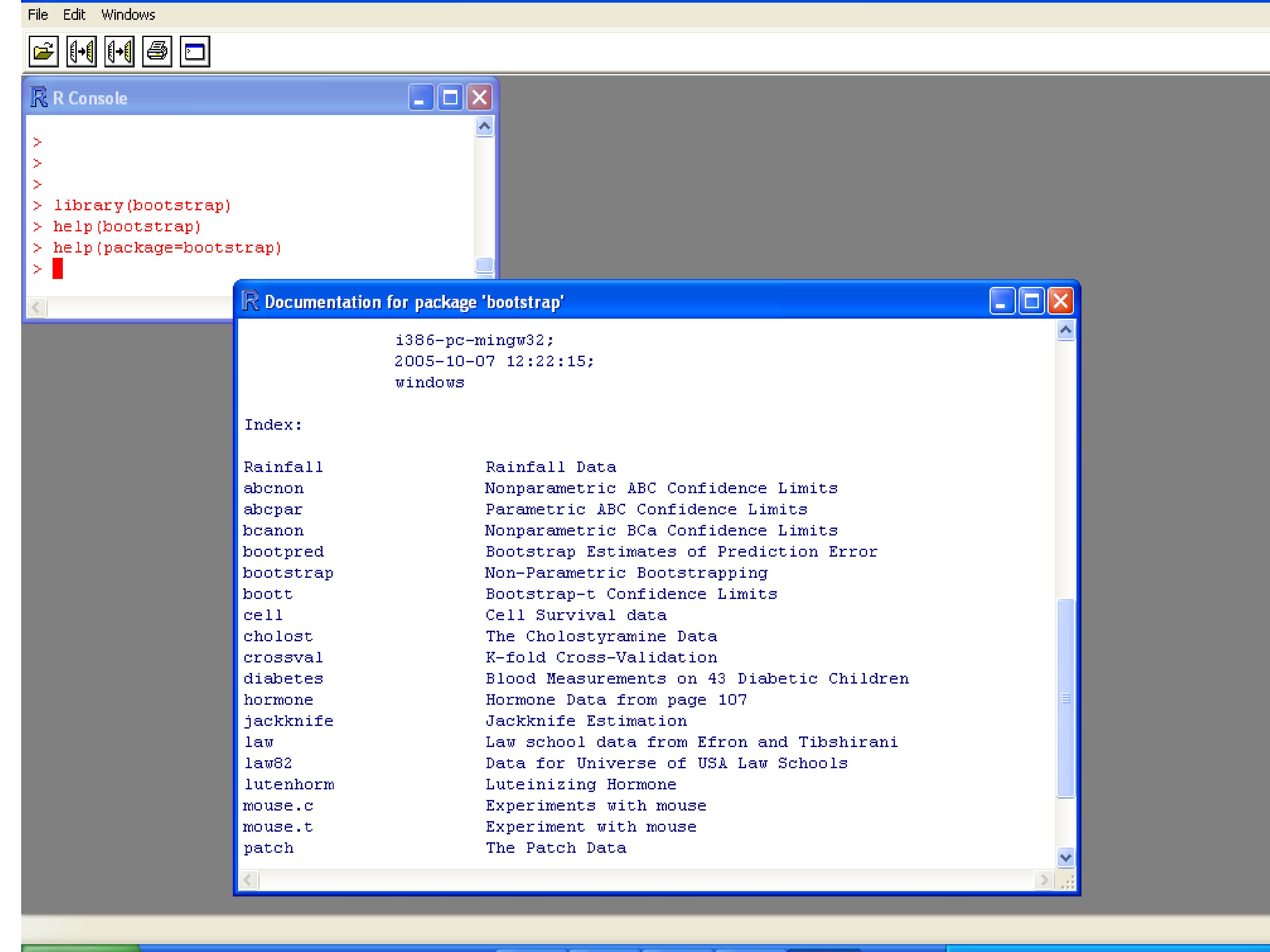
#Go to ***Packages***, ***Install Packages***, (select USA1, etc.) Bootstrap



```
R : Copyright                                   Statistical
Version 2.1.1                                   1-07-0

R is free sof                                   ELY NO WARR
You are welcome to redistribute it under certain con
Type 'license()' or 'licence()' for distribution det

  Natural language support but running in an English

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publi

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

[Previously saved workspace restored]

>
```



```
> he
> ut                              )
---                               irror for use in$
Warn
cann                              as '404 Not Foun$
Warn                              index for reposi$
```



```
>
>
>
> library(bootstrap)
> help(bootstrap)
> help(package=bootstrap)
>
```

```
                  i386-pc-mingw32;
                  2005-10-07 12:22:15;
                  windows

Index:

Rainfall                    Rainfall Data
abcnon                      Nonparametric ABC Confidence Limits
abcpar                      Parametric ABC Confidence Limits
bcanon                      Nonparametric BCa Confidence Limits
bootpred                    Bootstrap Estimates of Prediction Error
bootstrap                   Non-Parametric Bootstrapping
boott                       Bootstrap-t Confidence Limits
cell                        Cell Survival data
cholost                     The Cholostyramine Data
crossval                    K-fold Cross-Validation
diabetes                    Blood Measurements on 43 Diabetic Children
hormone                     Hormone Data from page 107
jackknife                   Jackknife Estimation
law                         Law school data from Efron and Tibshirani
law82                       Data for Universe of USA Law Schools
lutenhorm                   Luteinizing Hormone
mouse.c                     Experiments with mouse
mouse.t                     Experiment with mouse
patch                       The Patch Data
```

Problem Set 1

Reading Assignment 1

Chapter 1. The Statistical Sleuth: A Course in Methods of Data Analysis. Ramsey & Schafer

Reading Assignment 2

An Introduction to R : http://cran.r-project.org/doc/manuals/R-intro.pdf

Consider the Salary Data (Display 1.3) in Ramsey &Schafer, Chapter 1, and partially appended below.

i) Determine whether there are outliers in the combined data, using boxplots.

ii) Perform separate EDA, and compute appropriate measures of dispersion for the data in each group (i.e., Males and Females).

iii) For each of the estimates computed in (ii) above, determine the bias and variance using each of the following methods:

- Jackknife
- Bootstrap

SALARY	SEX
3900	FEMALE
4020	FEMALE
4290	FEMALE
4380	FEMALE
4380	FEMALE
4380	FEMALE
4380	FEMALE
4380	FEMALE
4440	FEMALE
4500	FEMALE
4500	FEMALE
4620	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4800	FEMALE
4980	FEMALE
5100	FEMALE
5100	FEMALE
5100	FEMALE
5100	FEMALE
5100	FEMALE
5100	FEMALE
5160	FEMALE
5220	FEMALE
5220	FEMALE
5280	FEMALE
5280	FEMALE
5280	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5400	FEMALE
5520	FEMALE
5520	FEMALE
4620	MALE
5040	MALE
5100	MALE
5100	MALE
5220	MALE
5400	MALE
5400	MALE
5400	MALE
5400	MALE
5400	MALE
5700	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6000	MALE
6300	MALE
6600	MALE
6600	MALE
6600	MALE
6840	MALE
6900	MALE
6900	MALE
8100	MALE