Advanced Data Analysis

Fall 2018

IMPORTANT NOTICE!

Due to over-enrollment, class open ONLY to Statistics students who will graduate in Dec 2018. NO EXCEPTION!

The course will be offered in the Spring 2019 semester.

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 and tests should reflect individual effort
- Projects may be done in groups of about 10.

- Active participation in class expected
  - Questions and answers
  - Project presentation: Mandatory for all group projects
    - For group projects, each group member is expected to present in class

GSAS Sample Statement on Academic Integrity

Columbia's intellectual community relies on academic integrity and responsibility as the cornerstone of its work. Graduate students are expected to exhibit the highest level of personal and academic honesty as they engage in scholarly discourse and research. In practical terms, you must be responsible for the full and accurate attribution of the ideas of others in all of your research papers and projects; you must be honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet source. Graduate students are responsible for knowing and correctly utilizing referencing and bibliographical guidelines. When in doubt, consult your professor. Citation and plagiarism-prevention resources can be found at the GSAS page on Academic Integrity and Responsible Conduct of Research (http://gsas.columbia.edu/academic-integrity).

Failure to observe these rules of conduct will have serious academic consequences, up to and including dismissal from the university. If a faculty member suspects a breach of academic honesty, appropriate investigative and disciplinary action will be taken following Dean's Discipline procedures (http://gsas.columbia.edu/content/disciplinary-procedures).

http://gsas.columbia.edu/content/sample-statement-academic-integrity

Instructor:

E-mail: da15@columbia.edu

Office Hours: Friday: 5:00 PM - 6 PM,
and by appointment

TAs:

Lydia Hsu	yh2692
Guanhua Fang	gf2340
William Reed Palmer	wrp2110

TA Office Hours: TBD

Course Objectives:

- Emphasis on hands-on experience with data analysis, involving case studies and using common statistical packages.

Prerequisites:

At least two applied stat courses:

- Stat GR5205 (Lin Reg), Stat GR5234 (Sample surveys), Stat GR5221 (Time Series), Stat GR5241(Stat Machine Learning), Stat GR5705 (Intro Data Sci), etc.

Topics:

- Exploratory data analysis (EDA)
- Basic stat procedures.
- Model formulation, goodness-of-fit testing
- Standard and non-standard statistical procedures, including:
  - Linear regression, classical and modern
  - Analysis of variance: ANOVA & ANCOVA
  - Generalized linear models
  - Nonlinear regression
  - Survival analysis
  - Time series analysis
  - Machine Learning: Random Forests, SVM, LASSO, Deep Learning, etc.

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 Monday, December 10, 2018.

Project:

- Students 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 projects will be accepted
- **Students expected to work in groups of about 10-15 students**
- Groups should be formed by September 15th , and the names of members listed on the first sheet at:

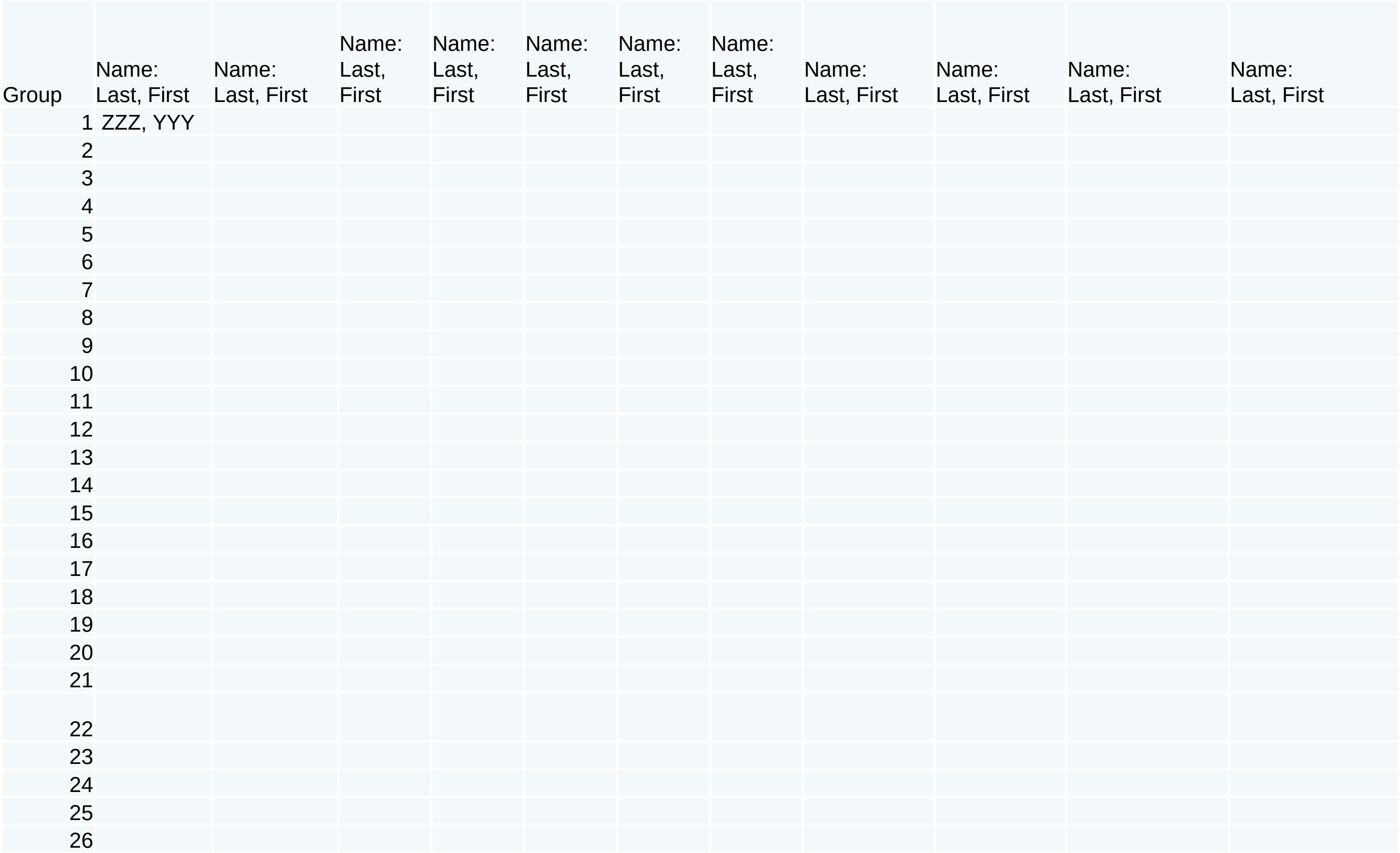
https://drive.google.com/file/d/1lwVdL018XwCPgtXePzQxBAKV2N9ednEv/view?usp=sharing

Each group will choose a date for the presentation and sign up in the 2nd sheet (sign-up sheet named: Date/Time)

  - Each group to sign up in only one time slot. If all slots are full, sign up in the Waitlist column

- During the presentation, each member of the group will be expected to participate and answer questions. The presenter will be selected at random from the group.
- All students are expected to attend the presentations and participate in the discussions.

Project Team and Presentation Sign-up (Sheet 1)



Group	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First	Name: Last, First
1	ZZZ, YYY										
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											

https://drive.google.com/file/d/1lwVdL018XwCPgtXePzQxBAKV2N9ednEv/view?usp=sharing

Date/Time of Presentation: Sign-up (Sheet 2)

Date	Time	Group Number	Group Number/ Waitlist
9/28	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
10/5	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
10/12	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
10/19	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
10/26	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
	8:30-8:40		
11/2	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
	8:30-8:40		

11/9	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
	8:30-8:40		
11/30	7:20-7:30		
	7:30-7:40		
	7:40-7:50		
	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		
12/7	6:50-7:00		
	7:00 -7:10		
	7:10-7:20		
	7:20-7:30		
	7:30-7:40		
	7:40-7:50		
	7:50-8:00		
	8:00 -8:10		
	8:10-8:20		
	8:20-8:30		

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 Duxbury

RECOMMENDED:

- *Hastie, et al (2009): The Elements of Statistical Learning:* Data Mining, Inference, and Prediction. http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Introduction Time Series and Forecasting Brockwell & Davis. Springer-Verlag

The following are also useful references

1. Miller, R. *Survival Analysis*. 1981, Wiley
2. McCullagh and Nedler. *Generalized Linear Models.* Chapman/Hall.
3. Hosmer and Lemshow. *Applied Logistic Regression*, Wiley
4. Neter, Wasserman and Kutner. *Applied Linear Statistical Models*. Wiley.

Types of Studies

- Controlled
- Observational 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

Data Generation

Experiments: Performed to generate data to help make decisions.

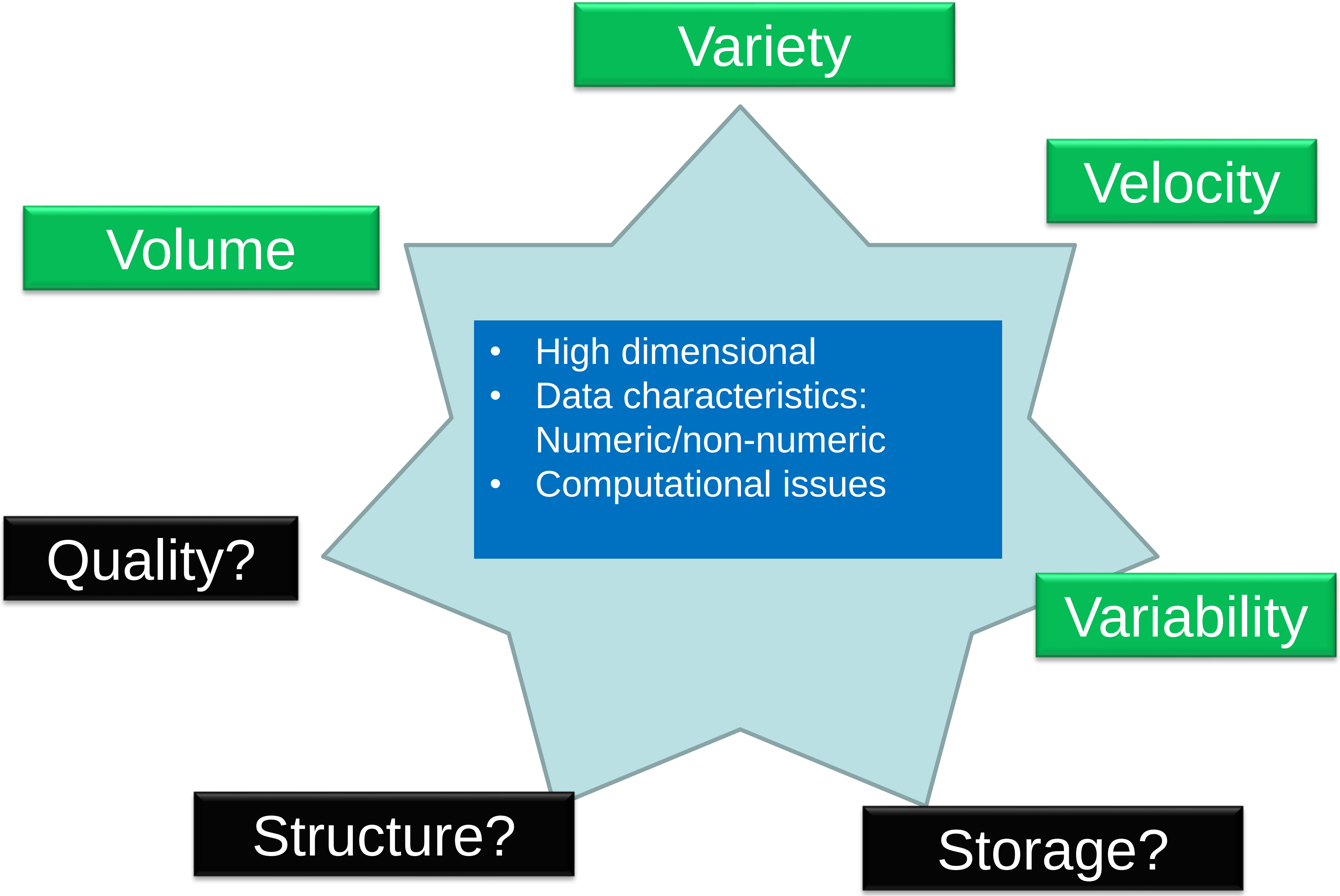
1. *Clinical Research*: Is a new therapy superior to the standard?
2. *Genetics*: Is there any association between genetic make-up and occurrence of a certain type of disease?
3. *Agriculture*: What is the effect of soil types on crop yield?
4. *Finance*: What factors affect the performance of a company's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

Generating 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?

10. Big Data?

Aspects of Big Data



Variety

Velocity

Volume

- High dimensional
- Data characteristics: Numeric/non-numeric
- Computational issues

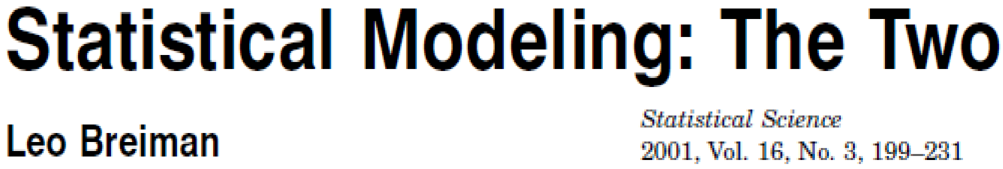
Quality?

Variability

Structure?

Storage?

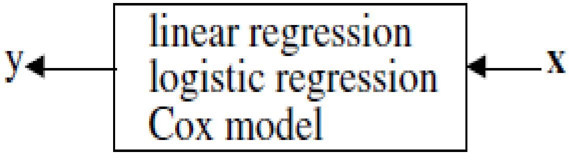
Data Analysis Paradigms


Statistical Modeling: The Two Cultures

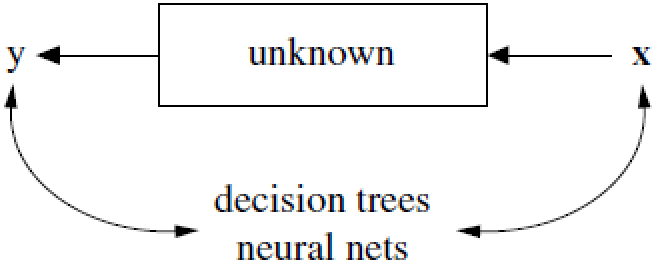
Leo Breiman

Statistical Science
2001, Vol. 16, No. 3, 199–231

Stochastic Models vs. Algorithmic Models



Knowledge of mechanism generating data

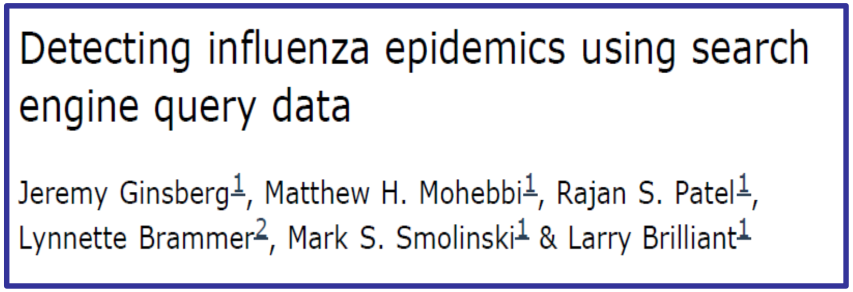


"… data characteristics are rapidly changing. In many of the most interesting problems, the idea of starting with a formal model is not tenable."
-Leo Breiman (2001)

"Automatic methods of model selection … are to be shunned or, if use is absolutely unavoidable, are to be examined carefully for their effects on the final conclusions." ..
-DR Cox (2001)

Example: Google Flu Trends (GFT)

nature



Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Nature (2009) **457**, 1012-1014

2009: Google published GFT in *Nature*:

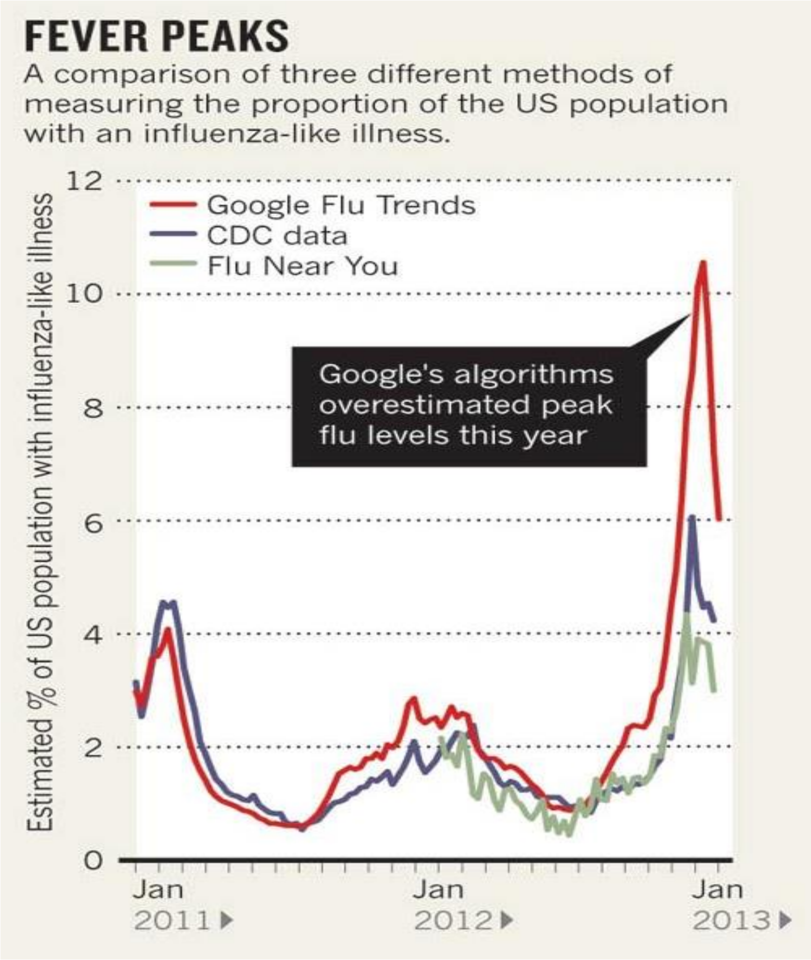
- Predicted trend without use of medical check-up data
- Could do it more quickly than the CDC: GFT a day's delay vs. a week or more CDC needed to get reports from doctors' offices
- GFT theory-free, all algorithmic based

Example: Google Flu Trends (GFT)

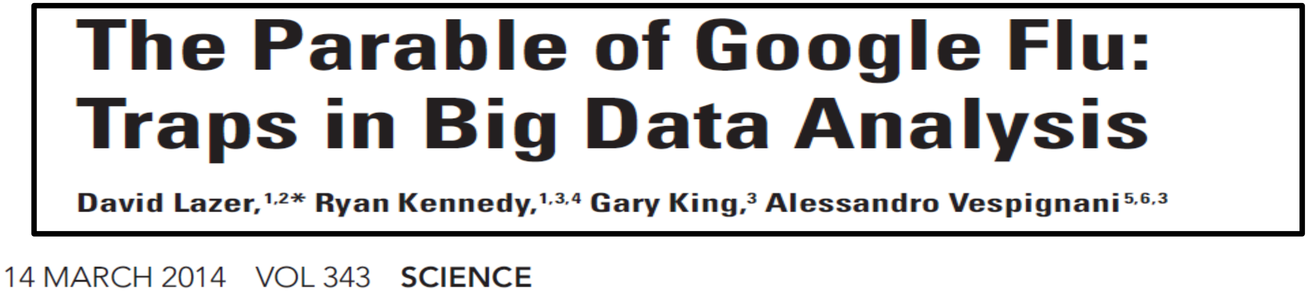
“When Google got flu wrong”

Nature Volume: 494, Pages: 155–156
Date published: (14 February 2013)
DOI: doi:10.1038/494155a

2013: GFT predicted a severe flu outbreak but CDC data showed GFT’s estimates overstated by almost a factor of two.



GFT Fiasco: What Went Wrong?



The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

14 MARCH 2014 VOL 343 **SCIENCE**

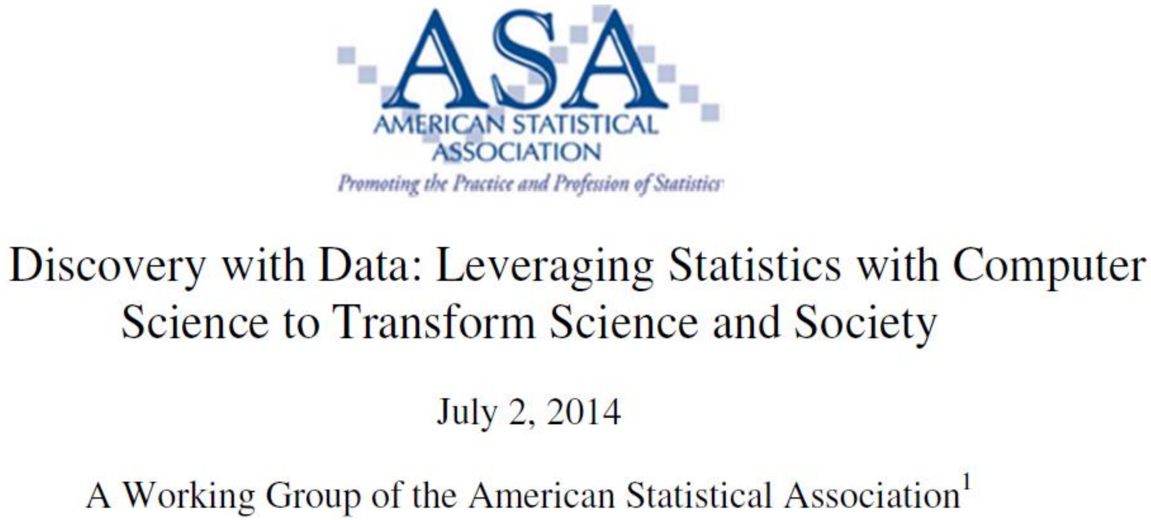
"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. ….

However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. …..

“ In general, … candidates (both at Columbia, and from other schools) were lacking in foundational statistical tools, such as hypothesis testing and identifying the assumptions of various modelling approaches. ….. many students had broad experience in the application of different techniques, especially machine learning, but very little experience in identifying the merits and drawbacks of certain approaches. Ultimately, …. we are looking for people that can differentiate between assumptions and methods and find the best tool for the task at hand.”

Feedback from a potential employer and alum

Emerging Trends



ASA
AMERICAN STATISTICAL
ASSOCIATION
Promoting the Practice and Profession of Statistics

Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society

July 2, 2014

A Working Group of the American Statistical Association[1]

http://www.amstat.org/asa/files/pdfs/POL-BigDataStatisticsJune2014.pdf

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.

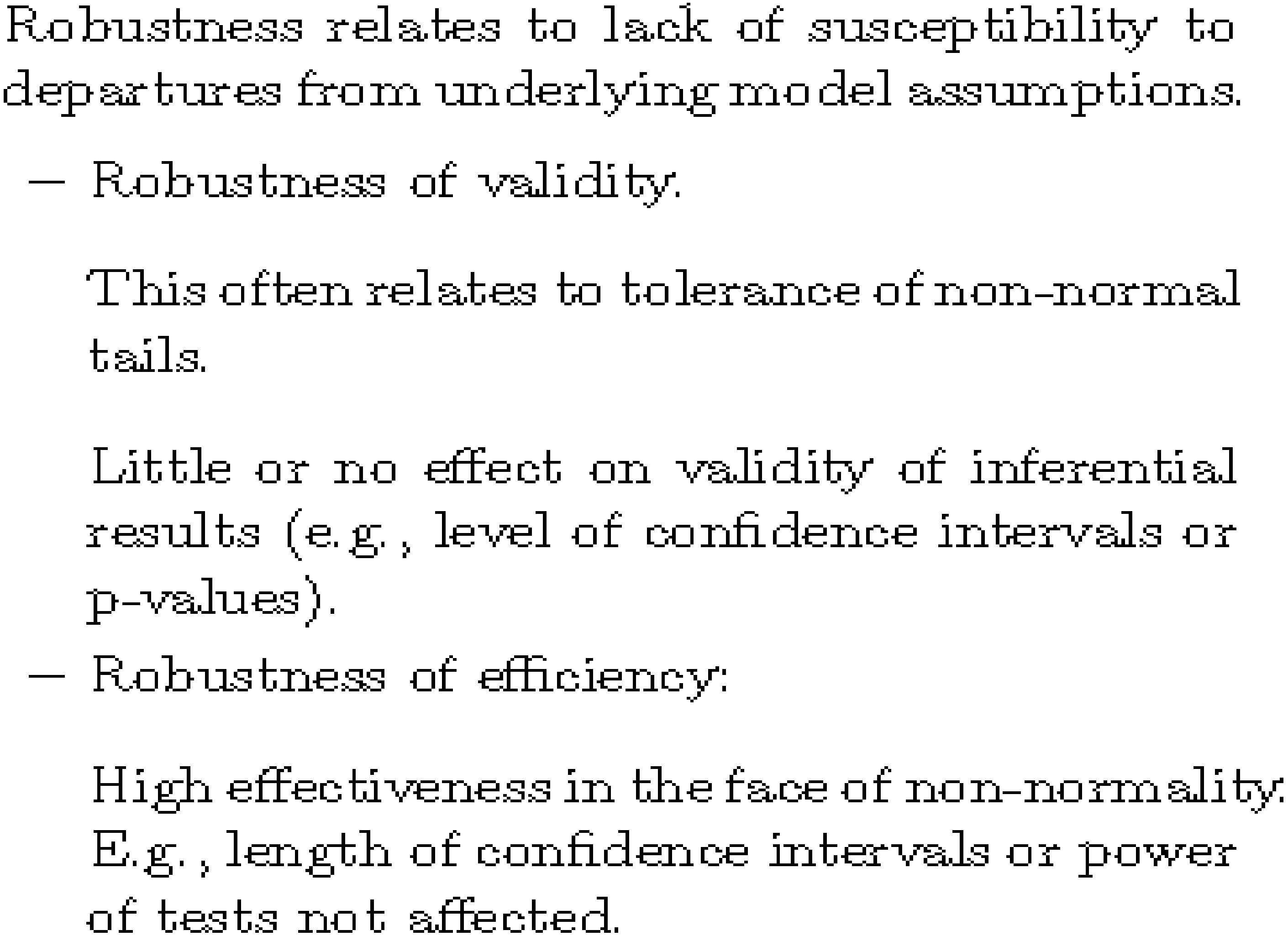
Measures of Dispersion

- 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

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

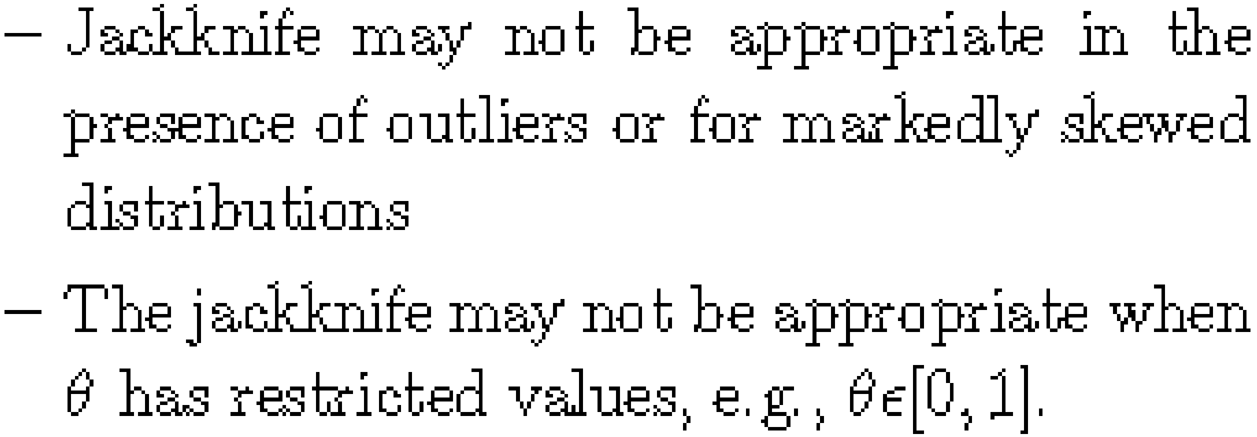
The bias reduced jackknife estimator is given by

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

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/Permutation 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(μ = 50, σ = 5).

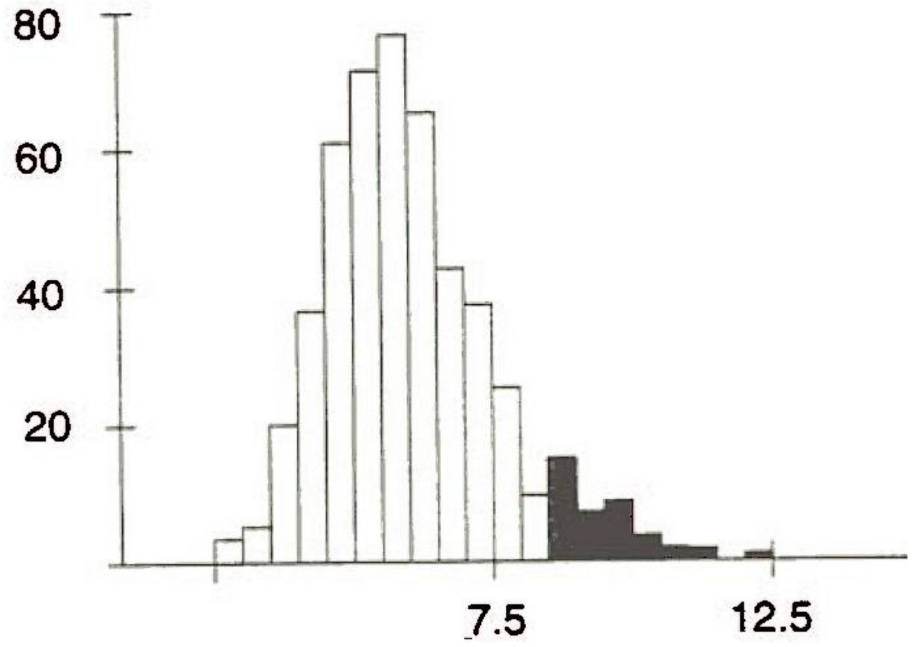
   Let $\mathfrak{S}$ be the original sample of size *N* = 20 students, and let IQR = IQR($\mathfrak{S}$) = 8.1 be our sample statistic.

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 IQR$_i$* = 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.

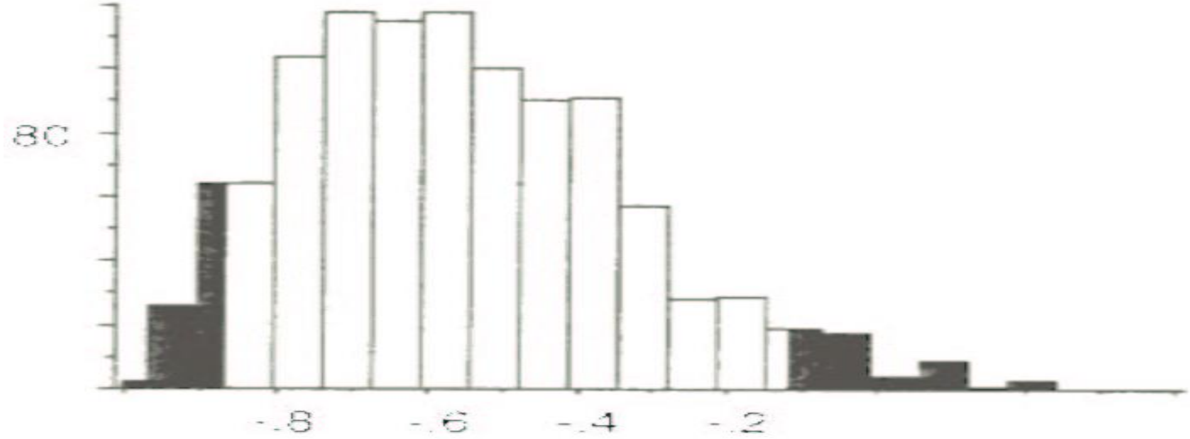
Bootstrap Methods (cont'd)

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.
- The primary hypothesis is “exchangeability”
  - Compare the observed data to all permutations

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 θ = $f(S_A, S_B)$ a statistic calculated from the two samples. Let S_{A+B} be the merge of S_A and S_B.

2. Do i = 1 … 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 $\theta_i^* = f(A_i^*, B_i^*)$ 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.

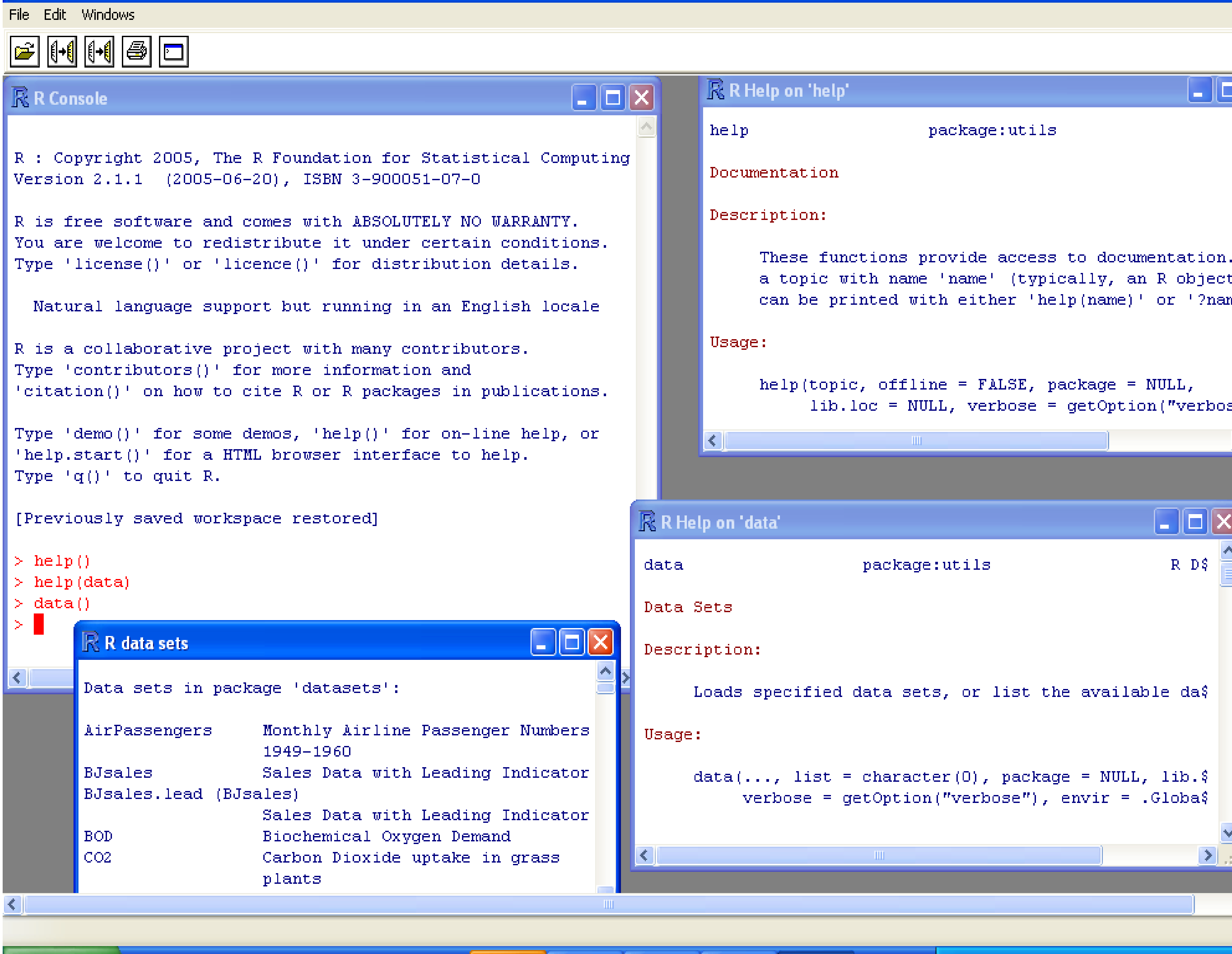
- A*pproximate randomization:* K iterations do not exhaust the space of all possible assignments of elements of S_{A+B} to A_i* and B_i*.
- *Exact randomization:* If can find exact probability by generating all possible outcomes.

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 <u>cannot</u> be used to draw inferences about population parameters, e.g., using 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.



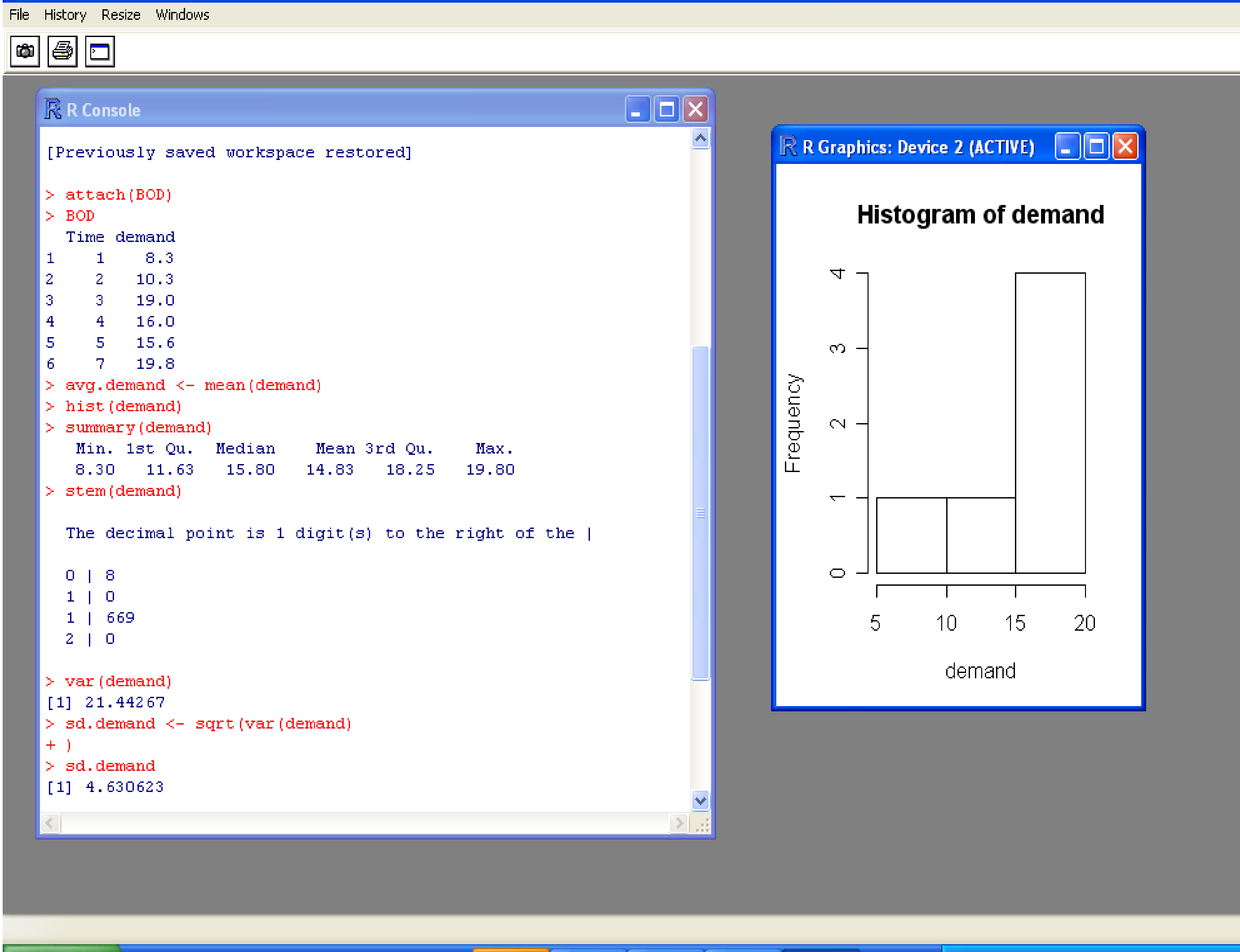
```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.1  (2005-06-20), ISBN 3-900051-07-0

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
```

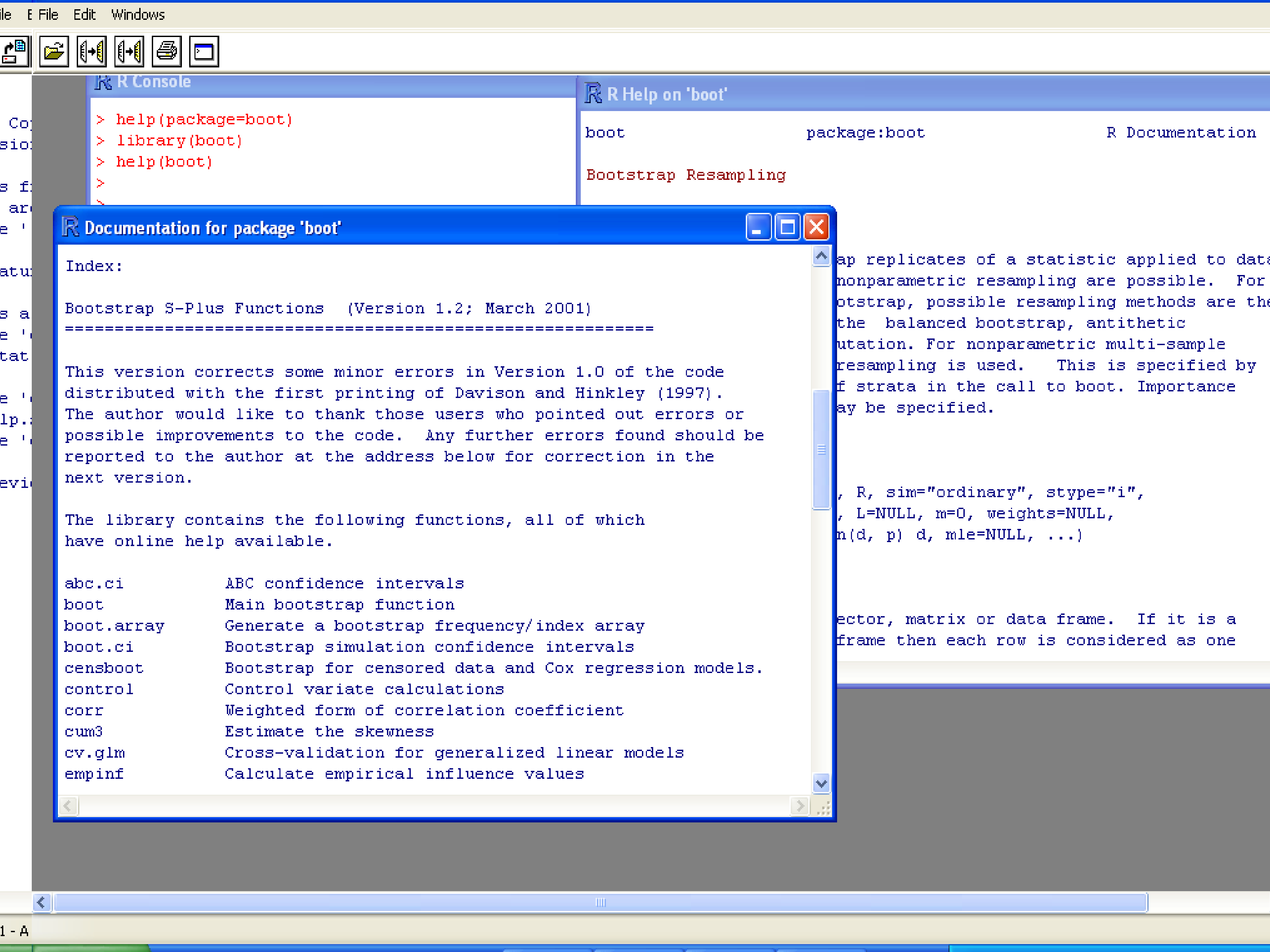
```
data                  package:utils                  R D$

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
```

```
read.table("C:\\Users\\alem\\Desktop\\zval.txt")
```



```
[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
```

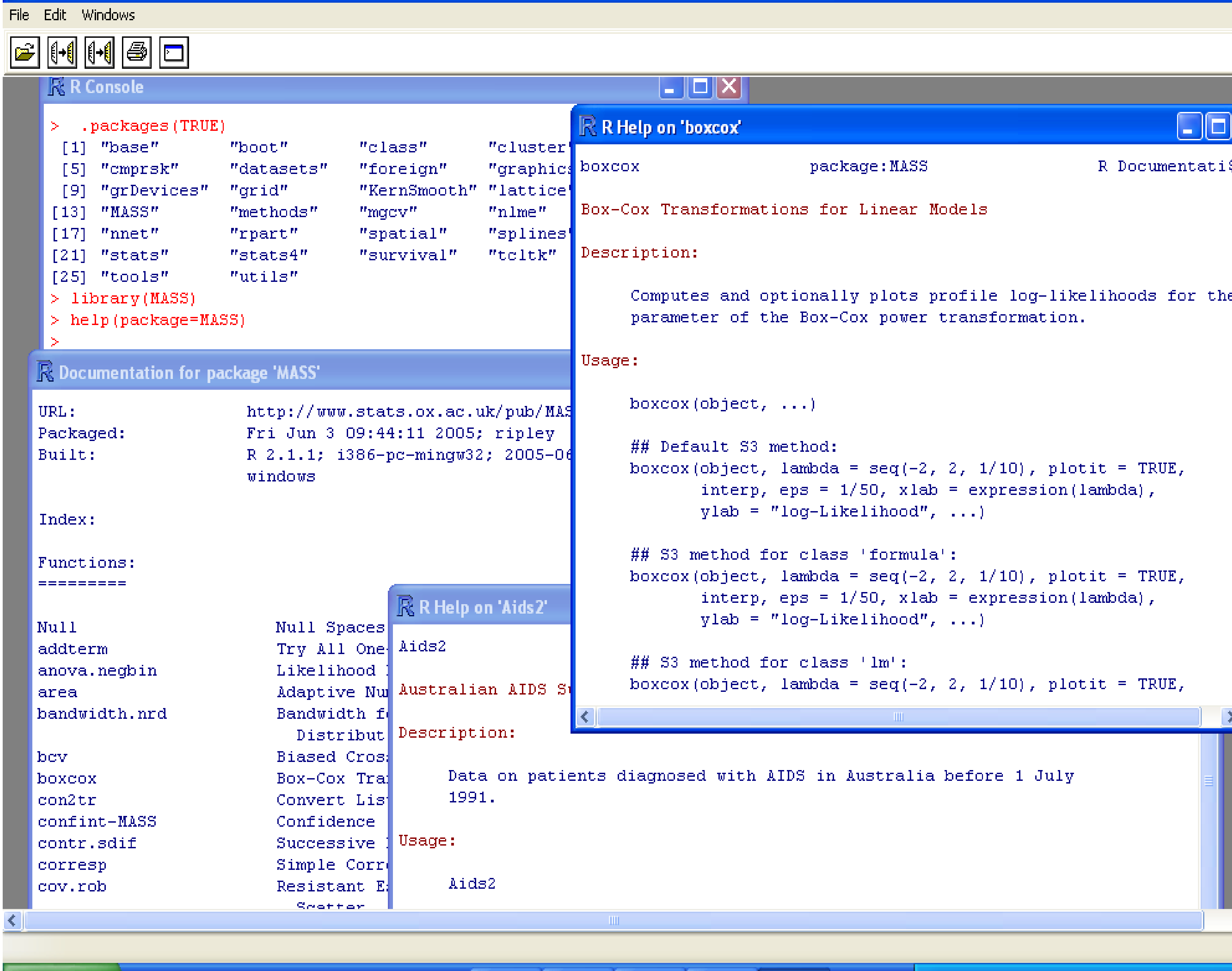
Index:

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
```

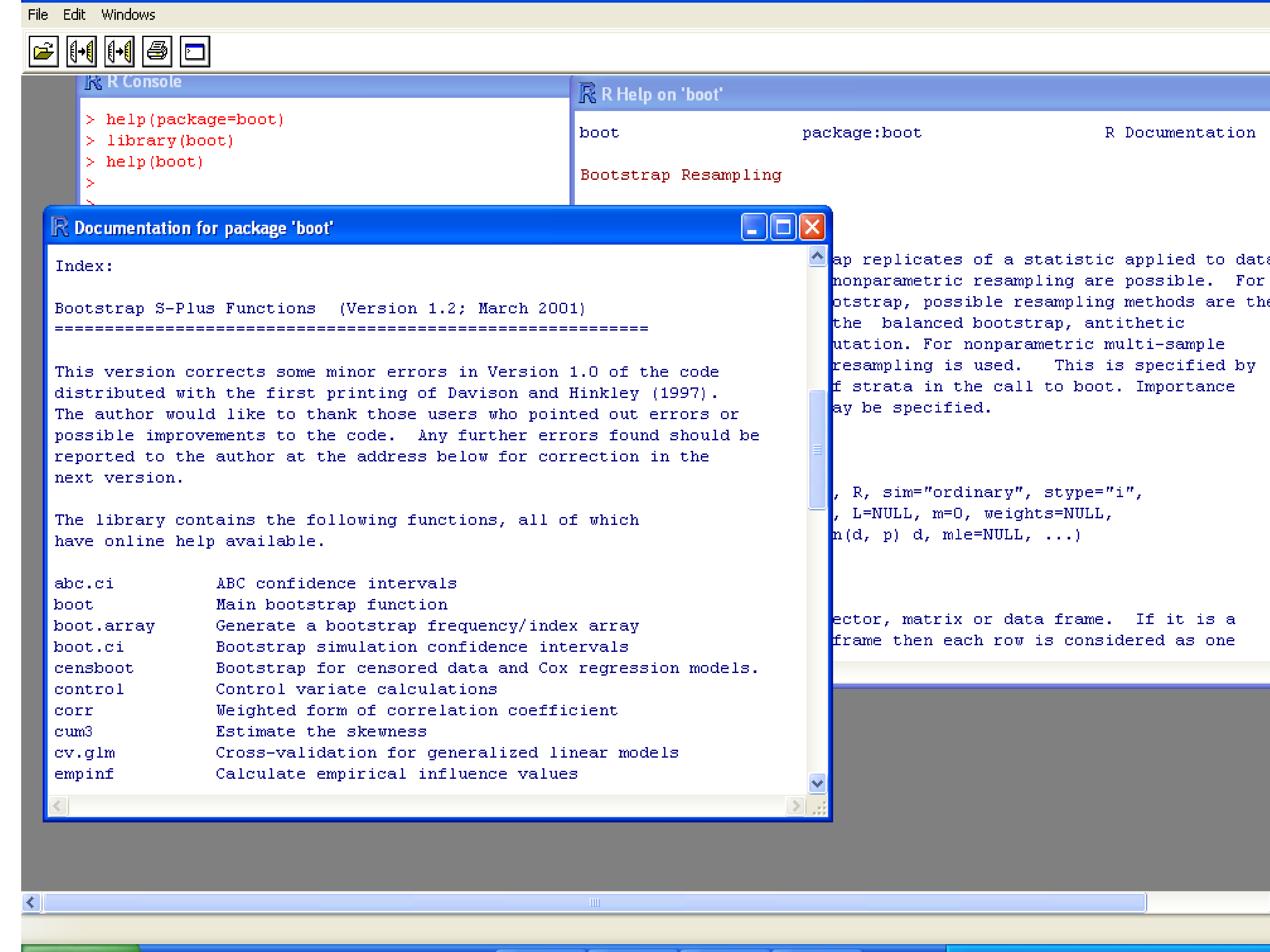


```
> .packages(TRUE)
 [1] "base"        "boot"        "class"       "cluster"
 [5] "cmprsk"      "datasets"    "foreign"     "graphics
 [9] "grDevices"   "grid"        "KernSmooth"  "lattice"
[13] "MASS"        "methods"     "mgcv"        "nlme"
[17] "nnet"        "rpart"       "spatial"     "splines"
[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
area                      Adaptive Nu
bandwidth.nrd             Bandwidth fo
                            Distribut
bcv                       Biased Cros
boxcox                    Box-Cox Tra
con2tr                    Convert Lis
confint-MASS              Confidence
contr.sdif                Successive
corresp                   Simple Corr
cov.rob                   Resistant E
                            Scatter
```

```
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,
```

```
Aids2

Australian AIDS Su

Description:

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

Usage:

     Aids2
```



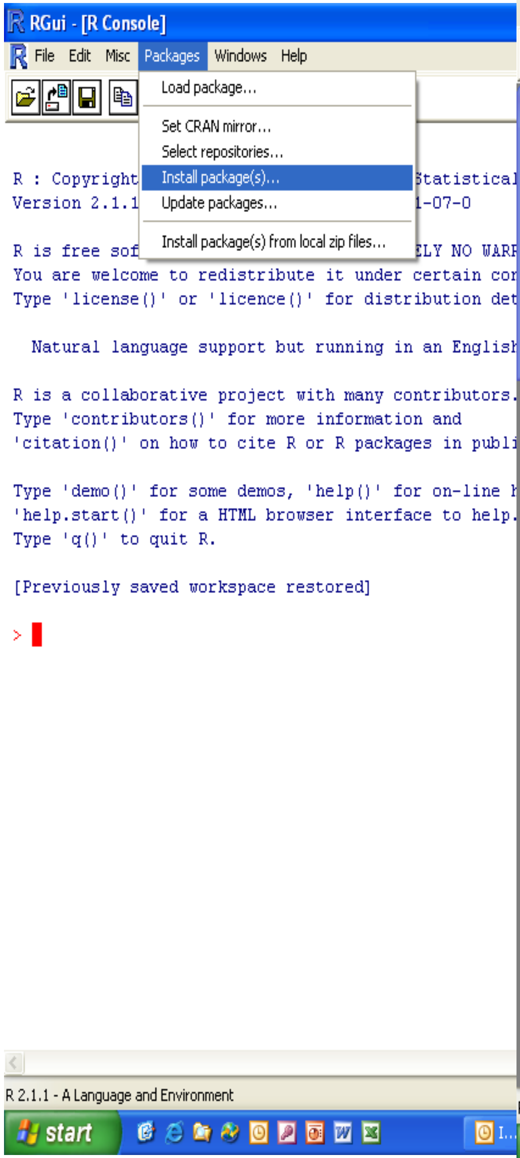
```
> help(package=boot)
> library(boot)
> help(boot)
>
```

```
boot                    package:boot                    R Documentation

Bootstrap Resampling
```

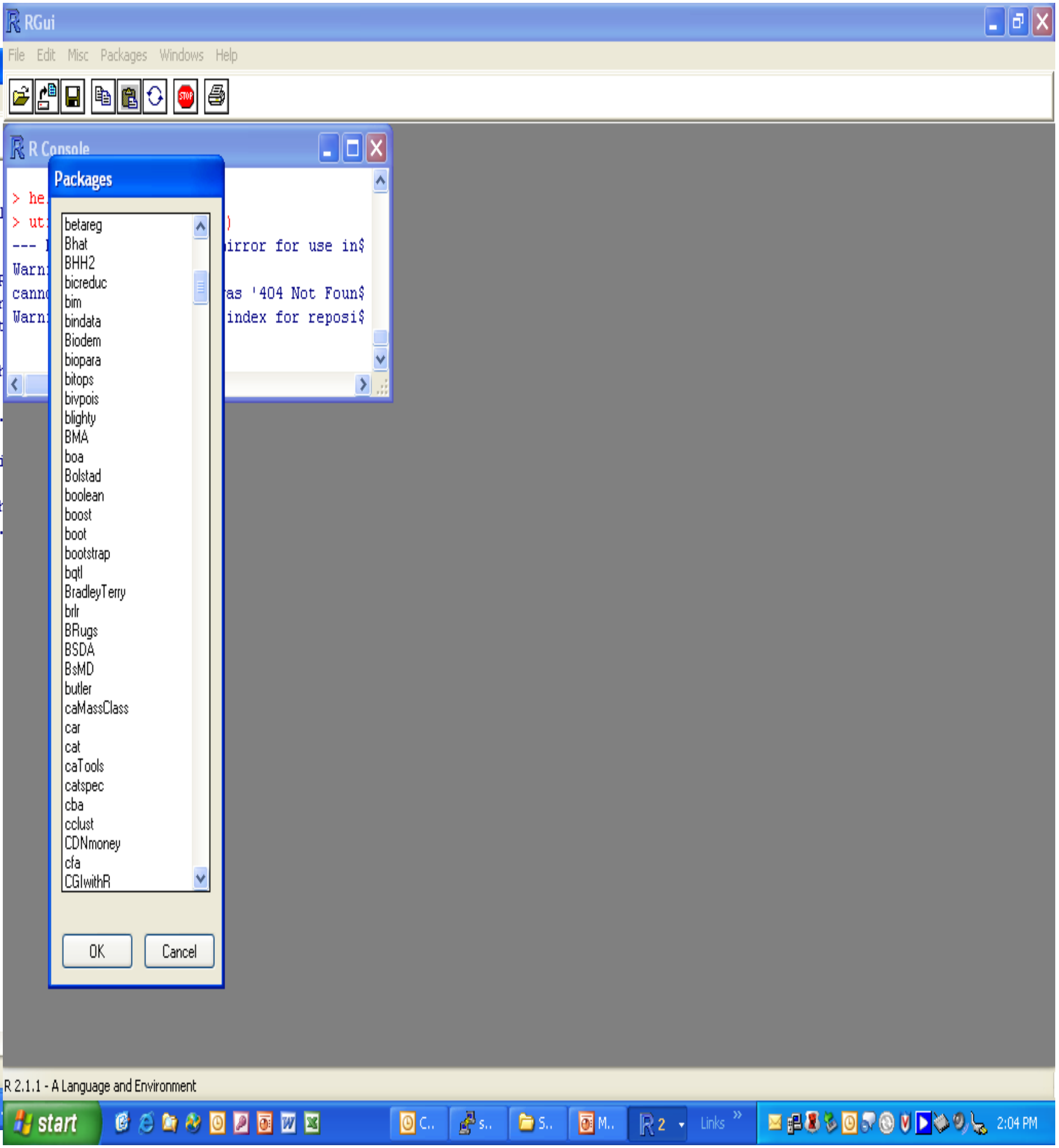
```
ap replicates of a statistic applied to data
nonparametric resampling are possible.  For
otstrap, possible resampling methods are the
the  balanced bootstrap, antithetic
utation. For nonparametric multi-sample
resampling is used.   This is specified by
f strata in the call to boot. Importance
ay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

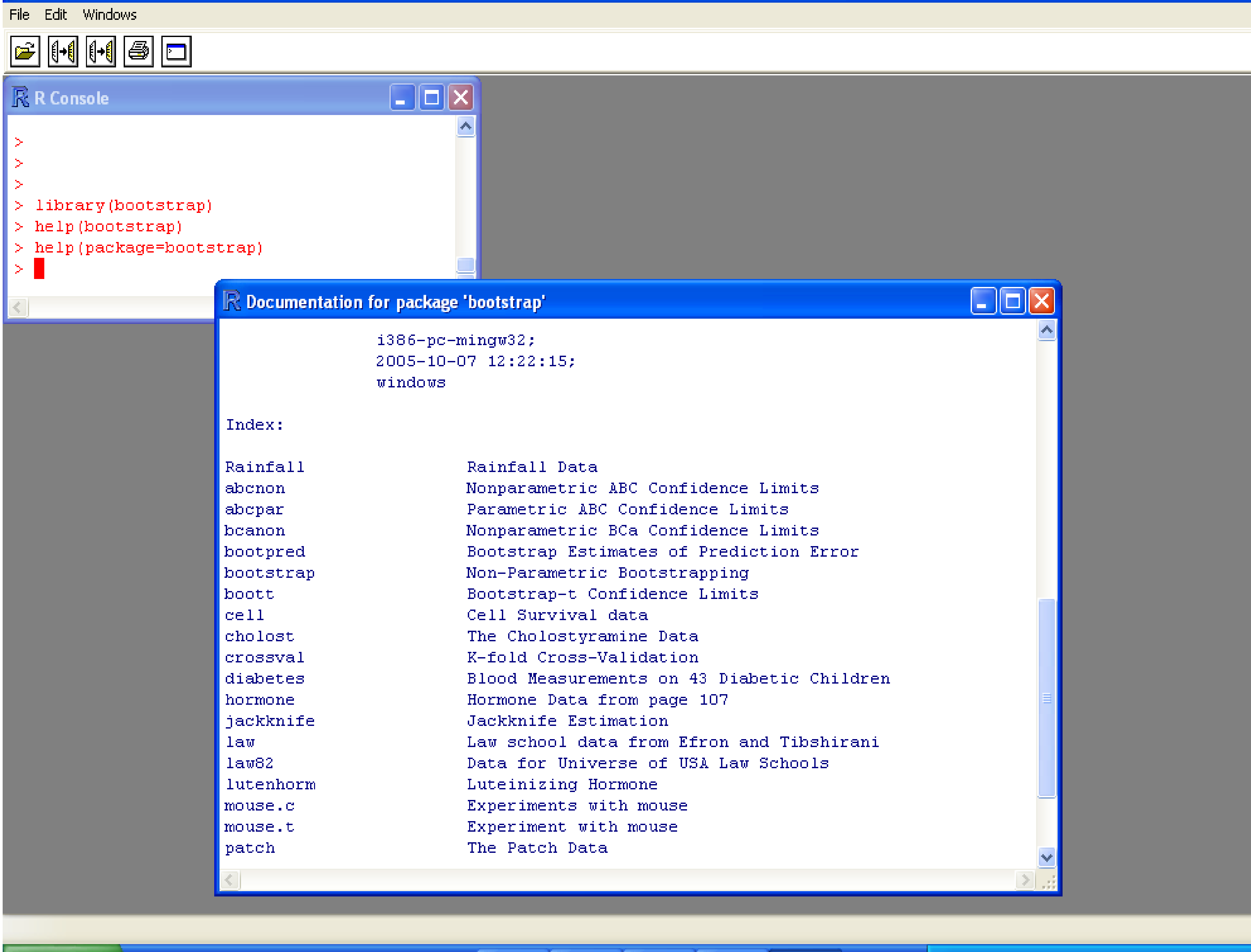
#Go to ***Packages***, ***Install Packages***, (select USA1, etc.) Bootstrap



```
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- Fall 2018

1. Form Project Teams, and sign up for presentation times

2. Reading Assignment 1

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

3. Consider the ToothGrowth data in R, concerning the Effect of Vitamin C on Tooth Growth in Guinea Pigs. Ignore 'dose', and

   i) Perform separate EDA for each "supplement" , (i.e., OJ /VC)., and compute the following for "len":

      a) Sample kurtosis
      b) Sample median
      c) Sample Inter Quartile Range (IQR)
      d) Sample IQR

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

      - Jackknife
      - Bootstrap

   iii) Bonus credit: Construct a 95% bootstrap confidence interval for the difference in IQRs between the "supplement" groups (i.e., OJ vs VC).