Using R for the Management of Survey Data and Statistics in Evaluation

Lindsey Dunn
Lauren Fluegge
Korinne Chiu

UNCG
Department of Educational Research Methodology

Introduction to R

- Importing Data
- Exporting Data
- Overview of screening and preliminary analysis of survey data with R package 'memisc'
- Overview of qualitative GUI 'RQDA'

What Can R Do?

- R is FREE
- R is flexible
  - R is both a stat package and a programming language
- Many packages have already been created
  - Everyday functions are available
  - Manuals are provided for each package
- There is a large and active development/user community
  - Tools are being added to R everyday

How to Get R

- http://cran.r-project.org/
  - Main R site
  - Downloads
    - Including packages
  - FAQs
  - Manuals / Help Files
  - **Contributed Documentation**
    - http://cran.r-project.org/other-docs.html

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

Source Code for all Platforms

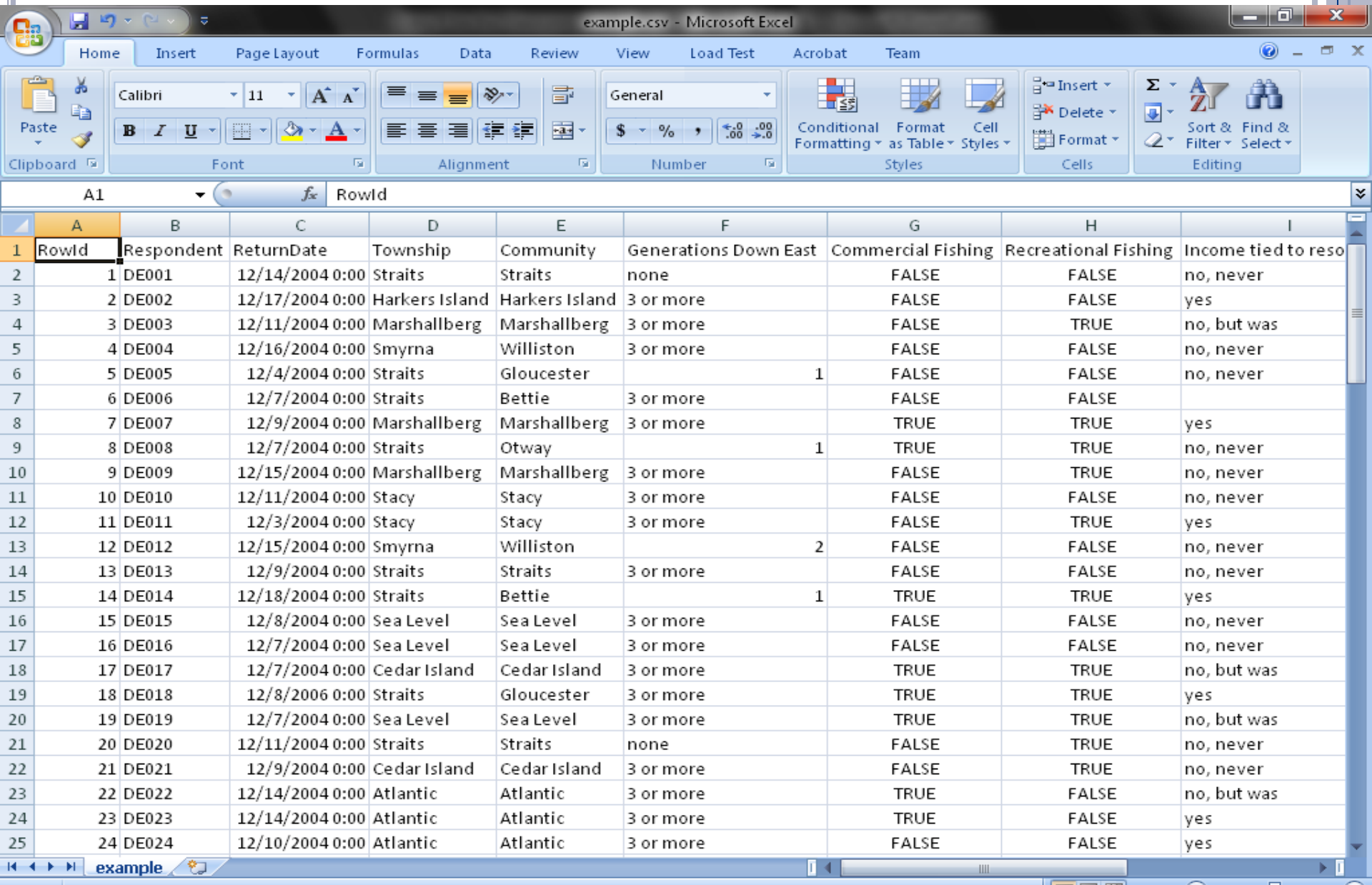
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2011-09-30): R-2.13.2.tar.gz (read what's new in the latest version).
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Reading and Writing Files

- Read in a data table
  - read.table("filename.txt")
- Start working with data.frames and data.set
  - Names, displaying and assigning values
- Exporting data in database/table form
  - write.table("filename.txt")

EXAMPLE.CSV



RowId	Respondent	ReturnDate	Township	Community	Generations Down East	Commercial Fishing	Recreational Fishing	Income tied to reso
1	DE001	12/14/2004 0:00	Straits	Straits	none	FALSE	FALSE	no, never
2	DE002	12/17/2004 0:00	Harkers Island	Harkers Island	3 or more	FALSE	FALSE	yes
3	DE003	12/11/2004 0:00	Marshallberg	Marshallberg	3 or more	FALSE	TRUE	no, but was
4	DE004	12/16/2004 0:00	Smyrna	Williston	3 or more	FALSE	FALSE	no, never
5	DE005	12/4/2004 0:00	Straits	Gloucester	1	FALSE	FALSE	no, never
6	DE006	12/7/2004 0:00	Straits	Bettie	3 or more	FALSE	FALSE	
7	DE007	12/9/2004 0:00	Marshallberg	Marshallberg	3 or more	TRUE	TRUE	yes
8	DE008	12/7/2004 0:00	Straits	Otway	1	TRUE	TRUE	no, never
9	DE009	12/15/2004 0:00	Marshallberg	Marshallberg	3 or more	FALSE	TRUE	no, never
10	DE010	12/11/2004 0:00	Stacy	Stacy	3 or more	FALSE	FALSE	no, never
11	DE011	12/3/2004 0:00	Stacy	Stacy	3 or more	FALSE	TRUE	yes
12	DE012	12/15/2004 0:00	Smyrna	Williston	2	FALSE	FALSE	no, never
13	DE013	12/9/2004 0:00	Straits	Straits	3 or more	FALSE	FALSE	no, never
14	DE014	12/18/2004 0:00	Straits	Bettie	1	TRUE	TRUE	yes
15	DE015	12/8/2004 0:00	Sea Level	Sea Level	3 or more	FALSE	FALSE	no, never
16	DE016	12/7/2004 0:00	Sea Level	Sea Level	3 or more	FALSE	FALSE	no, never
17	DE017	12/7/2004 0:00	Cedar Island	Cedar Island	3 or more	TRUE	TRUE	no, but was
18	DE018	12/8/2006 0:00	Straits	Gloucester	3 or more	TRUE	TRUE	yes
19	DE019	12/7/2004 0:00	Sea Level	Sea Level	3 or more	TRUE	TRUE	no, but was
20	DE020	12/11/2004 0:00	Straits	Straits	none	FALSE	TRUE	no, never
21	DE021	12/9/2004 0:00	Cedar Island	Cedar Island	3 or more	FALSE	TRUE	no, never
22	DE022	12/14/2004 0:00	Atlantic	Atlantic	3 or more	TRUE	FALSE	no, but was
23	DE023	12/14/2004 0:00	Atlantic	Atlantic	3 or more	TRUE	FALSE	yes
24	DE024	12/10/2004 0:00	Atlantic	Atlantic	3 or more	FALSE	FALSE	yes

Importing Data into R

- Reads data in and converts into a data frame
  - Flexible data organization of numbers and characters within the same database
- Data can be imported from Excel, SPSS, SAS, etc.
- read.table("file.txt")
- If reading in an SPSS file, you will need to install the package "foreign" and call that library first
  - Certain functions require the installation of a package prior to use

Importing Data into R

- Many options
  - read.table(file, **header = FALSE**, **sep = ""**, quote = "\"", dec = ".", **row.names, col.names**, as.is = !stringsAsFactors, na.strings = "NA", colClasses = NA, nrows = -1, skip = 0, check.names = TRUE, fill = !blank.lines.skip, strip.white = FALSE, blank.lines.skip = TRUE, comment.char = "#", allowEscapes = FALSE, flush = FALSE, stringsAsFactors = default.stringsAsFactors(), fileEncoding = "", encoding = "unknown")
  - ?read.table for complete syntax and descriptions

Importing Data into R

- Can assign the data set to an object to use it later
- survey<-read.csv("example.csv", header=TRUE)
- Survey
- 'memisc' uses importer mechanism to import data
- spss.fixed.file(file, varlab.file=NULL, codes.file=NULL, missval.file=NULL, count.cases=TRUE)

Importing Data into R

RGui

File Edit Packages Windows Help

R Console

```
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ?read.table
starting httpd help server ... done
> survey<-read.csv("example.csv", header=TRUE)
> survey
    RowId Respondent      ReturnDate       Township      Community
1       1       DE001 12/14/2004 0:00        Straits        Straits
2       2       DE002 12/17/2004 0:00 Harkers Island Harkers Island
3       3       DE003 12/11/2004 0:00   Marshallberg   Marshallberg
4       4       DE004 12/16/2004 0:00         Smyrna      Williston
5       5       DE005  12/4/2004 0:00        Straits     Gloucester
6       6       DE006  12/7/2004 0:00        Straits         Bettie
7       7       DE007  12/9/2004 0:00   Marshallberg   Marshallberg
8       8       DE008  12/7/2004 0:00        Straits          Otway
9       9       DE009 12/15/2004 0:00   Marshallberg   Marshallberg
10     10       DE010 12/11/2004 0:00          Stacy          Stacy
11     11       DE011  12/3/2004 0:00          Stacy          Stacy
12     12       DE012 12/15/2004 0:00         Smyrna      Williston
13     13       DE013  12/9/2004 0:00        Straits        Straits
14     14       DE014 12/18/2004 0:00        Straits         Bettie
15     15       DE015  12/8/2004 0:00      Sea Level      Sea Level
16     16       DE016  12/7/2004 0:00      Sea Level      Sea Level
17     17       DE017  12/7/2004 0:00   Cedar Island   Cedar Island
```

Untitled - R Editor

```
?read.table
survey<-read.csv("example.csv", header=TRUE)
survey
```

Exporting data from R

- Many options
  - write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ", eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE, qmethod = c("escape", "double"))
  - Can indicate how variables are separated (sep=“,”)
  - Can indicate the string to use for missing values (na=“NA”)
  - Can indicate column names (col.names=TRUE) or row names (row.names=TRUE)
  - ?write.table for complete syntax and descriptions

Project Example

- Sample Project from NVIVO9
  - Qualitative and Quantitative Data about a particular area in NC
    - Property
    - Environment
- Survey Portion-using memisc
- Interview Portion-using RQDA

Overview: R package 'memisc'

- Overview of 'memisc' Features
- Highlight the Advantages of 'memisc'
- Demonstrate select survey management functions from 'memisc'

'memisc' Features

- Documentation file versus Description File
  - help(package="memisc")
- Author/Maintainer:Martin Elff <melff at essex.ac.uk>
  http://www.martin-elff.net/software/memisc/
- Depends: lattice, grid, stats, methods, utils, MASS
- Vignettes: Analysing the American National Election Study of 1948
- CRAN checks: memisc results

'memisc' Features: General Function

- Provides tools for the management and analysis of survey data
- Creates tables of descriptive statistics and regression model estimates
- Graphics, programming, and simulation capabilities

R Package 'memisc' Advantages

- Includes a lot of functionality
- Provides flexible data visualization tools
- Well maintained package

http://www.martin-elff.net/software/memisc/

Demonstration: Presentation Focus

- Functions used in the following stages of an evaluation:
  - Data preparation
  - Data management
  - Preliminary analyses
  - Publication-ready table creation

Data preparation and management

str(object, vec.len = strO$vec.len, digits.d = strO$digits.d, nchar.max = 128, ...)

- Compactly displays internal structure of R object
- Must assign to a new object to print easily

str(survey)

```
> str(surveyDS)
Data set with 104 obs. of 13 variables:
 $ survey.RowId                    : Itvl. item  int  1 2 3
 $ survey.Respondent               : Nmnl. item w/ 104 label
 $ survey.ReturnDate               : Nmnl. item w/ 14 labels
 $ survey.Township                 : Nmnl. item w/ 9 labels
 $ survey.Community                : Nmnl. item w/ 13 labels
 $ survey.Generations.Down.East    : Nmnl. item w/ 4 labels
 $ survey.Commercial.Fishing       : Itvl. item  int  0 0 0
 $ survey.Recreational.Fishing     : Itvl. item  int  0 0 1
 $ survey.Income.tied.to.resources: Nmnl. item w/ 5 labels
 $ survey.Pace.of.development      : Nmnl. item w/ 5 labels
 $ survey.Age                      : Itvl. item  int  61 62
 $ survey.Gender                   : Nmnl. item w/ 2 labels
 $ survey.Education.Level          : Nmnl. item w/ 10 labels
```

str(survey$Income.tied.to.resources)

```
> str(surveyDS$survey.Income.tied.to.resources)
 Nmnl. item w/ 5 labels for 1,2,3,... + ms.v.  int
```

Data preparation and management

trimws(x,left=TRUE,right=TRUE)

- Removes leading or trailing whitespaces
- ‘left’ and ‘right’ arguments specify from which side of the character vector to remove the whitespace

Data preparation and management

names(x)

- Prints, removes, or assigns new names to all or part of the name attribute of a vector

rename(x,..., gsub = FALSE, warn = TRUE))

- Changes the names of a named object
- ‘gsub’ changes parts of row and column labels
- ‘warn’ issues a warning if old values are not found

names(survey)
Rename(survey, .="_", gsub=TRUE)

```
> names(survey)
 [1] "RowId"                          "Respondent"
 [3] "ReturnDate"                     "Township"
 [5] "Community"                      "Generations.Down.East"
 [7] "Commercial.Fishing"             "Recreational.Fishing"
 [9] "Income.tied.to.resources"       "Pace.of.development"
[11] "Age"                            "Gender"
[13] "Education.Level"
> surveyR <- rename(survey,
+                   .="_"
+                   ,gsub=TRUE)
> names(surveyR)
 [1] "RowId"                          "Respondent"
 [3] "ReturnDate"                     "Township"
 [5] "Community"                      "Generations_Down_East"
 [7] "Commercial_Fishing"             "Recreational_Fishing"
 [9] "Income_tied_to_resources"       "Pace_of_development"
[11] "Age"                            "Gender"
[13] "Education Level"
```

Data preparation and management

description(x)

- Prints or assigns new “variable label”

wording(x)

- Prints or assigns new question wording of an item

labels(x)

- Prints or associates character labels with possible values of an encoded survey item

Data preparation and management

missing.values(x)

- Specifies individual missing values and/or a range of missing values
- Can be updated using the + and - operators

annotation(x)

- objects of class "annotation", are character vectors

codebook(x)

- collects and returns formatted documentation about an item or data set

Syntax

```
surveyDS <- within(surveyDS,{
    description(survey.Income.tied.to.resources) <-
        "Income depends on land or water
         resources Down East"
    wording(survey.Income.tied.to.resources) <- "Is
          your income currently tied to land or water
          resources Down East?"
    labels(survey.Income.tied.to.resources) <- c(
           Never = 3, Yes = 5, Formerly = 2)
    annotation(survey.Income.tied.to.resources)["Re
           mark"] <- "Note: This item contained a
           mistype?"
    missing.values(survey.Income.tied.to.resources)<-
    c(1,4)
 })
```

codebook(surveyDS$Income.tied.to.resources)

```
   'Income depends on land or water resources
   Down East'

   "Is your income currently tied to land or
   water resources Down East?"

----------------------------------------------

   Storage mode: integer
   Measurement: nominal
   Missing values: 1, 4

   Values and labels     N    Percent

    2   'Formerly'      15   14.9 14.4
    3   'Never'         60   59.4 57.7
    5   'Yes'           26   25.7 25.0
      M (unlab.mss.)     3         2.9

   Remark:

       Note: This item contained a
       mistype
```

Preliminary Data Analyses

Descriptives(x)

- gives a vector of sample statistics

genTable(formula, data=parent.frame(), subset=NULL,...)

- constructs table of summaries conditional on given values of independent variables given by a formula

aggregate(x, data=parent.frame(),...)

- same as genTable but constructs a data frame

Descriptives(surveyDS$survey.Age)

```
> Descriptives(surveyDS$survey.Age)
        Min          Max         Mean    Std.Dev.     Skewness     Kurtosis
32.00000000 81.00000000 57.18269231 11.13100695 -0.01158689 -0.63259413
```

survey.table <- genTable(percent(survey.Income.tied.to.resources) ~ survey.Community.data=surveyDS)

```
> surveyDS.table
                                          survey.Community
percent(survey.Income.tied.to.resources) Atlantic   Bettie Cedar Island
                                Formerly 22.22222  0.00000     33.33333
                                Never    55.55556 50.00000     33.33333
                                Yes      22.22222 50.00000     33.33333
                                N         9.00000  8.00000      3.00000
                                          survey.Community
percent(survey.Income.tied.to.resources)    Davis Gloucester Harkers Island
                                Formerly 16.66667   11.11111       20.00000
                                Never    66.66667   44.44444       30.00000
                                Yes      16.66667   44.44444       50.00000
                                N         6.00000    9.00000       10.00000
                                          survey.Community
percent(survey.Income.tied.to.resources) Marshallberg    Otway Sea Level
                                Formerly     30.00000 11.11111  22.22222
                                Never        40.00000 77.77778  77.77778
                                Yes          30.00000 11.11111   0.00000
                                N            10.00000  9.00000   9.00000
```

```
survey.agg  <-
aggregate(percent(survey.Income.tied.to.resources)
~ survey.Community.data=surveyDS)
```

```
$rveyDS.agg
$urvey.Community Formerly    Never       Yes  N
$       Atlantic 22.22222 55.55556 22.22222  9
$         Bettie  0.00000 50.00000 50.00000  8
$   Cedar Island 33.33333 33.33333 33.33333  3
$          Davis 16.66667 66.66667 16.66667  6
$     Gloucester 11.11111 44.44444 44.44444  9
$ Harkers Island 20.00000 30.00000 50.00000 10
$   Marshallberg 30.00000 40.00000 30.00000 10
$          Otway 11.11111 77.77778 11.11111  9
$      Sea Level 22.22222 77.77778  0.00000  9
$         Smyrna 10.00000 70.00000 20.00000 10
$          Stacy  0.00000 60.00000 40.00000  5
$        Straits  0.00000 87.50000 12.50000  8
$      Williston 20.00000 80.00000  0.00000  5
```

toLatex(surveyDS.table)

```
> toLatex(surveyDS.table)
\begin{tabular}{lD{.}{.}{0}D{.}{.}{0}D{.}{.}{0}D{.}$
\toprule
 & \multicolumn{1}{c}{Atlantic} & \multicolumn{1}{c$
\midrule
Formerly & 22 & 0 & 33 & 17 & 11 & 20 & 30 & 11 & 2$
Never & 56 & 50 & 33 & 67 & 44 & 30 & 40 & 78 & 78 $
Yes & 22 & 50 & 33 & 17 & 44 & 50 & 30 & 11 & 0 & 2$
N & 9 & 8 & 3 & 6 & 9 & 10 & 10 & 9 & 9 & 10 & 5 & $
\bottomrule
\end{tabular}
```

http://www.r-bloggers.com/getting-started-with-sweave-r-latex-eclipse-statet-texlipse/

Summary of 'memisc'

- Once the user is familiar with the package it is quicker than point and click software
- Provides easy method of creating a codebook
- Flexible input and output file formats
- Creates publication-ready tables in a matter of seconds

Overview: R package 'RQDA'

- 'RQDA' is a qualitative package used to analyze text data
- Downloadable package works on Windows, Linux, and Mac systems
- Allows for qualitative and quantitative analyses in one place

Where to find RQDA

- http://rqda.r-forge.r-project.org/
- Be sure to first install dependent programs
  - The Windows system will do this automatically, but you have to prompt Mac
  - If packages do not load after installation, simply restart R

RQDA Program Information

- Title R-based Qualitative Data Analysis
- Version 0.2-1
- Date 2011-09-10
- Author HUANG Ronggui
- Maintainer HUANG Ronggui <ronggui.huang@gmail.com>
- Depends R (>= 2.8.0), DBI, RSQLite, gWidgets (>= 0.0-31),gWidgetsRGtk2 (>= 0.0-36)
- Imports RGtk2 (>= 2.20), methods, igraph
- URL http://rqda.r-forge.r-project.org/

Helpful Tools

- The manual for RQDA can be found at: http://cran.r-project.org/web/packages/RQDA/RQDA.pdf
- For tips, use http://rqda.r-forge.r-project.org/documentation.html#tips
- help(package='RQDA') provides the user with RQDA functions
- ?RQDA will send you to a website that provides basic program information, but not functions

RQDA Capabilities

- Both GUI platform and R functions
  - Presently, more work can be done in GUI
- Import multiple documents
  - Can support other languages
- Create coding indices
  - And code categories (helpful for theory building)
  - Export codings
- Analyze relationships between codings
- Create attributes
- Memo-ing
- Delete data and clean files

Demonstration

- Interview portion of Project

- Using GUI:
  - Opening a document
  - Creating codes
  - Coding an interview
  - Memo-ing

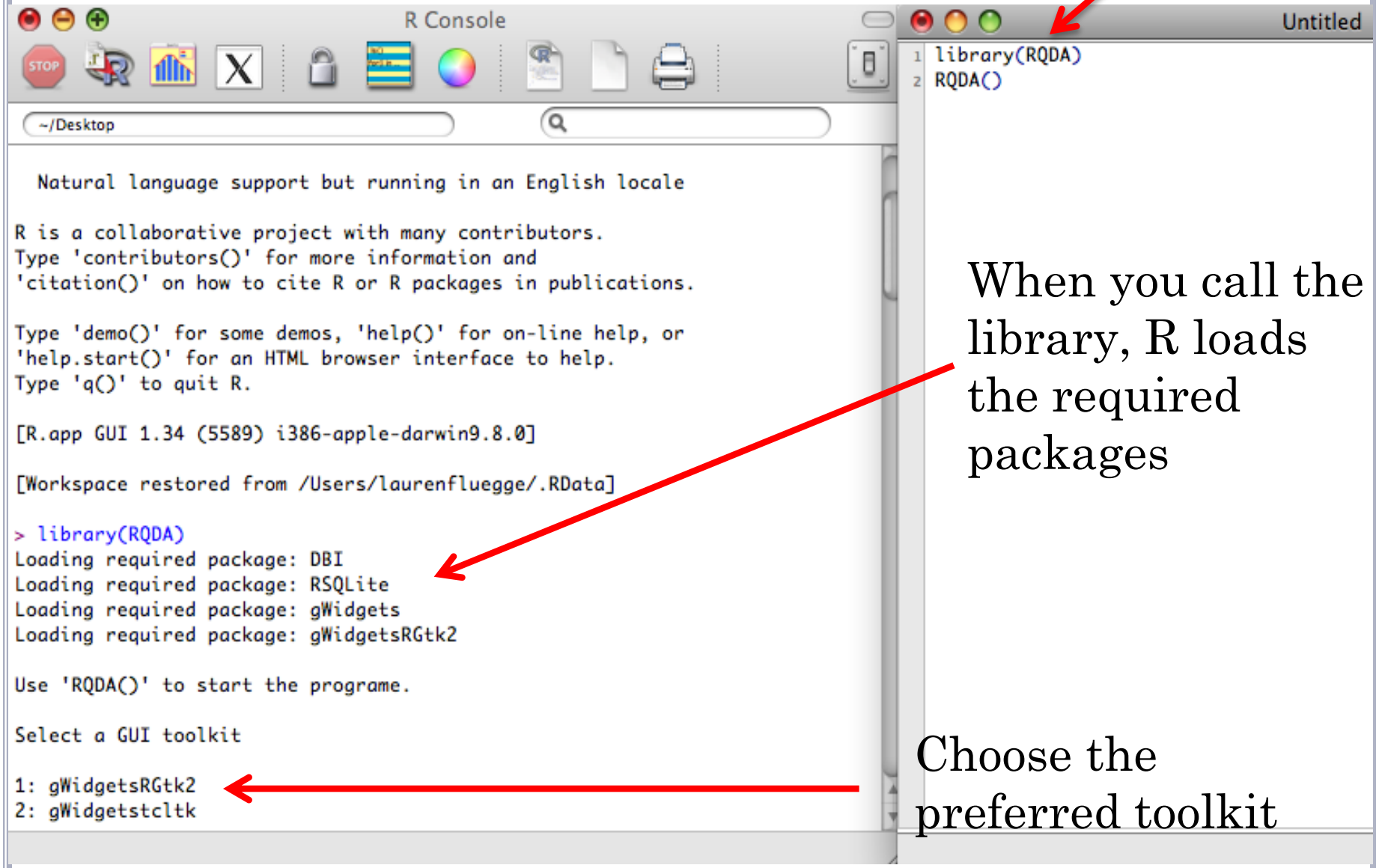
- Using R Functions
  - Retrieving codes
  - Getting coded text
  - Getting a table of codes for all files

Starting up RQDA

- Call the package using library(RQDA)
- To use GUI (graphical user interface), type RQDA()
- Choose from one of the toolkits, gWidgetsRGtk2 or gWidgetstcltk
  - Produce slightly different interfaces (preference)
  - RGtk2 has been interfacing with gWidgets longer than tcltk
  - tcltk is more appropriate for experienced users than beginners

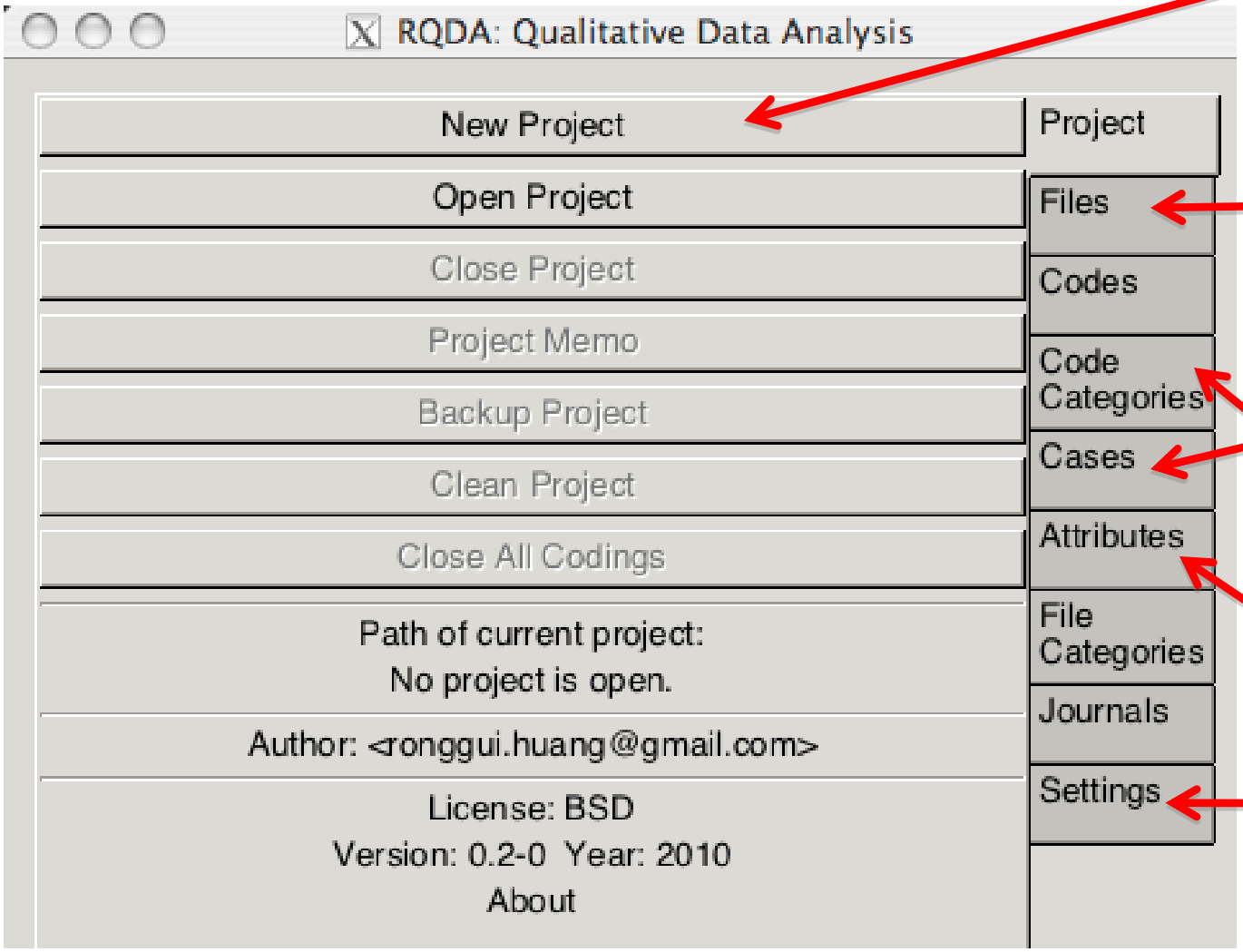
https://stat.ethz.ch/pipermail/r-sig-gui/2007-August/000706.html

R command file



When you call the library, R loads the required packages

Choose the preferred toolkit



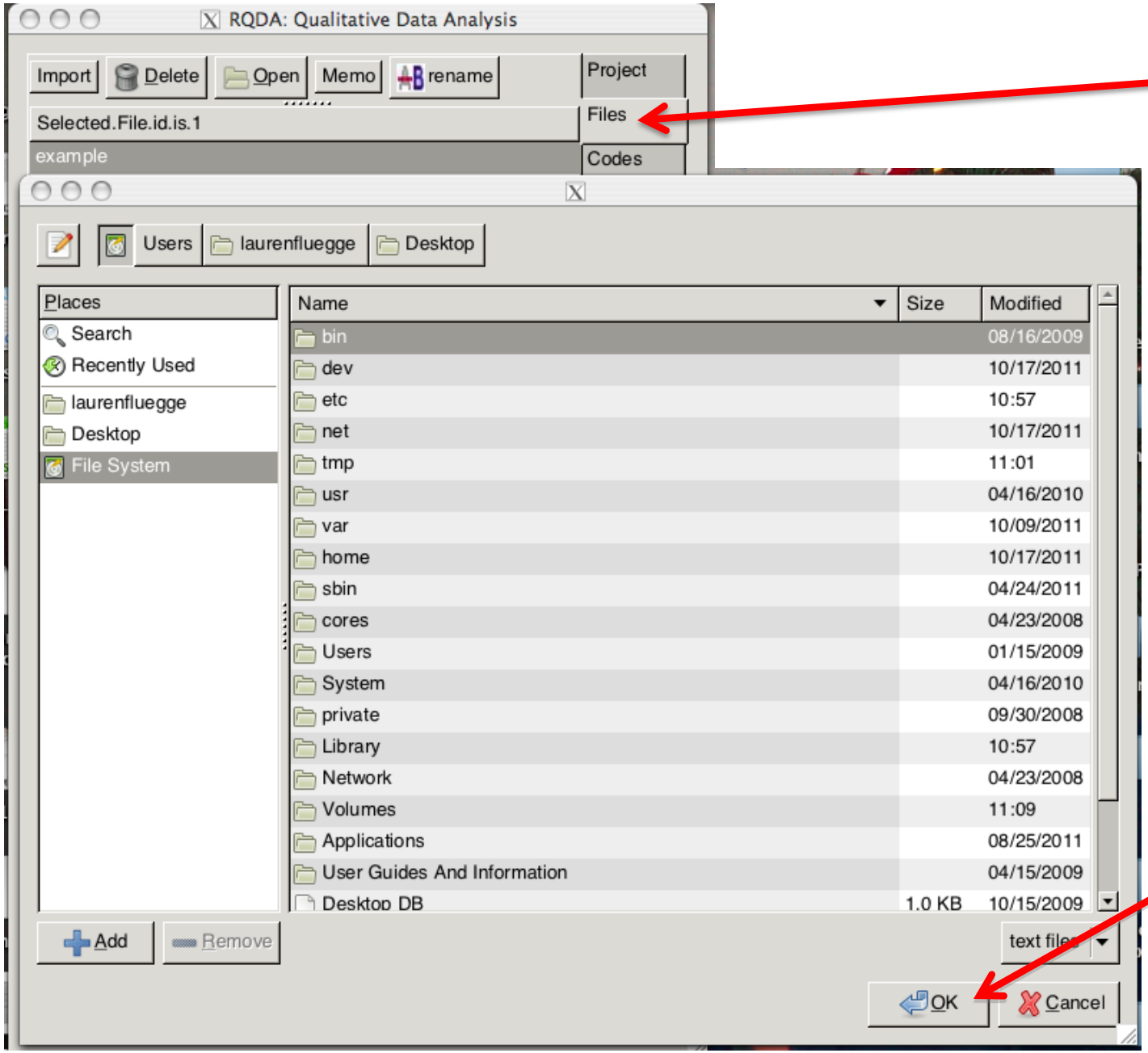
Open new or existing project

Import and manage files, create memos

Create codes and code categories

Add attributes

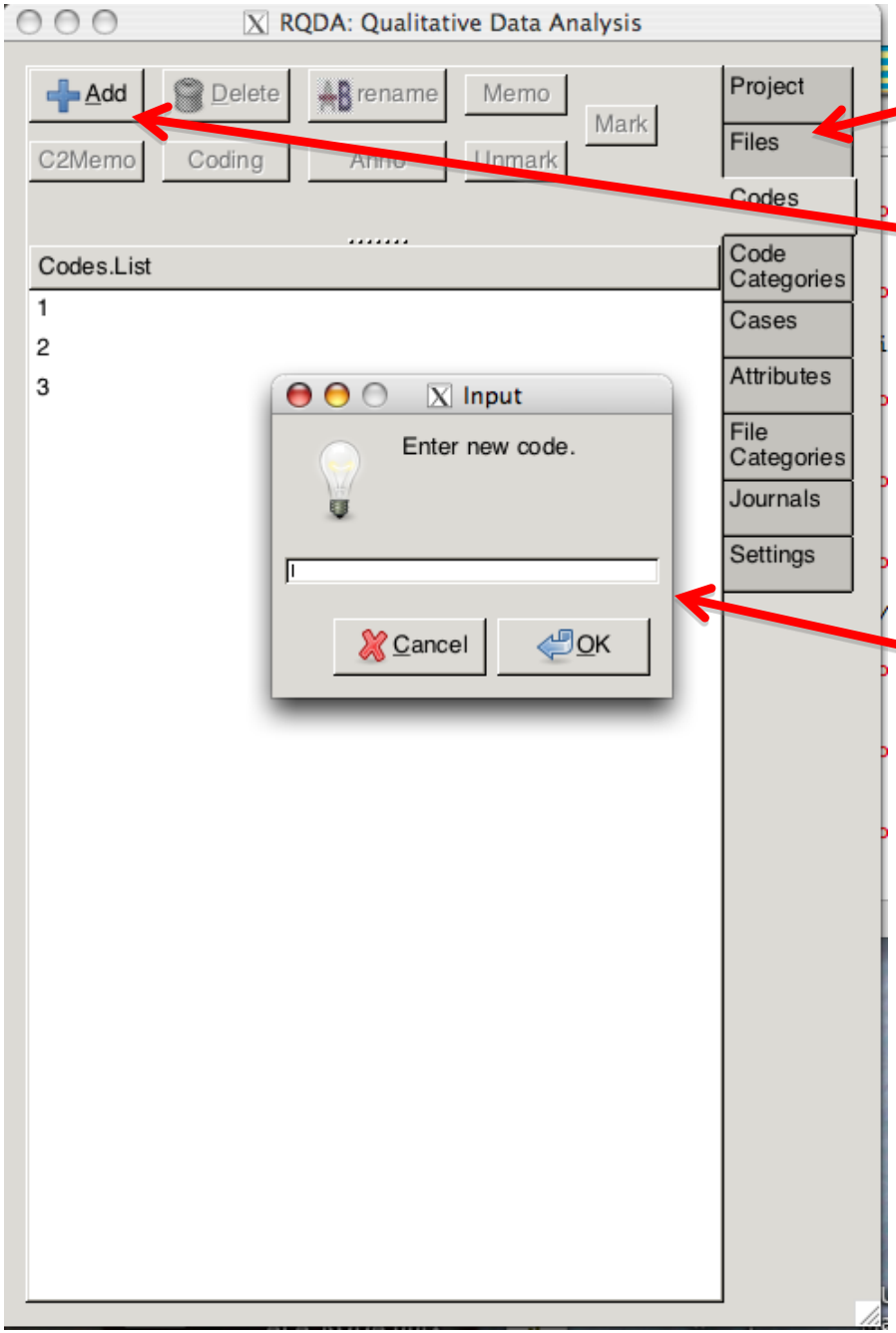
Control color of codes and cases, and coding table being managed



First, choose the "Files" tab

Then, choose "Import" to open your file

Once you've found the file, click "OK"

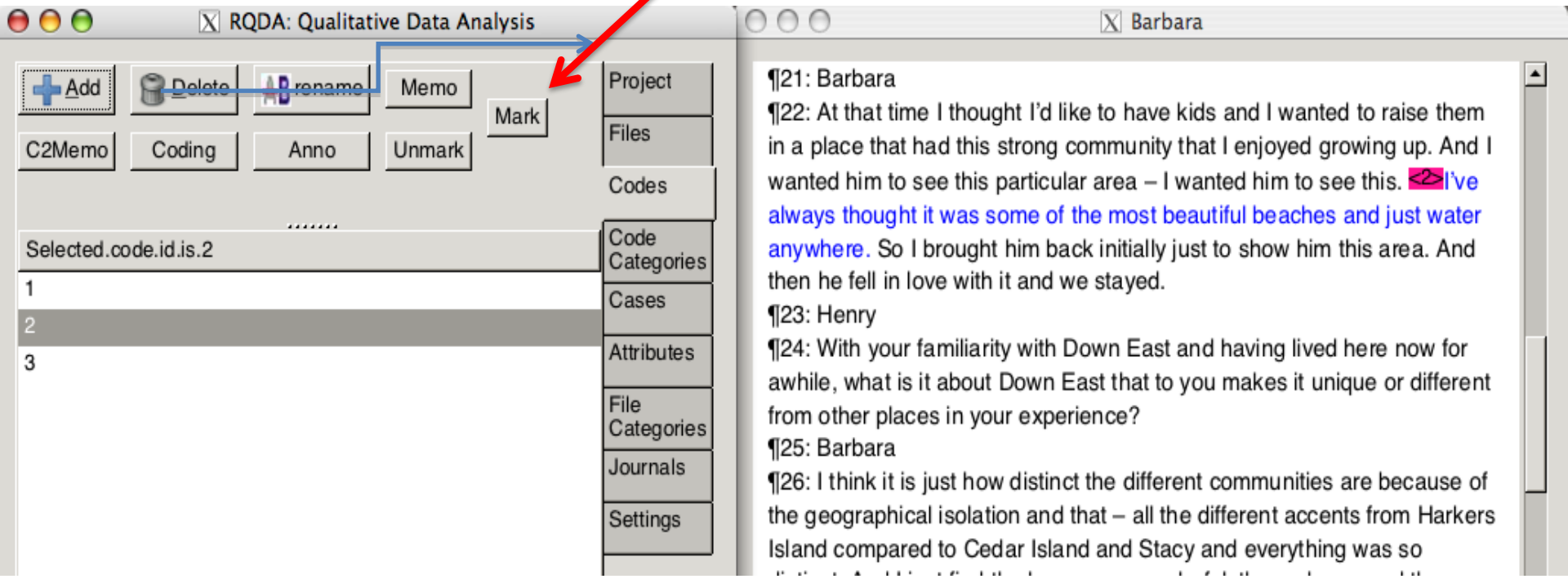


Click on "Codes"

Click on "Add" to create a code

This box will appear. Add one code at a time (it can be text or numeric, but you can also create memos to note what each code represents)

“Mark”



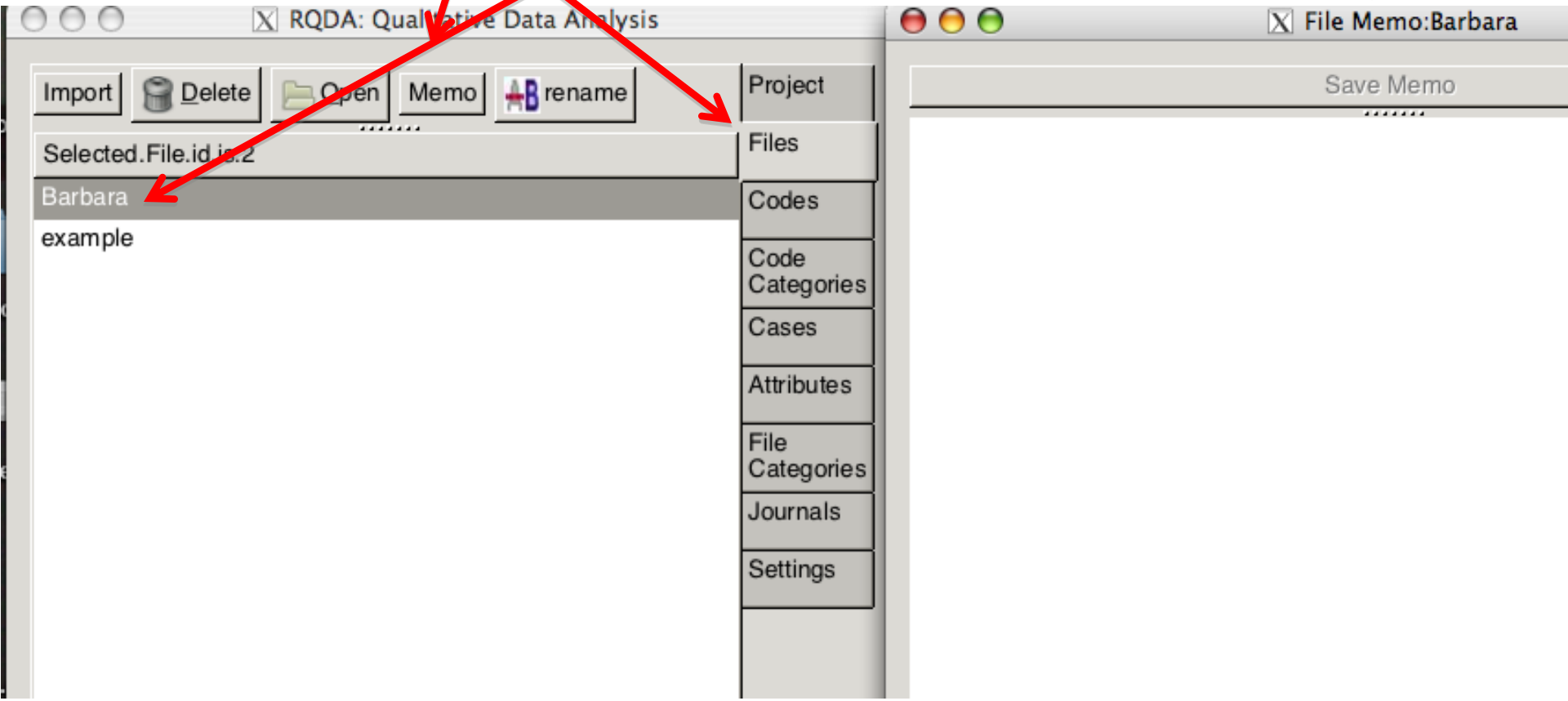
To code, simply open your document and click on the “Codes” tab. Click on the code you are using, highlight the part of the document that corresponds, and click “Mark” (you can control the color of the codes by right-clicking the code).

Click on “Files”

Click on “Memo”

Choose the file for which you want to memo

This box will appear. Type in your notes, and simply click “Save Memo” at the top.



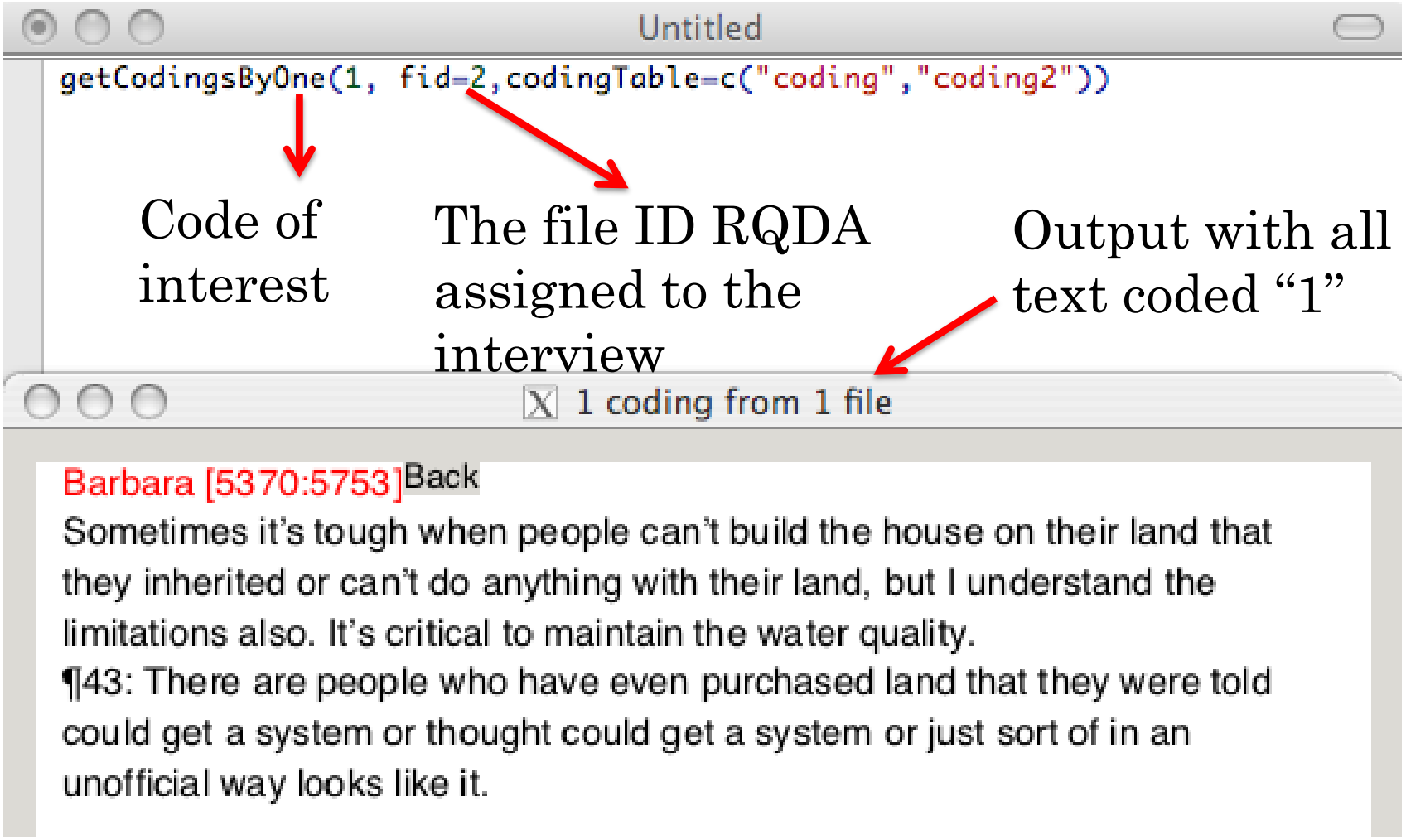
summaryCodings() provides you with all the codes used, as well as the number of codings for each code, the average number of words associated with each code, and the number of files that have that coding.

```
> d<-summaryCodings()
> d
---------------
Number of codings for each code.

1 2 3
1 3 2
---------------
Average number of words assciated with each code.

       1        2        3
383.0000 171.6667 234.0000
---------------
Number of files associated with each code.

1 2 3
1 1 1
```

R function

We are only using one file as an example, but you could have more

getCodingsByOne(cid, fid=NULL, codingTable=c("coding", "coding2")) provides you with the portions of the text that are coded with a particular code. For example, we wanted to view the portions of the interview that were coded by "1."



getCodingTable() provides the user with the total list of codes for all files in the project. The table below indicates in which file the code was found, the file name, the code name, and the number of characters for each code. This function code be helpful if you wanted to compare codings among different people.

```
> getCodingTable()
  rowid cid fid codename filename index1 index2 CodingLength
1     1   2   2        2  Barbara   3184   3305          121
2     2   2   2        2  Barbara   2240   2326           86
3     3   2   2        2  Barbara   3311   3619          308
4     4   1   2        1  Barbara   5370   5753          383
5     5   4   2        3  Barbara   8249   8456          207
6     6   4   2        3  Barbara   8577   8838          261
```

Summary of RQDA

- The GUI platform is intuitive and easy to navigate
- R provides a place to conduct both qualitative and quantitative analyses
- Though limited in number, R functions provide a quick way to retrieve coding and file information

Summary

- R is a free, flexible program that evaluators can use to analyze survey and interview data within the same software program
- Surveys can be screened, cleaned, and analyzed within the R environment
- Interviews can be coded and queried within the RQDA GUI and using a few syntax codes
- Data can be exported in a variety of formats (full data set, tables, specific codes)

Questions?

Thank You

- Lindsey Varner- l_dunn@uncg.edu
- Lauren Fluegge- lbfluegg@uncg.ed
- Korinne Chiu- k_chiu@uncg.edu