Loading Data into R

Heike Hofmann, Hadley Wickham

Loading data

- We will use .csv (comma separated values), because most software can both write and read this format
- somedataset <- read.csv(file.choose())
- Always check with str() that the file has loaded correctly

Your turn

- Open the Shangri La data in excel, save it as csv, and then load into R.
- Open the baseball data in excel, save it as csv, and then load into R.
- Check that they look OK using str()

- Advanced: Open the csv in Word. Try and break the data import, by adding odd characters (try #, , “,), read ?read.csv and figure out what’s going on.

Examining variables

- a
- head(a)
- summary(a)
- str(a)
- dim(a)

Your turn

- Download data from the Unites States Cancer Statistics at http://www.cdc.gov/cancer/npcr/uscs/2004/download_data.htm
- Unzip the archive (use Winzip, e.g.)
- Load ByArea.txt into Excel (2007)

in Excel

- Replace all ~ by NA. Are there other symbols that should be replaced by NAs?
- Delete all records for “2002-2004”, for “male and female”, and all other fields that represent sums of other fields
- Split rate, upper & lower CI, and count into two columns each according to event type
- Save as comma delimited text file (.csv)

Switch to R

- # Load data into R
  cancer <- read.csv(file.choose())
- Use
  head(cancer)
  dim(cancer)
  summary(cancer)
  to check that it worked
- Did you catch all the symbols for missing data in Excel?

Data frames

- A data frame is a list of vectors of the same length
- Create with **data.frame** using named arguments
- data.frame(a=1:10, b=c(TRUE,FALSE))
- Created by **read.csv** too

Extracting subsets

- One of the keys to mastering the R is learning to use the extraction (or subset) operators effectively.

[$

[

- By positive integers, select specified
- By negative integers, omit specified
- By logical vector, select T, omit F
- By character vector (by name)

“Sub”sets can be bigger

- a <- c(1, 5, 9)
- a[c(1,2,3)]
- a[c(1,1,1,2,2,3)]

[+ logical vectors

- The most complicated to understand, but the most powerful
- Lets you extract a subset defined by some characteristic of the data
- cancer$Site[cancer$Mortality.Rate > 100]
- cancer[cancer$Mortality.Rate > 100,]

Updating subsets

- You can take a subset and update the original data
- a <- 1:4
- a[2:3] <- 0
- a
- Very useful with logical subsetting

Practice

- Select the Race variable in three different ways
- Drop variables Age.Adjusted.Rate, Age.Adjusted.CI.Lower, and Age.Adjusted.CI.Upper from the dataset
- Replace all “~” by NA
- Replace all other symbols for missing data by NA

More about missings

- NA + x = NA, NA * x = NA
- x == NA
- **is.na** returns logical vector, for single vector
- **complete.cases** does the same for a data.frame
- Many functions have *na.rm*

Practice

- Remove all missings from the cancer data. Why might this be a problem?
- Remove all records with missing mortality rate.

Analysing the data

- What questions do we have about the data?
- Write down questions for 1 min, then get together with your neighbor and discuss.
- What data will you need to try to answer your questions? What graphics would you draw in support of your questions? Discuss again. Be ready to report.

Questions about the cancer data

Report

- Write a short report, which should include:
  - your question
  - your expectation before looking at the data
  - a graphic which answers the question
  - a conclusion based on the graphic
- Print/Email your report (don't forget to put your names on it, too)

Homework

- Pick one of the "Major Findings" from http://www.cdc.gov/cancer/npcr/uscs/2004/facts_major_findings.htm
  and find a graphic which supports this finding
- Write up a paragraph about what else the graphic also shows