

# An Introduction to R

Di Cook  
Iowa State University

Materials derived from notes by Hadley Wickham and Heike Hofmann for Statistics 480, ISU

## Outline

- An introduction to the R language
- The basics of handling data
- Producing data plots
- Simulation
- Reshaping data
- Being lazy with R
- Interactive graphics
- Statistical modeling of data
- Exploring multivariate data

# Setting up

- Open the R script in an editor, eg notepad, wordpad, emacs, ...
- Edit lines
- Cut and paste lines of code into the R interpreter window

# The R language

- Learning a new language: grammar, vocabulary
- Loading, examining, summarizing data
- Creating data
- Getting help
- Miscellaneous useful stuff



**Learning a new  
language is hard!**

## Learning a language

- Grammar / Syntax
- Vocabulary
- “Thinking in that language”

# Grammar

*Like mathematics*

- Basic algebra is the same
  - but  $2*x$  not  $2x$ ,  $2^p$  instead of  $2^P$
- Applying a function is similar
- Making a variable, use  $<-$  instead of  $=$
- Everything in R is a vector
  - Index a vector using  $[ ]$

# Examples

- $x = 2 / 3$
- $\sqrt{x}$
- $a = 2(x + 3)^2$
- $y = (1\ 2\ 3\ 5)^T$
- $y_1$
- $\sum y$
- $2y$
- $f(y, 2) = 2y$

# You try

- $x = (4 \ 1 \ 3 \ 9)^T$
- $y = (1 \ 2 \ 3 \ 5)^T$  (from examples, on previous slide)
- $d = \sqrt{x^2 + y^2}$
- $2(d_1 + d_4)$

# Vocabulary

- What verbs (=functions) do you need to know?
  - Loading data
  - Accessing parts of things
  - Statistical summaries
  - ...

# Loading data

- Import data with:
  - `read.csv()` for csv files
  - (and use `file.choose()` to help find your file)
  - Save from excel as csv files
  - Stored in a `data.frame`
    - a list of variables with the same length
- Check the data with:
  - `head(x)`

## Your turn

- Save shangri-la data as csv
- There are two sheets, so you will need one file for each sheet
- Load it into R  
(use `sla <- read.csv(file.choose())`)
- Use `head(sla)` to check it worked

# Examining variables

- `x`
- `head(x)`
- `summary(x)`
- `str(x)`
- `dim(x)`

# What do we have?

- A `data.frame` = a list of variables of the same length (but may be different types)
- Has row and column names

# Extracting parts

- `x$variable`
- `x[,"variable"]`
- `x[rows, columns]`
  - `x[1:5, 2:3]`
  - `x[c(1,5,6), c("age","height")]`
- `x$variable[rows]`

# Statistical summaries

- mean, median, min, max, range
- sd, var, cor
- summary

# Your turn

- Look at first 10 weight date records
- Compute the mean of weight change
- Compute the standard deviation of the weight change
- Compute correlation between days and weight change

# Random numbers

- Uniform: `runif(n, min, max)`
- Gaussian: `rnorm(n, mean, sd)`
- Exponential: `rexp(n, rate)`
- Poisson: `rpois(n, lambda)`

# Your turn

- Calculate the mean and sd for a sample of size 10 from a standard normal distribution.
- Calculate the mean and sd for a sample of size 100 from a standard normal distribution.
- Calculate the mean and sd for a sample of size 100 from a normal distribution with mean -8 and sd 4.
- Compute correlation between a sample of size 30 from a standard normal and a sample of size 30 from a uniform distribution.

# Getting information

- `?runif`
- `help(package="ggplot2")`
- `?"*"`

# Your turn

- Look at the documentation for the “mean” function.
- How would you write the command to calculate the mean for a variable that has missing values?
- Read the person sheet from the shangri-la data into R. Use you mean command to calculate the mean age, where missing values are ignored.

## Navigating the R interpreter window

- Up/down arrow keys to retrieve previous lines
- Left/right arrow keys to move cursor along line
- Mouse click to set cursor position
- Delete to remove and re-type parts of command

# Listing R objects

- `ls()`
- `ls(pos=2)`
- `search()`

# Generating sequences

- Equispaced;
  - `3:10`
  - `seq(3, 10)`
  - `seq(3, 10, by=1/3)`
  - `seq(3, 10, len=8)`
- Repeats:
  - `rep(3, 10)`
  - `rep(1:3, 5)`
  - `rep(c(1,2,3,4), c(2,3,2,3))`
  - `rep(1:3, rep(2,3))`

# Common functions

- abs, sign
- sqrt, exp, log, log10
- floor, ceiling, trunc, round, signif
- cos, sin, tan, acos, asin, atan
- modulus arithmetic
- strsplit
- Sys.time(), Sys.Date()

# Options

- options(digits=3)
- options(width=20)

# Writing your own functions

- Standardize all the columns of a matrix
- $f(x) = (x-m)/s$
- Apply to all the columns

# Quitting R

q()