Lab 1

Lab 1

Kyle N. Payne8/8/2015

Getting Started

Lab Time : Wed 12-1:50pm

Course Website : compass2g.illinois.edu

TA Name : Kyle N. Payne, (Kyle, K-Pax, "Hey... man")

TA Office Hours : Mon 8:30am-9:30am

TA Email : knpayne2@illinois.edu

Book

Title : An Introduction to Statistical Methods and Data Analysis

Author(s) : R Lyman Ott, Michael Longnecker

Edition : 6th

Can be found on most retailers like Amazon.com, abebooks.com, etc...

Also can be found by google-ing Title + 'pdf'

Software

Software is a huge component of statistical analysis

We will support SAS and R usage for this course

SAS is a suite of tools for statistical analysis

R is a programming language primarily for statistical analysis

SAS

SAS inc. is the worlds largest privately held software company.

Used by most fortune 500 companies http://archive.fortune.com/2010/01/21/technology/sas_best_companies.fo

SAS Software has over 200 components

Can perform many statistical analyses with limited effort

Presents a robust suite of industry standard tools. -https://en.wikipedia.org/wiki/SAS_(software)#cite_note-encycl-2

R

"Much closer to the metal" than SAS

Can be used to create your own functions, libraries, etc...

Requires more effort

But this results in arguably better skill development

Accessing SAS

N120

Many campus computer labs

Remotely through ACES

SAS OnDemand

Accessing R

Rstudio : https://www.rstudio.com/

On many workstation on campus

Can be downloaded on most computers

Very lightweight software

Reading Data SAS — Inline

Reading Data SAS — Inline

Reading Data SAS — Inline

Reading Data SAS — Inline

Reading Data R —- Inline

Reading Data R —- Inline

Reading Data SAS —- CSV

Reading Data SAS —- CSV

Let's try this example!

But first note,

$ in SAS means a character variable

By default, SAS assumes that a variable is numeric

If SAS encodes a variable as numeric, but encounters a character it will code is as ., which means missing.

Reading Data R —- CSV

Or just use the "import dataset" button in Rstudio

Reading Data R —- CSV

Let's try this example!

Basic Descriptive Statistics —- SAS

proc univariate

proc means

proc univariate —- SAS

proc means —- SAS

Basic Descriptive Statistics —- R

Basic Descriptive Statistics —- R

Basic Descriptive Statistics —- R

Basic Descriptive Statistics —- R

Basic Descriptive Statistics —- R

sapply is extremely useful for working with data in R. In order to understand sapply better we first need to understand the some basic data structures.

Data Structures —- R

R's base data structures can be organised by their dimensionality (1d, 2d, or nd) and whether they're homogeneous (all contents must be of the same type) or heterogeneous (the contents can be of different types). This gives rise to the five data types most often used in data analysis:

Homogeneous

Heterogeneous

1d

Atomic vector

List

2d

Matrix

Data frame

nd

Array

("Advanced R", Wickham 2014) http://adv-r.had.co.nz/

Data Structures —- R —- Vectors

1 dimensional data structures

Must all be one type i.e. "homogeneous"

Formed by the "c()" operator e.g.

Data Structures —- R —- Vectors

What happens with mixed types?

Data Structures —- R —- Lists

1 dimensional data structures

Can be mixed types i.e. “heterogeneous”

Formed by the “list()” function

Data Structures —- R —- Lists

Mixed Types?

Data Structures —- R —- data.frames

What are lists that contain vectors? Data Frames.

R structures data files e.g. .csv, .txt, etc... as data frames

These come with a bunch of useful functionality

You will see these alot.

Subsetting —- SAS —- Variables

keep

Subsetting —- SAS —- Observations

if

Subsetting —- R —- Variables

MouseID

DYRK1A_N

ITSN1_N

18899_1 : 1

Min. :0.1453

Min. :0.2454

18899_10: 1

1st Qu.:0.2881

1st Qu.:0.4734

18899_11: 1

Median :0.3664

Median :0.5658

18899_12: 1

Mean :0.4258

Mean :0.6171

18899_13: 1

3rd Qu.:0.4877

3rd Qu.:0.6980

18899_14: 1

Max. :2.5164

Max. :2.6027

(Other) :1074

NA's :3

NA's :3

Subsetting —- R —- Observations

Subsetting —- R —- Observations

MouseID

DYRK1A_N

ITSN1_N

1

309_1

0.5036439

0.7471932

2

309_2

0.5146171

0.6890635

3

309_3

0.5091831

0.7302468

16

311_1

0.7431179

0.8626527

17

311_2

0.7114799

0.8070539

18

311_3

0.7046332

0.8025372

Subsetting —- R

In general, one can select rows of a dataframe by subsetting on l.h.s. of [,]'s

one can select columns by subsetting on r.h.s. of [,]

We can also use the {r, eval=FALSE} subset() function

Activity

Using either R or SAS:

Import the Mouse cortex dataset

Select the variables H3MeK4_N, TIAM1_N, Ubiquitin_N, Genotype, MouseID

Remove missing values from the dataset

Calculate the means of all of the numeric variables

Basic Statistical Graphics

Visualization is hugely important

Histogram, Boxplot

Basic Statistical Graphics —- Histogram

Basic Statistical Graphics —- Boxplot

Activity

Using either R or SAS:

Take the variable H3MeK4_N

Make a histogram and boxplot

Getting help with software

Google.com

Stackoverflow.com

SAS documentation (at your own risk!)

R documentation

TAs