Statuatory Warning

Will Speak too fast and mumble words

Please ask Siva to repeat if you do not understand

R

- R is an open-source compute programming language and runs on Linux, Windows, and Mac-OS.
- R is FREE.
- The R project web page
  http://www.r-project.org

R

- R is modeled after S and S-Plus. The S language was developed in the late 1980s at AT & T labs.
- The R project was started by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland in 1995. [*Journal of Computational and Graphical Statistics,5:3, pp. 299-314. 1996.*]
- R is now a collaborative project with many contributors and is maintained by the R core-development team.

Installing R

- To download `R` visit https://cloud.r-project.org

- Rstudio is an Integrated Development Environment, or IDE for R. To download `Rstudio` visit http://www.rstudio.com/download

Getting Started on R– as a calculator

You can try the following commands:

```
> 9 / 44
> 0.6 * 0.4 + 0.3 * 0.6
> log(0.6 * 0.4 + 0.3 * 0.6)
```

Getting Started on R– as a calculator

Your output should look like:

```
> 9 / 44
[1] 0.2045455
> 0.6 * 0.4 + 0.3 * 0.6
[1] 0.42
> log(0.6 * 0.4 + 0.3 * 0.6)
[1] -0.8675006
```

`[1]` at the beginning of each answer is there for a good reason.

Any data is stored in R as a *vector*. `[1]` represents the position of that element in the vector.

R- c function

Suppose we wish to enter ScoresinDS of 10 students.

40, 39, 15, 6, 18, 22, 30, 21, 15, 23

```
> ScoresinDS= c(40, 39, 15, 6, 18, 22, 30, 21, 15, 23)
```

R- c function

The output should be

```
> ScoresinDS = c(40, 39, 15, 6, 18, 22, 30, 21, 15, 23)
```

In the above, we have assigned values to a variable called `ScoresinDS`.

The assignment operator is `=`.

The values do not get displayed automatically unless we call it with `ScoresinDS` as below.

```
> ScoresinDS
 [1] 40 39 15  6 18 22 30 21 15 23
```

R-inbuilt functions

R has many in-built functions

```
> meancomputation = (40+ 39+ 15+ 6+ 18+ 22+ 30+ 21+ 15+ 23)/10
> meancomputation
```

```
[1] 22.9
```

```
> meaninbuilt = mean(ScoresinDS)
> meaninbuilt
```

```
[1] 22.9
```

R-Slicing

Changing one element of ScoresinDS: Suppose we want to change the entry of student 4 from 6 to 16.

```
> ScoresinDS

 [1] 40 39 15  6 18 22 30 21 15 23

> ScoresinDS2 = ScoresinDS # create a copy of ScoresinDS
> ScoresinDS2[4] = 16
> ScoresinDS2

 [1] 40 39 15 16 18 22 30 21 15 23

> ScoresinDS[c(1,3,5)]

[1] 40 15 18
```

R- Logical Operators ==, <=, >=, <, >

Selecting few elements of ScoresinDS: Suppose we want to see which students ScoresinDS are equal to 30 marks

```
> ScoresinDS

 [1] 40 39 15  6 18 22 30 21 15 23

> y = which(ScoresinDS == 30)
> y

[1] 7
```

or those with ScoresinDS lesser than or equal to 20.

```
> z = which(ScoresinDS <= 20)
> z

[1] 3 4 5 9
```

R- Creating sequence of vectors

```
> x = 1:100
> x

  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100

> x[x < 10 | x > 90]

 [1]   1   2   3   4   5   6   7   8   9  91  92  93  94  95  96  97  98  99 100
```

Data

- Data and its analysis has a rich and wide literature.
- Three kinds of Data:
  - Categorical Data
  - Discrete Numeric Data
  - Continuous Numeric Data
- *On the Theory of Scales of Measurement* By S. S. Stevens Science 07 Jun 1946: Vol. 103, Issue 2684, pp. 677-680, gave a broad classification of data from measurements into 9 categories.

Discrete Numerical Data

- Many data are described in terms of numbers.
- Many variables naturally take on only discrete values.
- Boxplot and Histograms are used to visualise such data.

Discrete Numerical Data: Key features

- Center
- Spread
- Shape

Discrete Numerical Data: Key features

- Center Widely used measure of centre is the `mean` or the average of the data set. Other measures include the `median` and the `mode`
- Spread Understanding variabiity of the given data is very important. If one were to understand `mean` as specifying the center then the range of the data set around it is determined by its variability or spread. It is often measured by the variance(`var`) or standard deviation (`sd`) or the inter-quartile range (`IQR`).
- Shape To understand various distributional aspects of the dataset one needs to understand its "shape". For e.g. if it is symmetric or skewed around its mean. Other aspects include among the data points which are more likely than others.

Datasets in R

R has a lot inbuilt Datasets that one can use. The command :

```
> data()
```

will list currently installed data sets.