

Foundations of R Software

Lecture 18

Basics of Calculations

⋮

Missing Data Handling

Shalabh

Department of Mathematics and Statistics

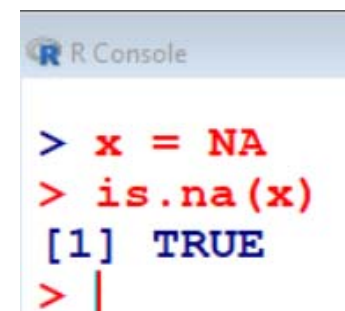
Indian Institute of Technology Kanpur

Missing data

R represents missing observations through the data value **NA**

We can detect missing values using **is.na**

The command **is.na()** returns a logical vector with **TRUE** in the element locations that contain missing values represented by **NA**.



```
R Console
> x = NA
> is.na(x)
[1] TRUE
> |
```

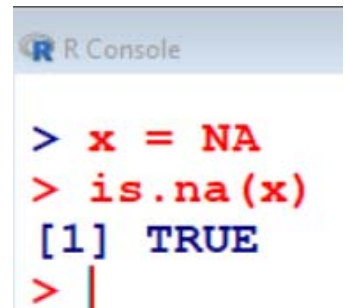
Missing data

`is.na()` will work on vectors, lists, matrices, and data frames.

```
> x = NA    # assign NA to variable x
```

```
> is.na(x)  # is it missing?
```

```
[1] TRUE
```

A screenshot of an R console window. The title bar says "R Console". The console shows the following commands and output:

```
> x = NA
> is.na(x)
[1] TRUE
> |
```

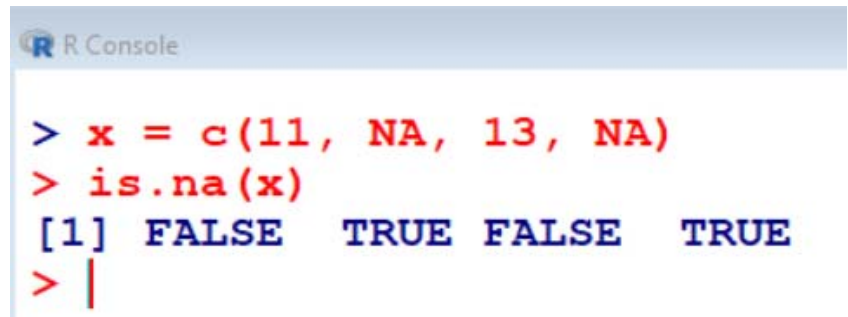
Missing data

Now try a vector to know if any value is missing?

```
> x = c(11, NA, 13, NA)
```

```
> is.na(x)
```

```
[1] FALSE TRUE FALSE TRUE
```

A screenshot of an R console window. The title bar says "R Console". The console shows the following text: a red prompt character ">" followed by the command "x = c(11, NA, 13, NA)" in red; another red prompt character ">" followed by the command "is.na(x)" in red; and the output "[1] FALSE TRUE FALSE TRUE" in blue. A red prompt character ">" followed by a vertical bar "|" is on the next line, indicating the cursor is ready for more input.

```
R Console  
  
> x = c(11, NA, 13, NA)  
> is.na(x)  
[1] FALSE TRUE FALSE TRUE  
> |
```

Example : How to work with missing data

```
> x = c(11, NA, 13, NA) # vector
```

```
> mean(x)  $\frac{11 + NA + 13 + NA}{4}$   
[1] NA
```

```
> mean(x, na.rm = TRUE) # NAs can be removed  
[1] 12  $\frac{11 + 13}{2} = 12$ 
```

```
R Console  
  
> x = c(11, NA, 13, NA)  
> mean(x)  
[1] NA  
> mean(x, na.rm = TRUE)  
[1] 12  
> |
```

NA versus NULL

The null object, called **NULL**, is returned by some functions and expressions.

Note that **NA** and **NULL** are not the same.

NA is a placeholder for something that exists but is missing.

NULL stands for something that never existed at all.

Missing data: Location of missing values

To identify the location of **NAs**, use **which()** function as

which(is.na())

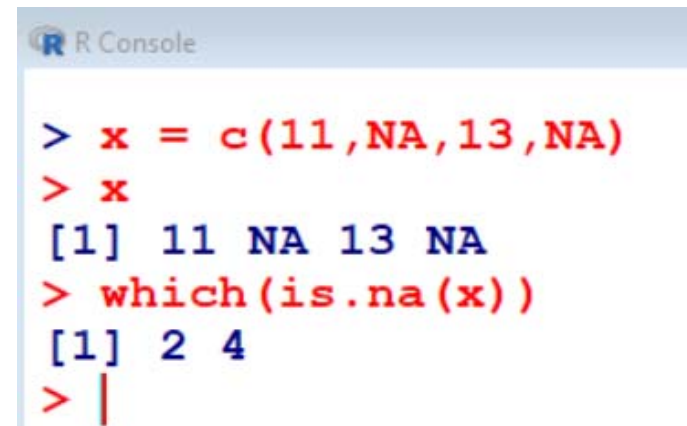
```
> x = c(11, NA, 13, NA)
```

```
> x
```

```
[1] 11 NA 13 NA
```

```
> which(is.na(x))
```

```
[1] 2 4
```



```
R Console
> x = c(11, NA, 13, NA)
> x
[1] 11 NA 13 NA
> which(is.na(x))
[1] 2 4
> |
```

It indicates that the missing values occur at 2nd and 4th places.

Missing data

To count of NAs the number of **NAs** , use **sum()** function as

sum(is.na())

```
> x = c(11, NA, 13, NA)
```

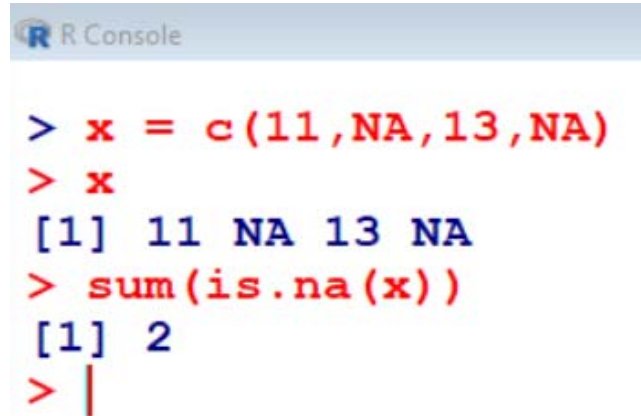
```
> x
```

```
[1] 11 NA 13 NA
```

```
> sum(is.na(x))
```

```
[1] 2
```

It indicates that there are 2 missing values.



```
R Console  
  
> x = c(11, NA, 13, NA)  
> x  
[1] 11 NA 13 NA  
> sum(is.na(x))  
[1] 2  
> |
```

Missing data: Finding complete cases

To find complete cases, use `complete.cases()` function which returns a logical vector identifying rows which are complete cases.

```
> x = c(11, NA, 13, NA)
```

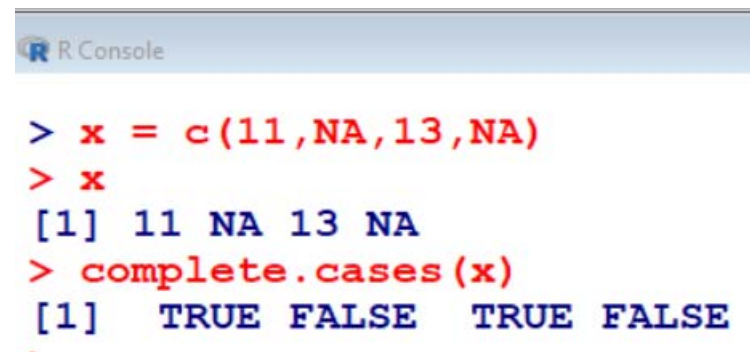
```
> x
```

```
[1] 11 NA 13 NA
```

```
> complete.cases(x)
```

```
[1] TRUE FALSE TRUE FALSE
```

It indicates that the values at 1st and 3rd places are not missing.

A screenshot of an R console window showing the execution of R code. The code defines a vector 'x' with values 11, NA, 13, and NA. It then prints 'x' and uses the 'complete.cases()' function to check for missing values. The output shows the vector 'x' and a logical vector where TRUE indicates non-missing values and FALSE indicates missing values (NA).

```
R Console
> x = c(11, NA, 13, NA)
> x
[1] 11 NA 13 NA
> complete.cases(x)
[1] TRUE FALSE TRUE FALSE
```

Missing data: Finding complete data set

The function `na.omit()` returns the object with listwise deletion of missing values. Drop out any rows with missing values anywhere in them and forgets them forever.

```
> x = c(11, NA, 13, NA)
```

```
> x
```

```
[1] 11 NA 13 NA
```

```
> y = na.omit(x)
```

```
> y
```

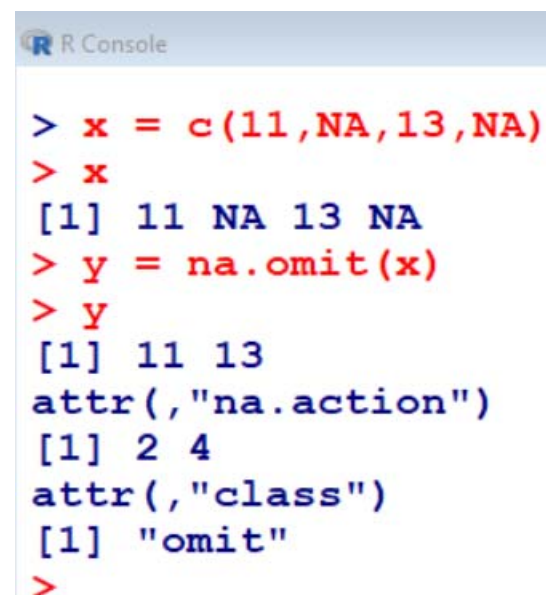
```
[1] 11 13
```

```
attr(,"na.action")
```

```
[1] 2 4
```

```
attr(,"class")
```

```
[1] "omit"
```



```
R Console
> x = c(11, NA, 13, NA)
> x
[1] 11 NA 13 NA
> y = na.omit(x)
> y
[1] 11 13
attr(,"na.action")
[1] 2 4
attr(,"class")
[1] "omit"
>
```

Missing data: Handling values

```
> x = c(11, NA, 13, NA)
```

```
> y = na.omit(x)
```

```
> y
```

```
[1] 11 13
```

```
attr(,"na.action")
```

```
[1] 2 4
```

```
attr(,"class")
```

```
[1] "omit"
```

```
> mean(x)
```

```
[1] NA
```

```
> mean(y)
```

```
[1] 12
```

R Console

```
> x = c(11, NA, 13, NA)
```

```
> x
```

```
[1] 11 NA 13 NA
```

```
> y = na.omit(x)
```

```
> y
```

```
[1] 11 13
```

```
attr(,"na.action")
```

```
[1] 2 4
```

```
attr(,"class")
```

```
[1] "omit"
```

```
>
```

```
> mean(x)
```

```
[1] NA
```

```
> mean(y)
```

```
[1] 12
```

```
> |
```