Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 29

Coefficient of Variation and Boxplots

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Coefficient of Variation (CV)

The Coefficient of Variation (CV) measures the variability of a data set without reference to the scale or units of the data.

Useful in comparing the results from two different surveys or tests in which the values are collected on different scales.

Suppose there are two data sets with

- sample means $\bar{x}_1$ and $\bar{x}_2$
- standard errors s_1 and s_2

How to compare the two data sets?

Coefficient of Variation (CV)

The sample based coefficient of variation measure of variation which uses both the arithmetic mean and standard deviation.

$$\text{C V} = \frac{s}{\bar{x}}$$

It is properly defined only when $\bar{x} > 0$.

The data with higher CV is said to be more variable than the other.

Coefficient of Variation (CV)

For example, suppose two experimenters measure the heights of same group of children in meters and centimetres (cms.).

Experimenter	Average height	Standard deviation	CV
First	$\bar{x}_1$= 1.50 meters	s_1 = 0.3 meters	CV_1 =0.3/1.50 = 0.2
Second	$\bar{x}_2$= 150 cms.	s_2 = 30 cms.	CV_2 = 30/150 = 0.2.

Both answers are the same.

How to report it correctly?

Apparently, s_1 appears to be much smaller than s_2.

Coefficient of Variation (CV)

The CV helps in comparing data sets on two completely different measurements. These variables are measured in different scales but their dimensionless CV enables the comparison of the variation of these variables.

Example: Rents of houses in a metro city and in a village.

Example: Rents of houses in Mumbai (in INR) and rent of houses in London (in Pound).

How to compare?

CV helps.

Variance

Decision Making

The data set having higher value of coefficient of variation (CV) has more variability.

The data set with lower value of cv is preferable.

If we have two data sets and suppose their coefficients of variations are CV_1 and CV_2.

If $CV_1 > CV_2$ then the data in CV_1 is said to have more variability (or less concentration) around mean than the data in CV_2.

Coefficient of Variation (CV)

R command:

Data vector: `x`

```
sqrt(var(x))/mean(x)
```

If `x` has missing values as NA, say `xna`, then R command is

```
sqrt(var(xna, na.rm = TRUE))/mean(xna, na.rm =
TRUE)
```

Note:

Similar definition can be defined for grouped data.

Coefficient of Variation (CV)

Example: Ungrouped data

Following are the time taken (in seconds) by 20 participants in a race: 32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> sqrt(var(time))/mean(time)

[1] 0.3005991
```

Coefficient of Variation (CV)

Example: Ungrouped data - Handling missing values

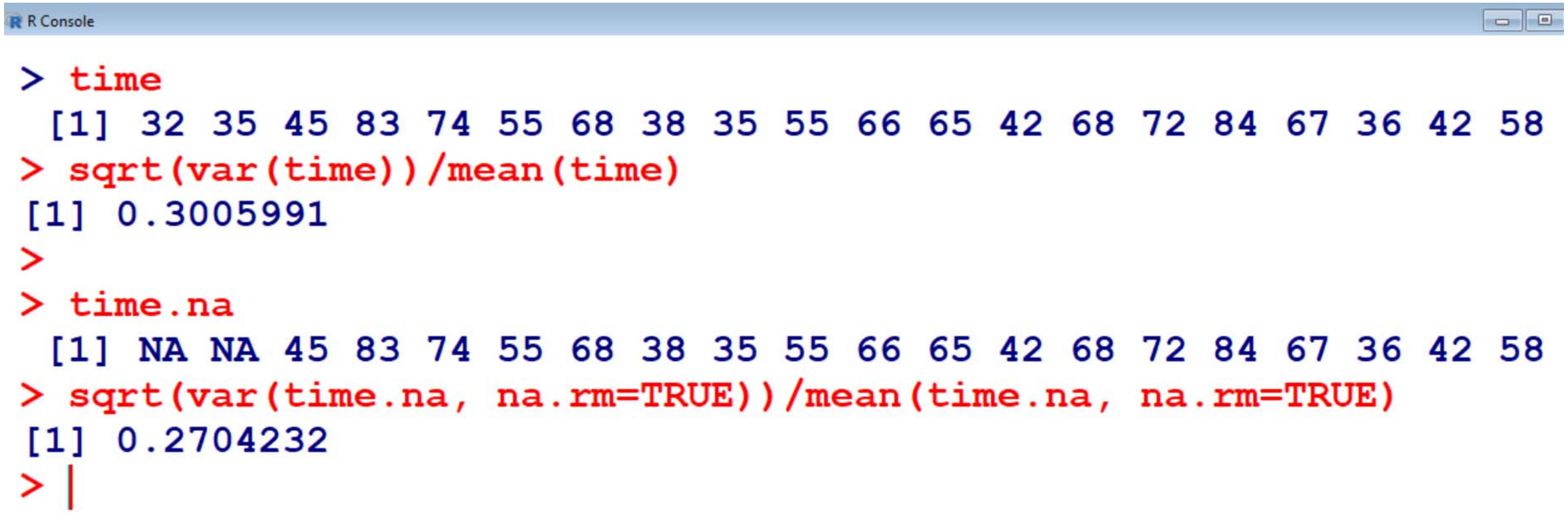
Suppose two data points are missing in the earlier example where the time taken (in seconds) by 20 participants in a race. They are recorded as NA

NA, NA, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time.na
 [1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68
72 84 67 36 42 58

>  sqrt(var(time.na,  na.rm=TRUE))/mean(time.na,
na.rm=TRUE)
[1] 0.2704232
```

Coefficient of Variation (CV)

Example: Ungrouped data



```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> sqrt(var(time))/mean(time)
[1] 0.3005991
>
> time.na
 [1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> sqrt(var(time.na, na.rm=TRUE))/mean(time.na, na.rm=TRUE)
[1] 0.2704232
> |
```

Summary of Observations

In R, quartiles, mean, minimum and maximum values can be easily obtained by the `summary` command.

```
x: data vector
summary(x)
```

It gives information on

- ❖ minimum,
- ❖ maximum,
- ❖ mean,
- ❖ first quartile,
- ❖ second quartile (median) and
- ❖ third quartile.

Summary of Observations

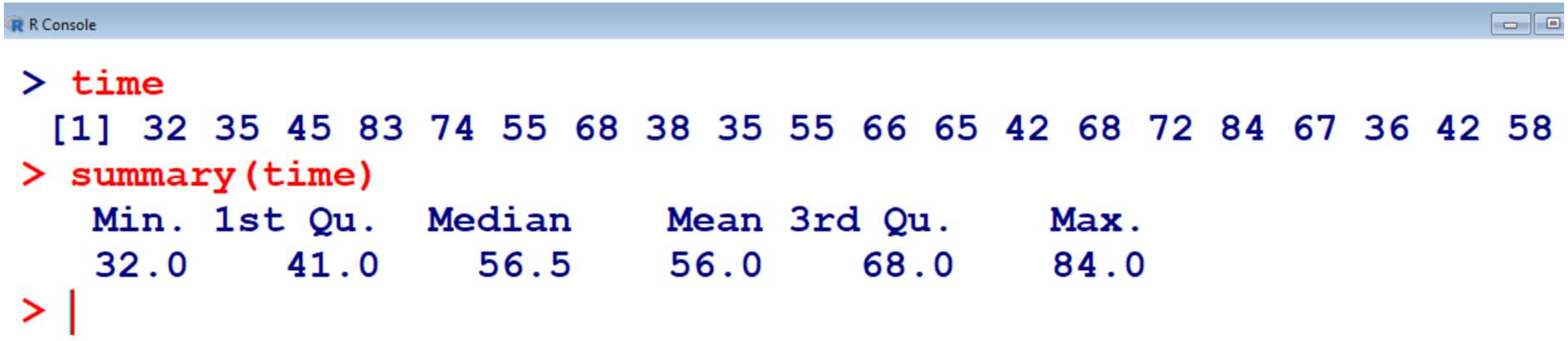
Example:

Following are the time taken (in seconds) by 20 participants in a race: 32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> summary(time)
 Min. 1st Qu.  Median   Mean   3rd Qu.  Max.
 32.0   41.0     56.5   56.0     68.0   84.0
```

Summary of Observations

Example:

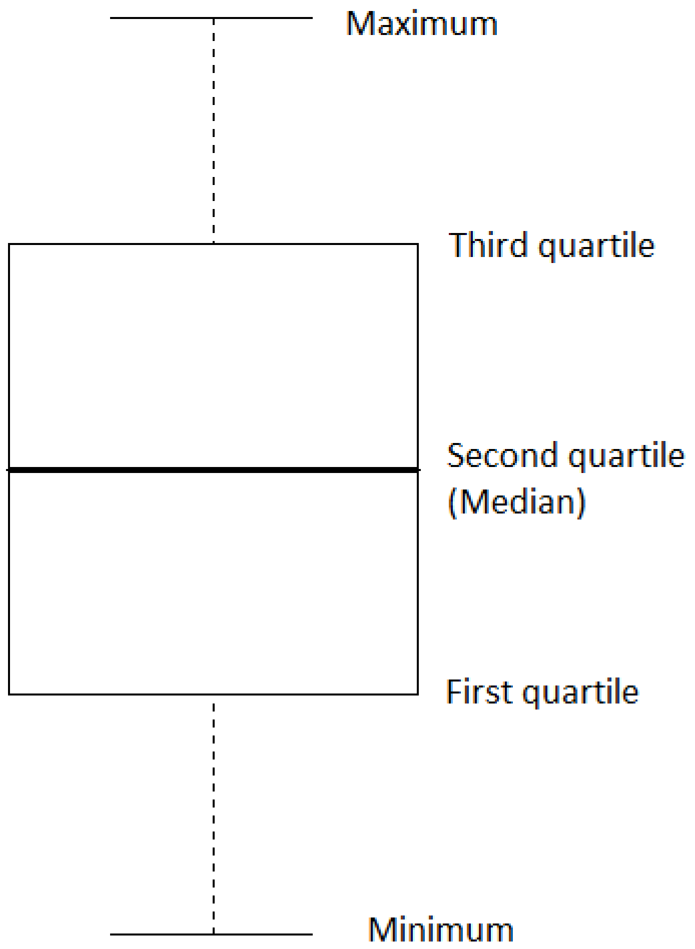


```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> summary(time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   32.0    41.0    56.5    56.0    68.0    84.0
> |
```

Boxplot

Box plot is a graph which summarizes the distribution of a variable by using its median, quartiles, minimum and maximum values.

Useful in comparing different datasets.



Boxplot

R Command:

`boxplot()`draws a box plot.

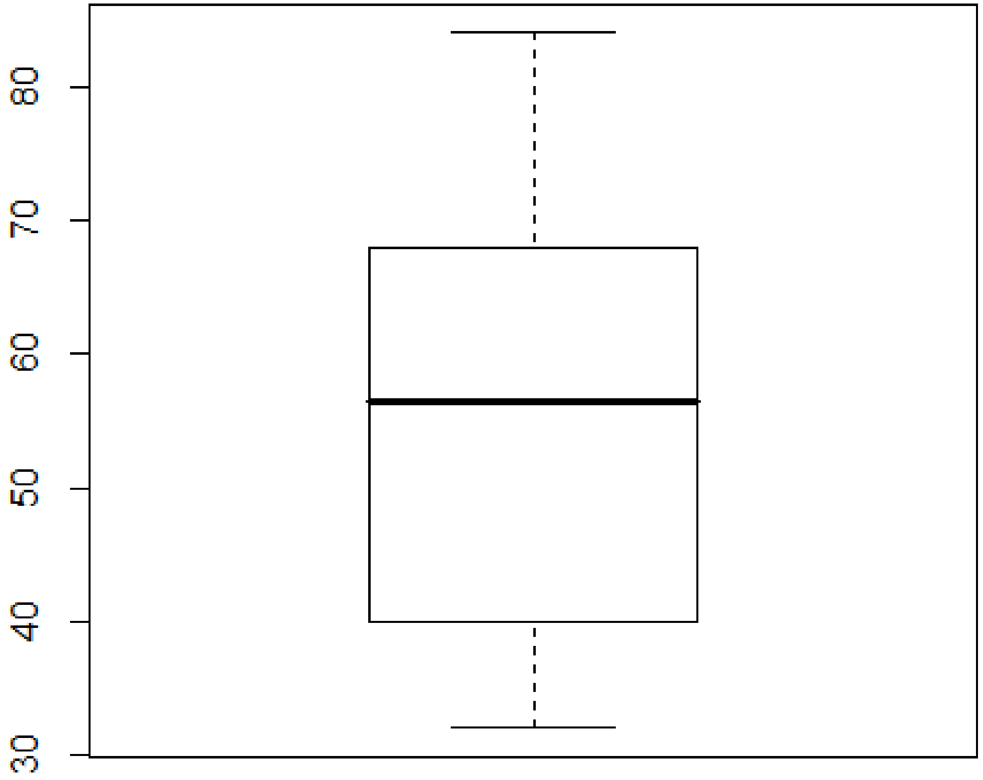
Various options are available which can be given inside the arguments.

See help on `boxplot`.

Boxplot

Example

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> boxplot(time)
```



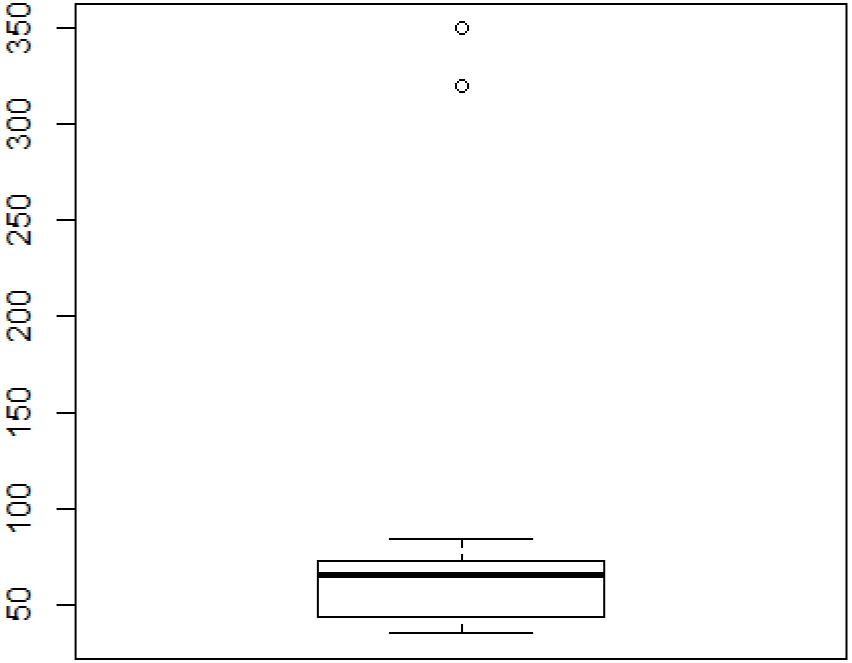
Boxplot

Example

Make first two observations to high.

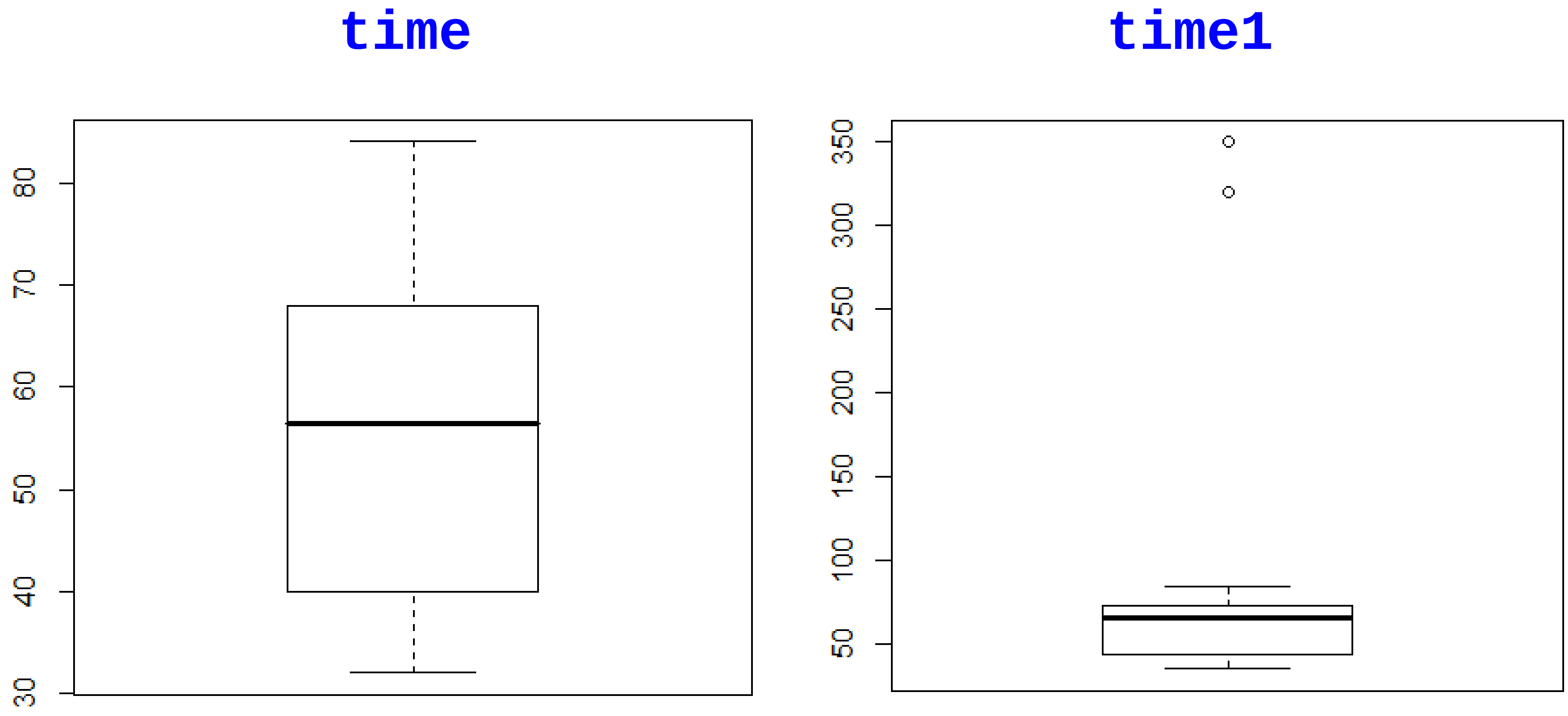
```
> time1 = c(320, 350, 45, 83, 74, 55, 68, 38,
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42,
58)
```

```
> boxplot(time1)
```



Boxplot

Example: Comparison of datasets through boxplots



time

time1

Different scales on y-axis.

Grouped Boxplot

Combine the data for which the boxplots are to be plotted in the format of Data Frame

Suppose the data vectors are `x`, `y` and `z`.

Create the dataframe as

```
data.frame(x, y, z)
```

Construct the grouped box plot as

```
boxplot(data.frame(x, y, z))
```

Grouped Boxplot

Example

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

```
> time1 = c(320, 350, 45, 83, 74, 55, 68, 38,
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

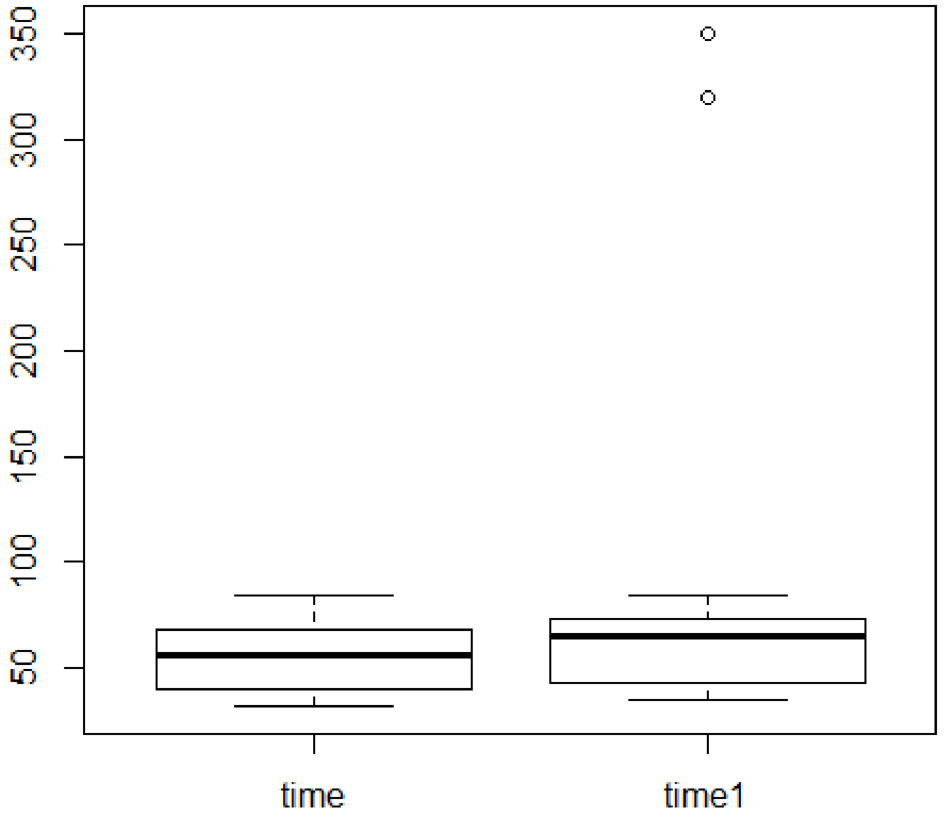
Create the data frame as follows:

```
> databoxplot = data.frame(time, time1)
```

Grouped Boxplot

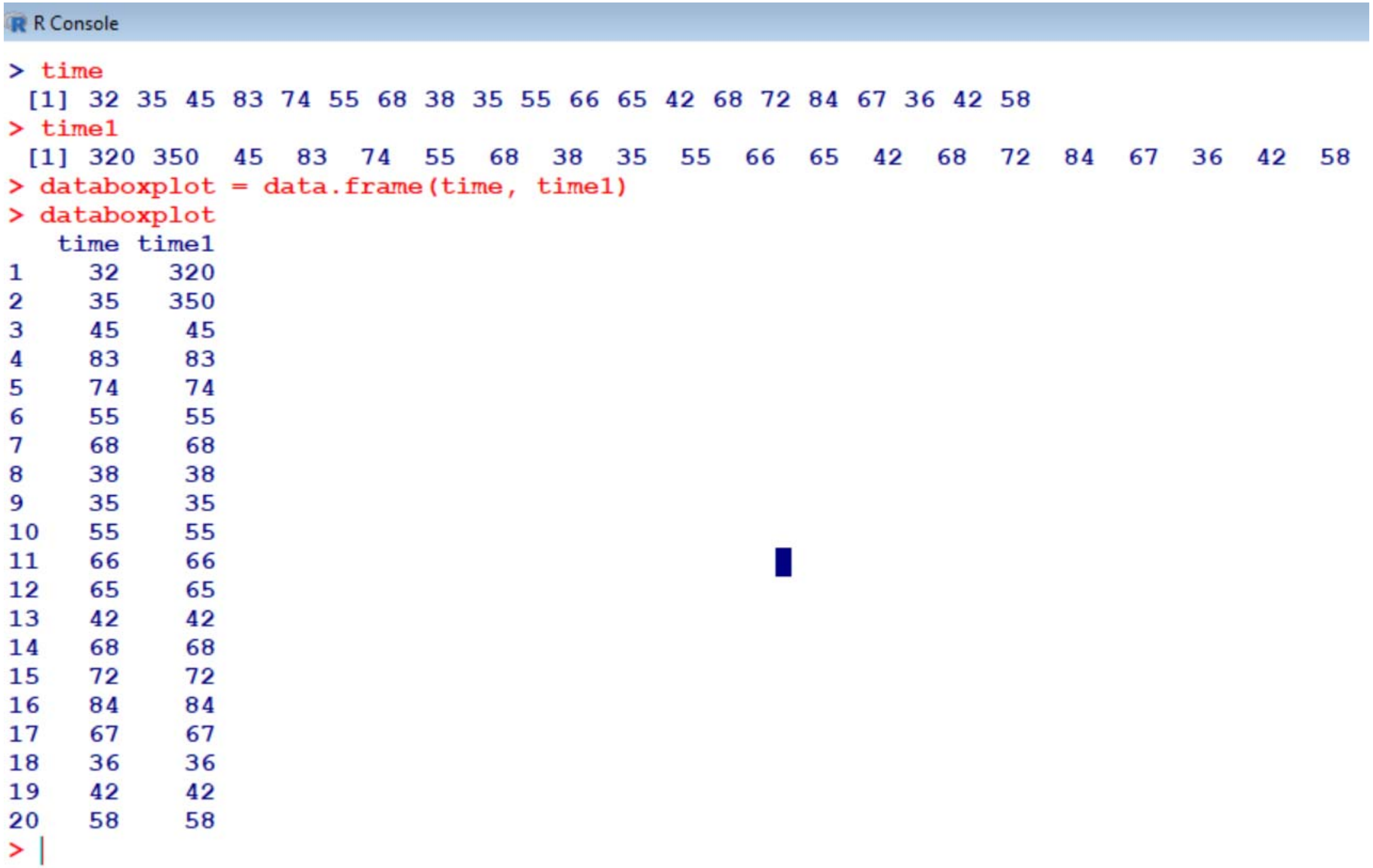
Example

```
> boxplot(databoxplot)
```



Grouped Boxplot

Example



```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
> time1
 [1] 320 350  45  83  74  55  68  38  35  55  66  65  42  68  72  84  67  36  42  58
> databoxplot = data.frame(time, time1)
> databoxplot
   time time1
1    32   320
2    35   350
3    45    45
4    83    83
5    74    74
6    55    55
7    68    68
8    38    38
9    35    35
10   55    55
11   66    66
12   65    65
13   42    42
14   68    68
15   72    72
16   84    84
17   67    67
18   36    36
19   42    42
20   58    58
> 
```

Grouped Boxplot

Example

Marks of 10 students in two different examinations are obtained as follows. We compare them using the boxplots.

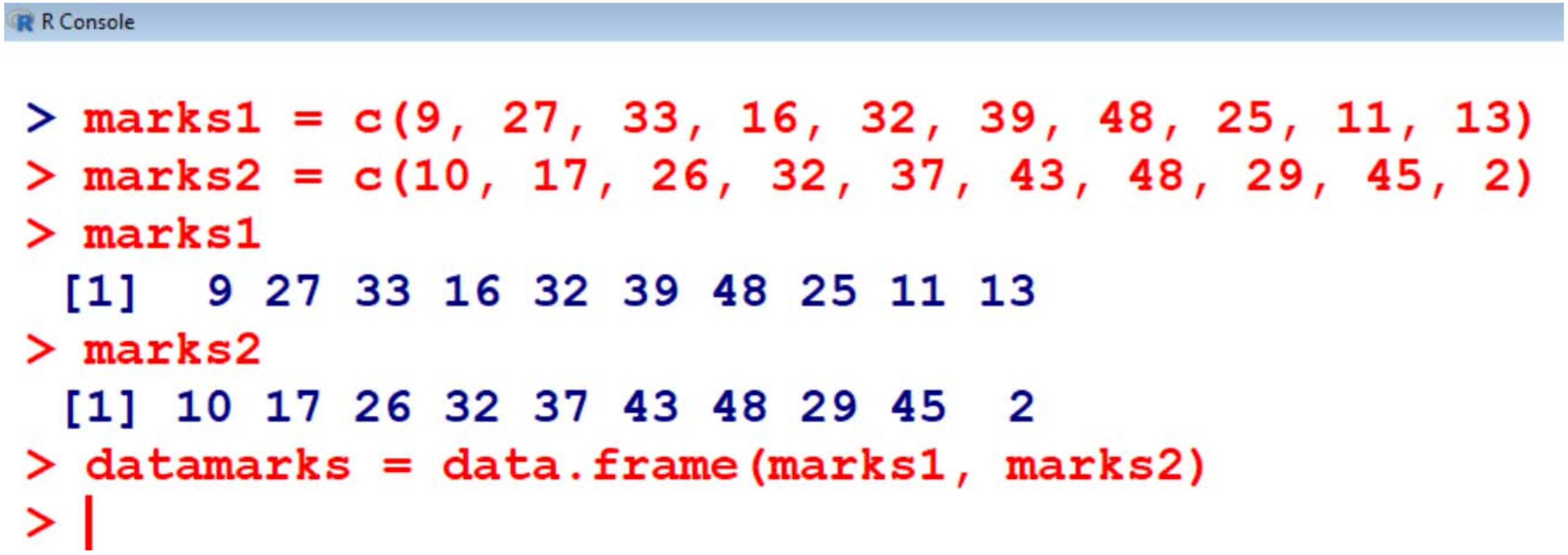
```
> marks1 = c(9,27,33,16,32,39,48,25,11,13)

> marks2 = c(10,17,26,32,37,43,48,29,45,2)
```

Create the data frame as follows:

```
> datamarks = data.frame(marks1, marks2)
```

Grouped Boxplot

Example

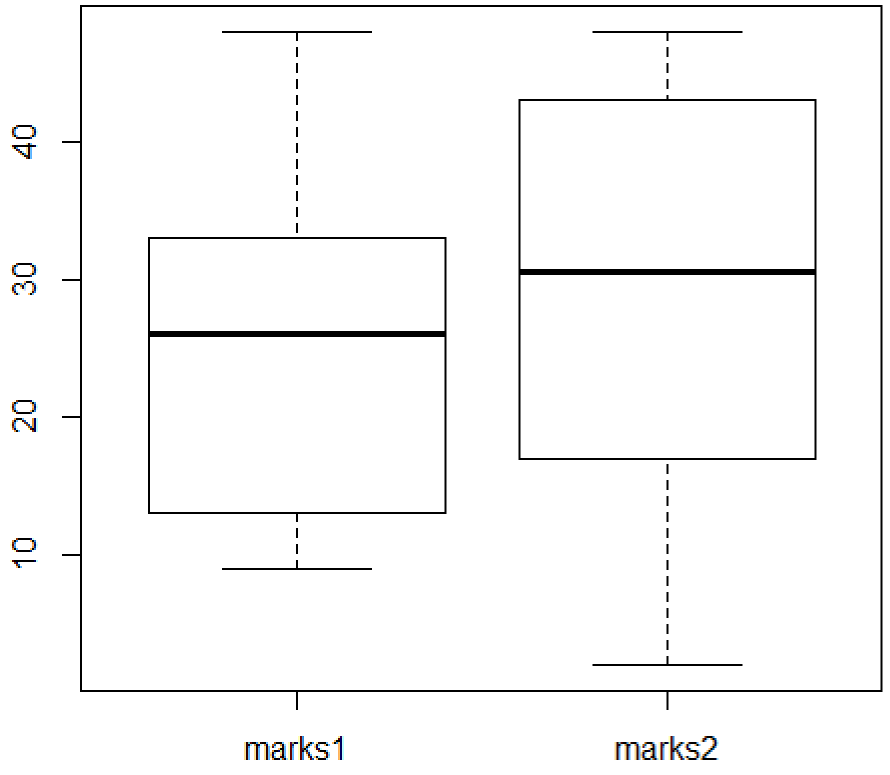


```
> marks1 = c(9, 27, 33, 16, 32, 39, 48, 25, 11, 13)
> marks2 = c(10, 17, 26, 32, 37, 43, 48, 29, 45, 2)
> marks1
 [1]  9 27 33 16 32 39 48 25 11 13
> marks2
 [1] 10 17 26 32 37 43 48 29 45  2
> datamarks = data.frame(marks1, marks2)
> |
```

Grouped Boxplot

Example

```
> boxplot(databoxplot)
```


marks1
marks2