# Descriptive Statistics With R Software

## Variation in Data

## ::

## Range, Interquartile Range and Quartile Deviation

**Shalabh**

**Department of Mathematics and Statistics**

**Indian Institute of Technology Kanpur**

# Measures of Variation (or Dispersion)

**Measures of central tendency gives an idea about the location where most of the data is concentrated.**

**Two different data sets may have same arithmetic mean but they may have different concentrations around mean.**

# Measures of Variation (or Dispersion)

**Example: The temperature of three cities in degree centigrade on 6 days are recorded as follows:**

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 |
|---|---|---|---|---|---|---|
| City 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| City 2 | -15 | -15 | -15 | 15 | 15 | 15 |
| City 3 | 11 | 9 | 10 | 8 | 12 | 10 |

**Arithmetic mean of the city 1 = 0**

**Arithmetic mean of the city 2 = 0**

**Arithmetic mean of the city 3 = 10**

# Measures of Variation (or Dispersion)

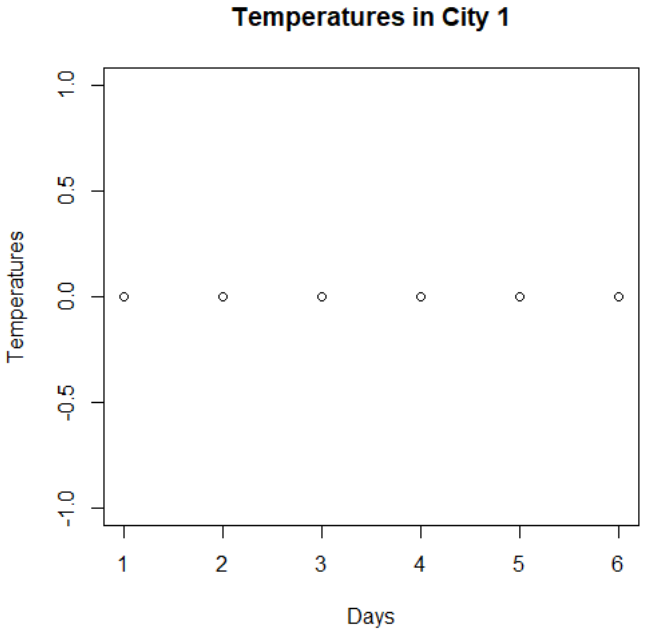
**Mean Temperatures in City 1 and City 2 are the same as 0 but does this makes any sense?**

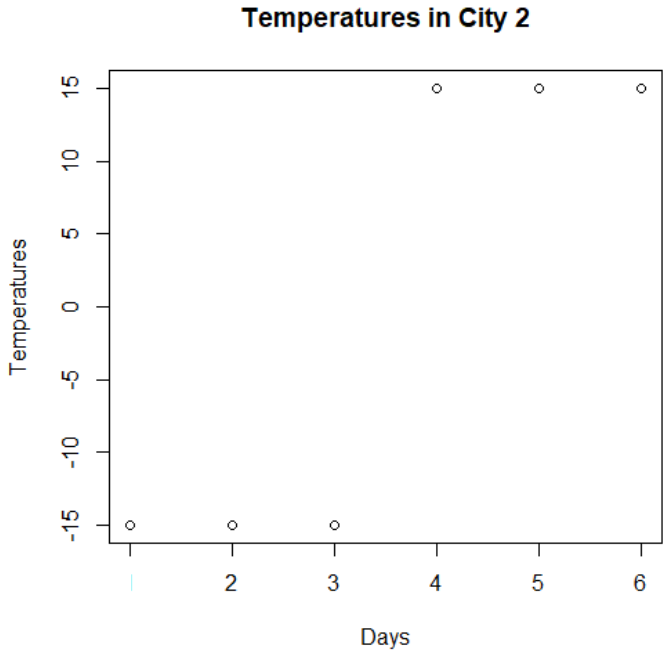
**City 2 has two extreme temperatures on -15 and 15.**

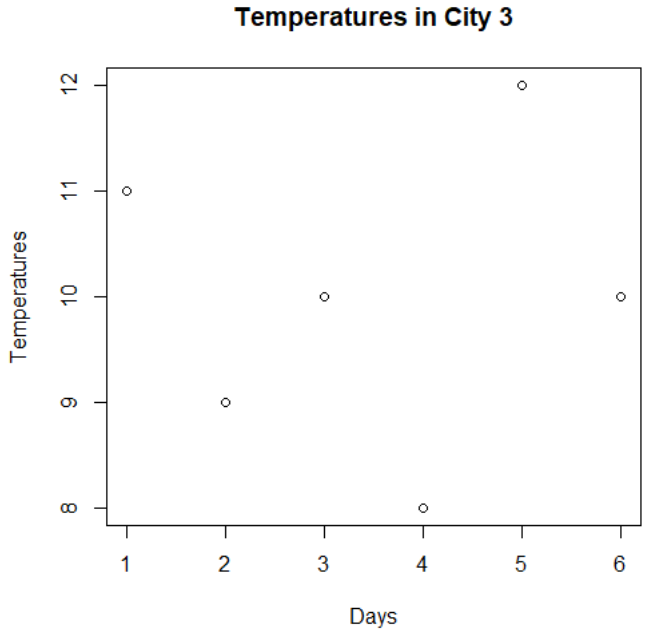
**City 3 has variation among the values. Do you think, is it more reliable temperature?**

**Let us have a graphical look.**

# Measures of Variation (or Dispersion)



# Measures of Variation (or Dispersion)

**Location measures are not enough to describe the behaviour of data.**

**Concentration or dispersion of observations around any particular value is another property to characterize the data.**

**How to capture this variation?**

**Various statistical measures of variation or dispersion are available.**

# Measures of Variation (or Dispersion)

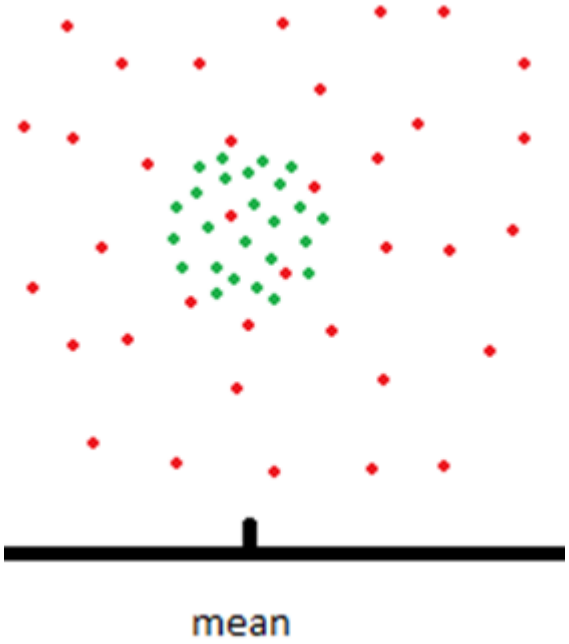
**Two data sets – Green and red dots,**

**Same mean of red and green colour data points.**

**Whose variation is more?**

**Which data is more dispersed?**

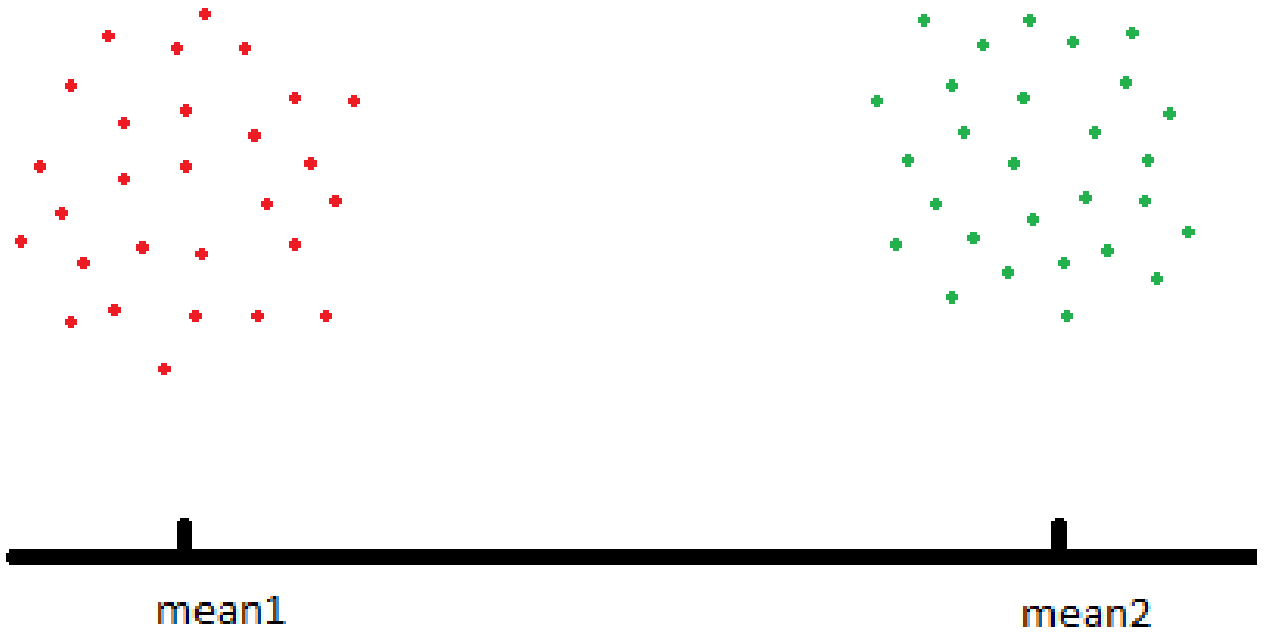
**Which data is more concentrated around mean?**



# Measures of Variation (or Dispersion)

**Two data sets – Green and red dots.**

**Same variability but different means.**




**Spread and scatterdness of data can be measured around any point but the mean value is more preffered.**

# Measures of Variation (or Dispersion)

**Measures of variation or dispersion helps in measuring the spread and scatterdness of data around any point, preferebly the arithmetic mean value.**

## Various measures of variation are available:

- **Range,**
- **Interquartile range,**
- **Quartile deviation,**
- **Absolute mean deviation,**
- **Variance,**
- **Standard deviation etc.**

# Range

**Observations: $x_1, x_2, \ldots, x_n$**

**Range: Difference between the maximum and minimum values of the data**

**$R = max\,(x_1, x_2, \ldots, x_n) - min\,(x_1, x_2, \ldots, x_n)$**

# Range

## Decision Making

**The data set having higher value of range has more variability.**

**The data set with lower value of range is preferable.**

**If we have two data sets and suppose their ranges are $Range_1$ and $Range_2$.**

**If $Range_1 > Range_2$ then the data in $Range_1$ is said to have more variability than the data in $Range_2$.**

## Range

**R command:**

**Data vector:** `x`

```
max(x) - min(x)
```

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

```
max(xna, na.rm = TRUE) - min(xna, na.rm = TRUE)
```

## Caution:

**Command** `range` **returns a vector containing the minimum and maximum of all the given arguments.**

# Range

## 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.**

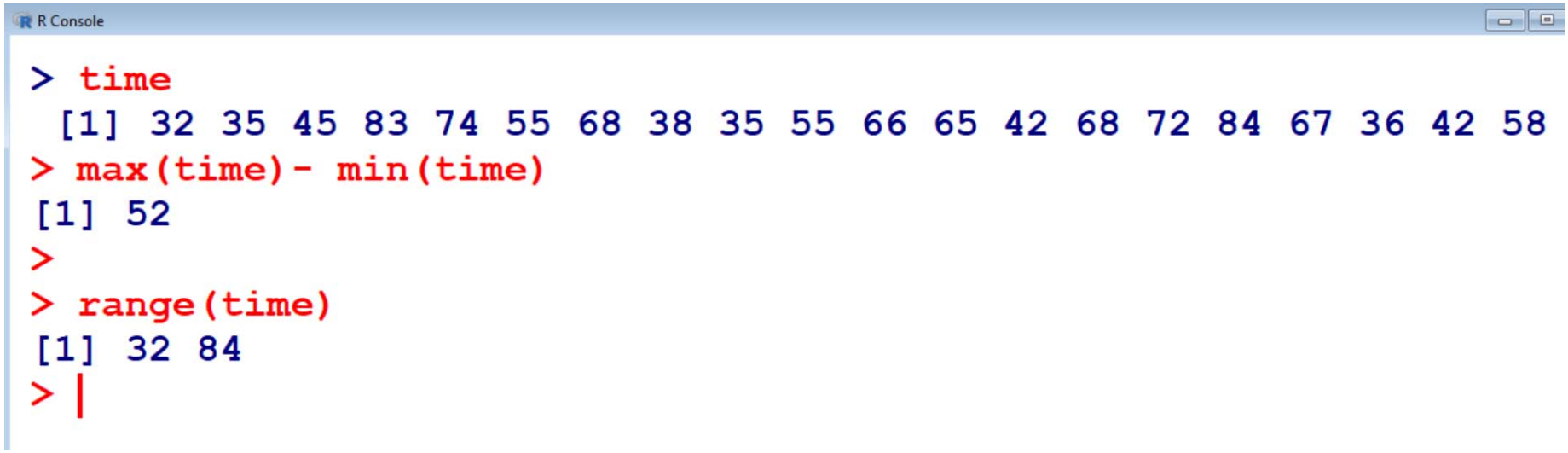
```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)

> max(time)- min(time)
[1] 52
```

## Caution:

```
> range(time)
[1] 32 84
```

# Range

## Example:



# Range

**Example:** Handling missing values

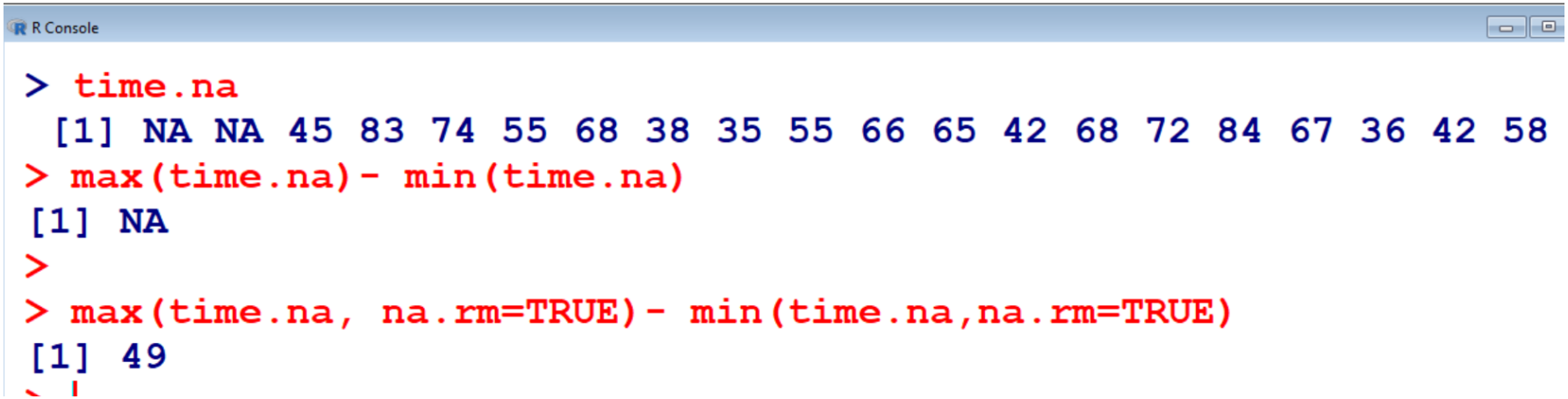
Suppose two data points are missing in the earlier example where the time taken (in seconds) by 20 participants in a race. They are recorded as NA

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

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

> max(time.na)- min(time.na)
[1] NA

> max(time.na,na.rm=TRUE)-min(time.na,na.rm=TRUE)
[1] 49
```

# Range

## Example: Handling missing values



## Interquartile Range

**Difference between the $75^{th}$ and $25^{th}$ quartile (or equivalently $3^{rd}$ and $1^{st}$ quartile).**

$$IQR = Q_3 - Q_1$$

**It covers centre of the distribution and contains 50% of the observations.**

# Interquartile Range

## Decision Making

**The data set having higher value of interquartile range has more variability.**

**The data set with lower value of interquartile range is preferable.**

**If we have two data sets and suppose their interquartile ranges are $IR_1$ and $IR_2$.**

**If $IR_1 > IR_2$ then the data in $IR_1$ is said to have more variability than the data in $IR_2$.**

## Interquartile Range

**R command:**

**Data vector: `x`**

```
IQR(x)
```

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

```
IQR(xna, na.rm = TRUE)
```

## Quartile Deviation

**Half difference between the $75^{th}$ and $25^{th}$ quartile (or equivalently $3^{rd}$ and $1^{st}$ quartile).**

**Half of Interquartile range.**

**Quartile deviation is defined as**

$$\frac{1}{2}(Q_3 - Q_1) = \frac{IQR}{2}$$

## Decision Making

**The data set having higher value of quartile deviation has more variability.**

# Quartile Deviation

**R command:**

**Data vector: `x`**

```
IQR(x)/2
```

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

```
IQR(xna, na.rm = TRUE)/2
```

# Interquartile Range and Quartile Deviation

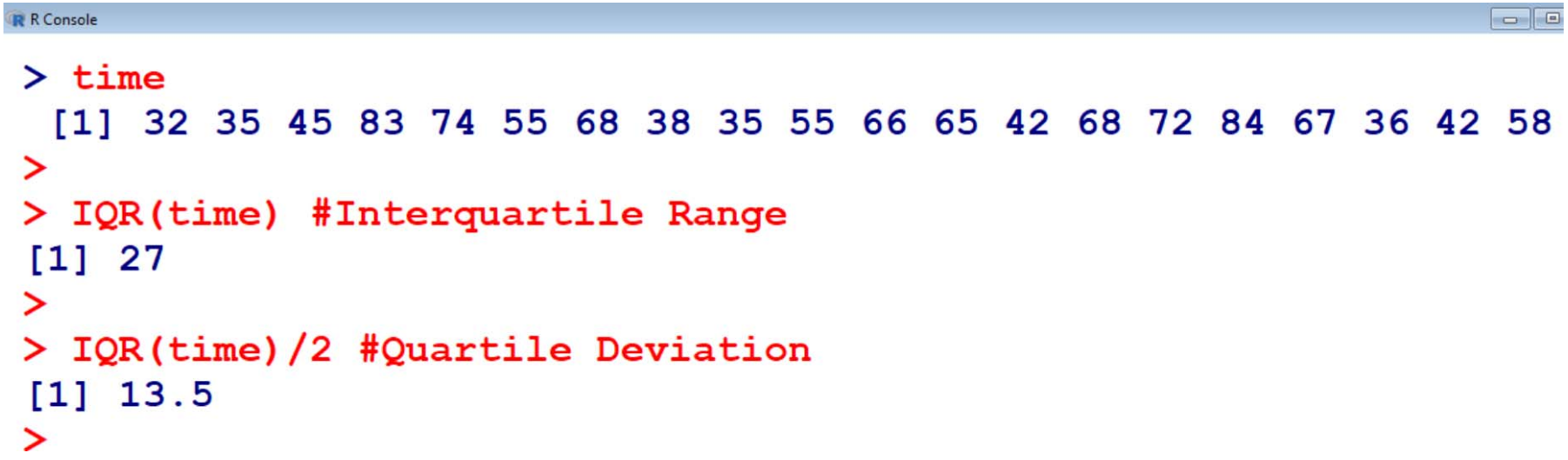
## Example:

**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
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68
72 84 67 36 42 58

> IQR(time) #Interquartile Range
[1] 27

> IQR(time)/2 #Quartile Deviation
[1] 13.5
```

# Interquartile Range and Quartile Deviation

## Example:



# Interquartile Range and Quartile Deviation

**Example:** 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 = c(NA, NA, 45, 83, 74, 55, 68, 38,
35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

# Interquartile Range and Quartile Deviation

**Example:** Handling missing values

```
> IQR(time.na) #Interquartile Range
Error in quantile.default(as.numeric(x),
c(0.25, 0.75), na.rm = na.rm, : missing values
and NaN's not allowed if 'na.rm' is FALSE

> IQR(time.na, na.rm = TRUE) #Interquartile Range
[1] 25.25

> IQR(time.na, na.rm = TRUE)/2 #Quartile Deviation

[1] 12.625
```