Introduction to R Software

Swayam Prabha

Lecture 36

Central Tendency and Variation in Data

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

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

Descriptive statistics:

First hand tools which gives first hand information.

- **Central tendency of data (Mean, median, mode, geometric mean, harmoninc mean etc.)**

- **Variation in data (variance, standard deviation, standard error, mean deviation etc.)**

Data:

```
84  73 133  38 115 183 157  96 108 146  58 144  56  15  65  65 154 103 181  10  25 130 160
147 151  12 131  97  76  94 185  93  41 171  66  55 175 104  80 131  38  86  81 182 179  47
178  28 154 147 135  70 145 200  87  57 137 173 143 148 126  98 114  98 195 112  27 116 125
185  39 133  87 153 179 134  10 110  62 122  88 102  59 184  21  15 178  93 107 155 123  22
119  82 164  58 170  46  88  13  42  38 126 153 187  65 182 155 172 198  46 115 145 152 138
200  18 161  57 182 173 194 153 190  97 132  70  53 170  16 54 148  20 155  58 125  83  30
146 163 166 133 129 184 168 173 139 119  69 105 173 109 175 124  59 196  94  73 188 155  41
158  94 131  22 137  14  36 159 166  67 181  64  63 167 118  87  95  95 101  65 110 116 155
117  60  74 151 107  77  66 126  42  30 196   51  30  32  17 172 100 161 193  54 184  65 32
128  41  26 131 111 101  13 137  56  29 164  41  42  32  60  54 153 121  22  92  75 121 147
58 104 188 124  82  37  39 200  34 109 142  50  14 176 137  50  83 168 117  12  85 158  12
188 114  88  13 140 109 144 26  21 149 165  32 195  42 164 164  60 136  86  41 189 145 182
30  87 110  41 132 156 156 172  29 199 103 185  79  86 140  59  70 183 114  53 169 172  13
197 143 152  87  72 125  11 197  63  67  64  38  53 160 118 177 106 151  84  36 137  77 166
116  27 162 51  23 109  63 145  37  45 179 112 177 124 153  52  42 185  45  36  76 101 151
82 181 126 118 184 200 130  25  71 131  38  98  83 171  53  97  62  90  48 152 130 128  71
181  98  84 174 160 110  44  70  48  40  94 134  25  38 104  21 196 104 198 151  65 111 161
181 127  77 175 150 113 172  79  66 144 127 158  82  42  66  56 119 133 100 105 177  33 192
157 150 108  82 166  28  76  19  15 161 173 158  51  90  84 168 120  44 120 163 161  49 162
164 141 142 170  43  71  71 119  41  13  90  85  78  79 106 176 178 192 64 108 187 162  19
33 175 118 174  14  24  53  34 180 169 187  85 175  19  44  99 139 190  13 151 157 144 143
123  28 145 185  38  94  57  13 164  34  26 169  52 138 195 128 105  73  10 158  37 188  99
117 137  13 139 115  91  16 151  21 193 153  50  91  35 124  54 152 197  92  73 136  71 138
74  22 197 198 151  69 199 200 142 123 123  35 183 191 194 174 173 190 102  29  59  85 165
29 159 147  71 150 115  26  58  63 131 126 140  45  45 124 192  39 200  27 126 192 160  84
114 171 156  72 112  10  71 112 110 188 91 111 115 183 125 187 136 129 158 134  65  33 146
70 141 196 177 107 200 146  76  89 176 117 192 141 182 194  39 197  94 138 133  85 111 165
27  36  87 134 107  38  47 118 130 129 154  85 149 116 150  27 143 134 130  34 162 161  92
159 190 188 169  23  45  95 181 109 156  14 162  32  40 189  68  99 151  22 169 187 150 124
147 182  37  11 127  23 173  31  81 140 174  57  77 126 118  86  25 109  15 189 179 191 143
69 177  74 197  30  70 183 112 160  79  31 113 139  66 122  87  74  48 173 200  58  92  57
168 197  59  45  79  79  16  31  47  13 198 182  67 152 148 185 166  48 182  31  74 105  58
125 119  59  66  69 197 106 146 181 149  42 168 150  58  34 184 127 145  40  65  23 197 163
53 155 103 147  82  75 190  29  92 164  41 131 152  38 195 110  86 164 198  33  60
 87 141 121  98  19 150 133 184 158  34  94 129  40 101 101  54 109 159  67 172 197  60 157
```

Central tendency of the data

Gives an idea about the mean value of the data

The data is clustered around what value?

Data: $x_1, x_2, \ldots, x_n$

x : Data vector

Arithmetic mean (mean) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

```
mean(x)
```

Central tendency of the data

Geometric mean

$$\bar{x}_{GM} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

```
prod(x)^(1/length(x))
```

(`length(x)` is equal to the number of elements in x)

Harmonic mean

$$\bar{x}_{HM} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

```
1/mean(1/x)
```

Central tendency of the data

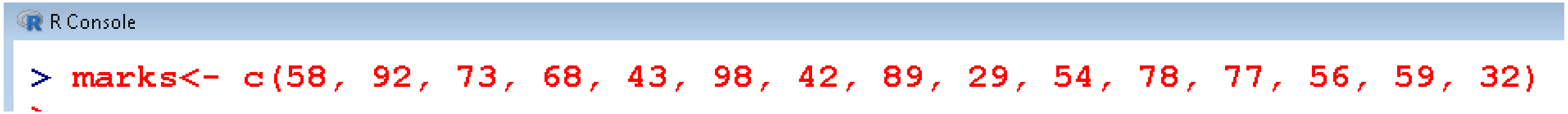
Median:

Value such that the number of observation above it is equal to the number of observation below it.

```
median(x)
```

Example

```
> marks<- c(58, 92, 73, 68, 43, 98, 42, 89,
29, 54, 78, 77, 56, 59, 32)
```

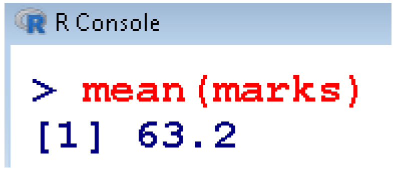


```
> marks<- c(58, 92, 73, 68, 43, 98, 42, 89, 29, 54, 78, 77, 56, 59, 32)
```

Arithmetic mean:

```
> mean(marks)
[1] 63.2
```

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



```
> mean(marks)
[1] 63.2
```

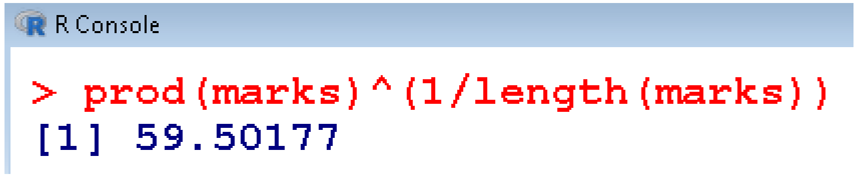
Example

```
> marks<- c(58, 92, 73, 68, 43, 98, 42, 89,
29, 54, 78, 77, 56, 59, 32)
```

Geometric mean:

$$\bar{x}_{GM} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

```
>  prod(marks)^(1/length(marks))
[1] 59.50177
```



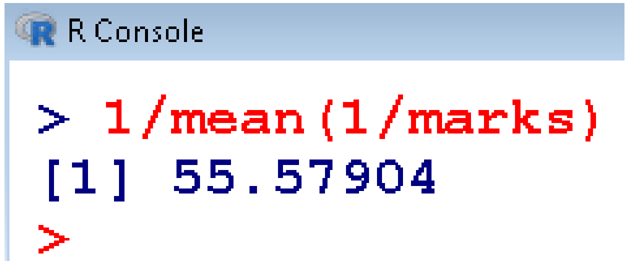
```
> prod(marks)^(1/length(marks))
[1] 59.50177
```

Example

Harmonic mean:

```
> 1/mean(1/marks)
[1] 55.57904
```

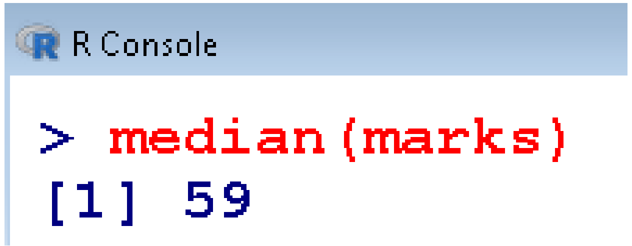
$$\overline{x}_{HM} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$



```
> 1/mean(1/marks)
[1] 55.57904
>
```

Median:

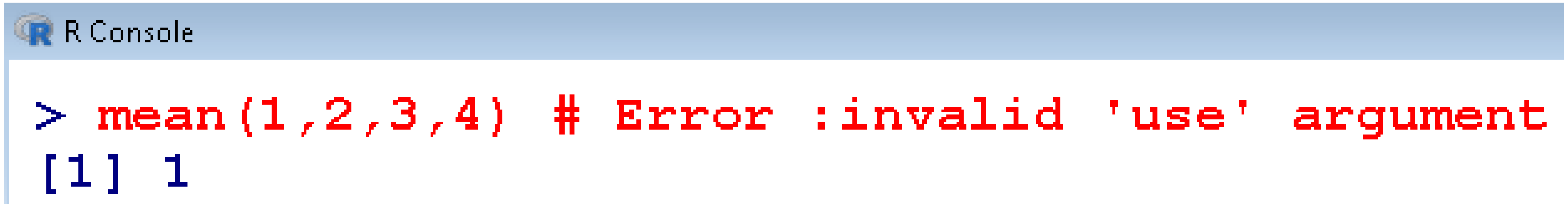
```
> median(marks)
[1] 59
```



```
> median(marks)
[1] 59
```

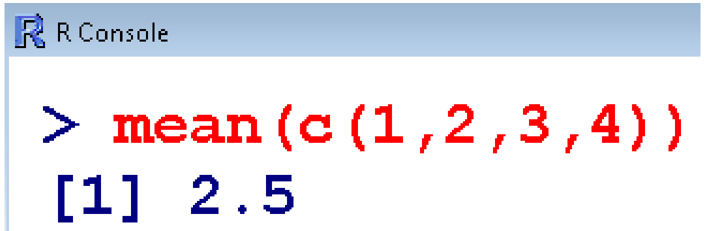
Doesn't do what you would expect:

```
> mean(1,2,3,4) # Error :invalid 'use' argument
 [1] 1
```


R Console

```
> mean(1,2,3,4) # Error :invalid 'use' argument
[1] 1
```

```
> mean(c(1,2,3,4))
 [1] 2.5
```


R Console

```
> mean(c(1,2,3,4))
[1] 2.5
```

Variability

Spread and scatterdness of data around any point, preferebly the mean value.

Data set 1: 360, 370, 380

mean = (360 + 370 + 380)/3 = 370

Data set 2: 10, 100, 1000

mean = (10 + 100 + 1000)/3 = 370

How to differentiate between the two data sets?

Variability

Variance $$\mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

x: data vector

```
var(x)
```

Positive square root of variance : standard deviation

```
sqrt(var(x))
```

Variability

Variance

Another variant,

$$\operatorname{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

If we want divisor to be n, then use

```
((n – 1)/n)*var(x)
```

where `n = length(x)`

Variability

Range:

$$\text{maximum}(x_1, x_2, \ldots, x_n) - \text{minimum}(x_1, x_2, \ldots, x_n)$$

```
max(x) – min(x)
```

Variability

Interquartile range:

Third quartile ($x_1, x_2, \ldots, x_n$) – First quartile ($x_1, x_2, \ldots, x_n$)

```
IQR(x)
```

Variability

Quartile deviation:

[Third quartile ($x_1, x_2, \ldots, x_n$) − First quartile ($x_1, x_2, \ldots, x_n$)]/2

= Interquartile range/2

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

Variability

Mean deviation:

$$MD(x) = \frac{1}{n}\sum_{i=1}^{n} | x_i - \overline{x} |$$

```
sum(abs(x-mean(x)))/length(x)

mean(abs(x-mean(x)))
```

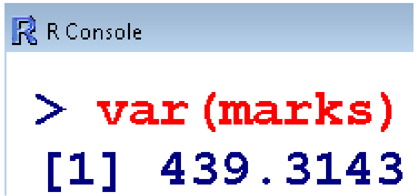
Example

x: data vector

```
> marks <- c(56, 59, 42, 68, 89, 29, 51, 82,
63, 86, 34, 96, 41, 75, 77 )
```

Variance:

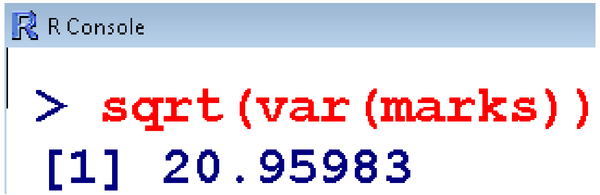
```
> var(marks)
[1] 439.3143
```



```
> var(marks)
[1] 439.3143
```

Standard deviation:

```
> sqrt(var(marks))
[1] 20.95983
```

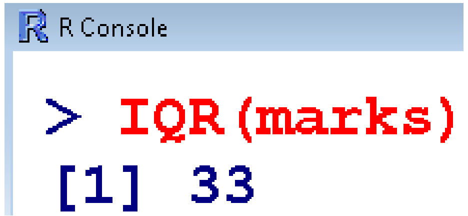


```
> sqrt(var(marks))
[1] 20.95983
```

Example

Interquartile Range:

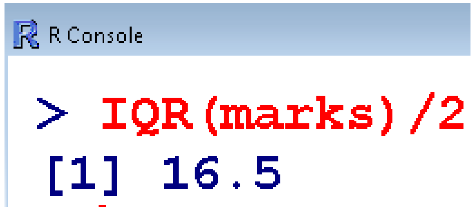
```
> IQR(marks)
[1] 33
```



```
> IQR(marks)
[1] 33
```

Quartile deviation :

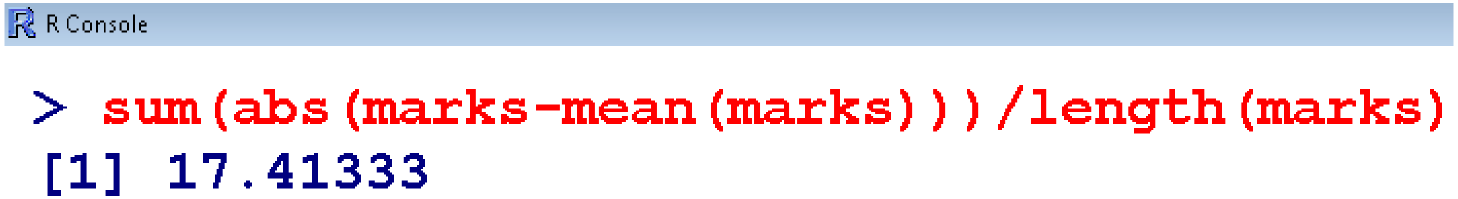
```
> IQR(marks)/2
[1] 16.5
```



```
> IQR(marks)/2
[1] 16.5
```

Mean deviation:

```
> sum(abs(marks-mean(marks)))/length(marks)
[1] 17.41333
```



```
> sum(abs(marks-mean(marks)))/length(marks)
[1] 17.41333
```

Example

Data set 1: 360, 370, 380

mean = (360 + 370 + 380)/3 = 370

```
> var(c(360, 370, 380 ))
[1] 100
```

Data set 2: 10, 100, 1000

mean = (10 + 100 + 1000)/3 = 370 **Same as of Data set 1**

```
> var(c(10, 100, 1000))
[1] 299700
```