Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 20

Quantiles

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

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

Quantiles

Median: Value which splits the data into two equal parts.

Quantile: Partitions the data into other proportions.

Quantiles

25% Quantile: Splits the data into two parts such that atleast 25% of the values are less than or equal to quantile and atleast 75% of the values are greater than or equal to the quantile.

50% Quantile: Splits the data into two parts such that atleast 50% of the values are less than or equal to quantile and atleast 50% of the values are greater than or equal to the quantile.

50% Quantile: Median

Quantiles

Let $0 \leq \alpha \leq 1$.

$(\alpha \times 100)\%$ quantile: Value which divides the data in proportions of $(\alpha \times 100)$ % and $(1 - \alpha) \times 100$ % such that at least $(\alpha \times 100)\%$ of the values are less than or equal to the quantile and at least $(1 - \alpha) \times 100\%$ of the values are greater than or equal to the quantile.

Quantiles

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

Order the Observations : $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

where $x_{(1)} = \min(x_1, x_2, \ldots, x_n)$, $x_{(n)} = \max(x_1, x_2, \ldots, x_n)$

(α x 100)% quantile:

$$\overline{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer, choose } k \text{ as the smallest integer} > n\alpha \\ \dfrac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2} & \text{if } n\alpha \text{ is an integer.} \end{cases}$$

Quantiles

Quartiles

The values which divide the given data into four equal parts, say,

Q_1, Q_2, Q_3, Q_4

Q_1 : First quartile which has 25% of the observations.

Q_2 : Second quartile which has 50% of the observations – median.

Q_3 : Third quartile which has 75% of the observations.

Q_4 : Fourth quartile which has 100% of the observations.

Quantiles

Deciles

The values which divide the given data into ten equal parts, say,

$D_1, D_2, \ldots, D_{10}$

D_1 : First decile which has 10% of the observations.

D_2 : Second decile which has 20% of the observations.

D_5 : Fifth decile which has 50% of the observations – median.

D_9 : Ninth decile which has 90% of the observations.

Quantiles

Percentiles

The values which divide the given data into hundred equal parts, say, P_1, P_2,..., P_{100}

P_1 : First percentile which has 1% of the observations.

P_2 : Second percentile which has 2% of the observations.

P_{50} : Fiftieth percentile which has 50% of the observations – median.

P_{90} : Ninetieth percentile which has 90% of the observations.

Quantiles

R Command :

```
quantile(x, ...)

quantile(x, probs =, na.rm =, type =, ...)
```

Arguments

`x` numeric vector whose sample quantiles are wanted,

`probs` numeric vector of probabilities with values in [0, 1].

`na.rm` value `TRUE` if data in `x` is `NA` otherwise default is `FALSE`

`type` an integer between 1 and 9 selecting one of the nine quantile algorithms

Quantiles

R Command :

R offers nine different ways to obtain quantiles, each of which is chosen by the `type` argument.

Type 1 : Inverse of empirical distribution function.

Type 2 : Similar to type 1 but with averaging at discontinuities.

Type 3 : Nearest even order statistic.

… …

Quantiles

Example

Height of 50 persons in centimetres are recorded as follows:

166,125,130,142,147,159,159,147,165,156,149,164,137,166,135,142,

133,136,127,143,165,121,142,148,158,146,154,157,124,125,158,159,

164,143,154,152,141,164,131,152,152,161,143,143,139,131,125,145,

140,163

```
> height = c(166,125,130,142,147,159,159,147,
165,156,149,164,137,166,135,142,133,136,127,143,
165,121,142,148,158,146,154,157,124,125,158,159,
164,143,154,152,141,164,131,152,152,161,143,143,
139,131,125,145,140,163)
```

Quantiles

Example: Quantiles

```
> quantile(height)
   0%   25%   50%   75%  100%
121.0 137.5 146.5 158.0 166.0
```

Default outcome is quartiles: Q_1, Q_2, Q_3, Q_4

Quantiles

Example: Quartiles Q_1, Q_2, Q_3, Q_4

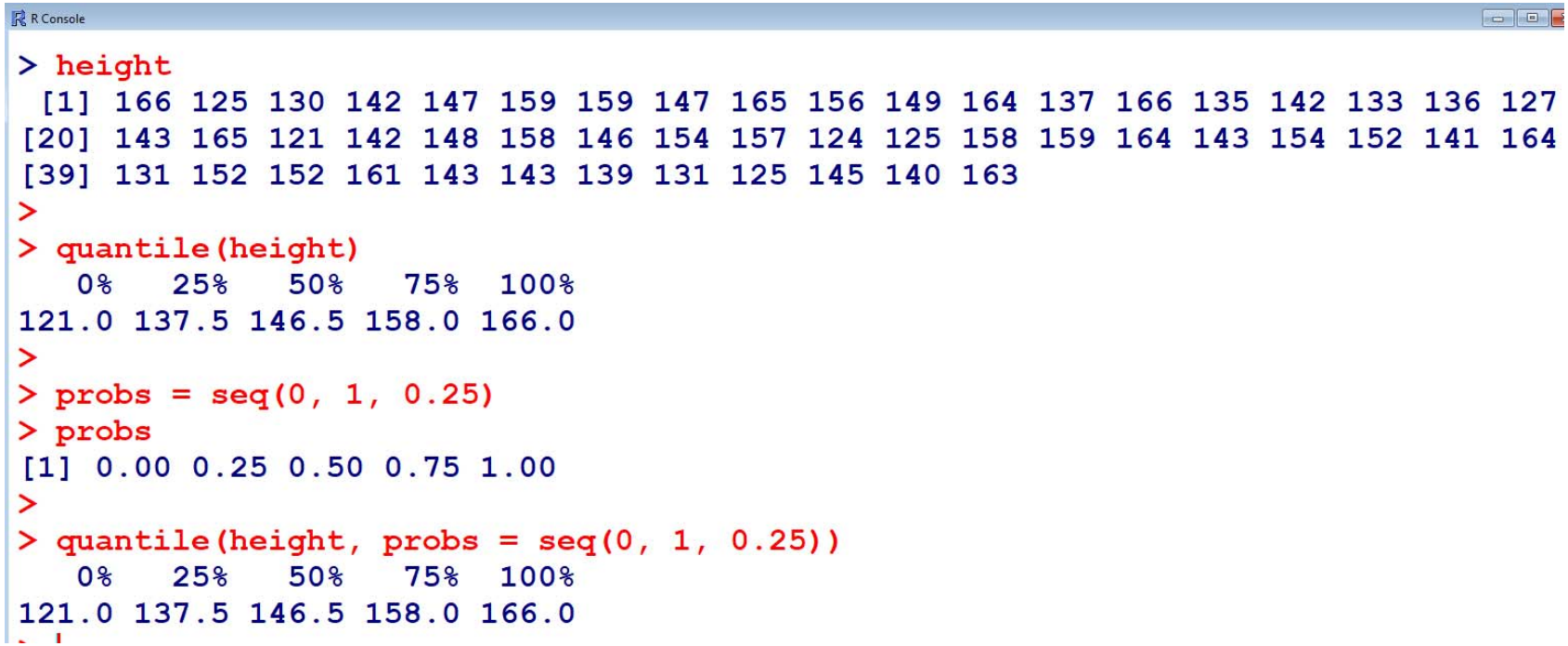
```
> probs = seq(0, 1, 0.25) # probs for quartiles
> probs
[1] 0.00 0.25 0.50 0.75 1.00

> quantile(height, probs = seq(0, 1, 0.25))
   0%   25%   50%   75%  100%
121.0 137.5 146.5 158.0 166.0
```

Same as earlier using `quantile` function.

Quantiles

Example: Quartiles Q_1, Q_2, Q_3, Q_4



Quantiles

Example: Deciles $D_1, D_2, \ldots, D_{10}$

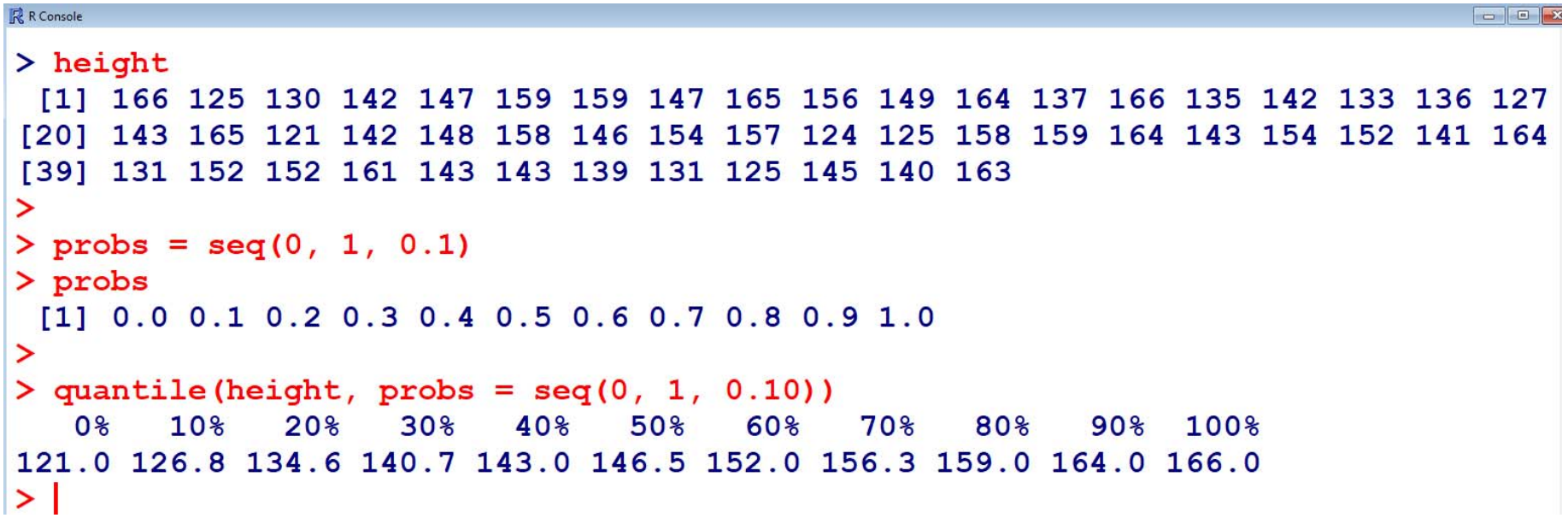
```
> probs = seq(0, 1, 0.1) # probs for deciles
> probs
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

> quantile(height, probs = seq(0, 1, 0.10))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
121.0 126.8 134.6 140.7 143.0 146.5 152.0 156.3 159.0 164.0 166.0
```

Need to change the `probs` function only.

Quantiles

Example: Deciles D_1, D_2,..., D_{10}



Quantiles

Example: Percentiles P_1, P_2,..., P_{100}

```
> probs = seq(0, 1, 0.01) # probs for percentiles
> probs = seq(0, 1, 0.01)
> probs
  [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
 [16] 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
 [31] 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
 [46] 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
 [61] 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
 [76] 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
 [91] 0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

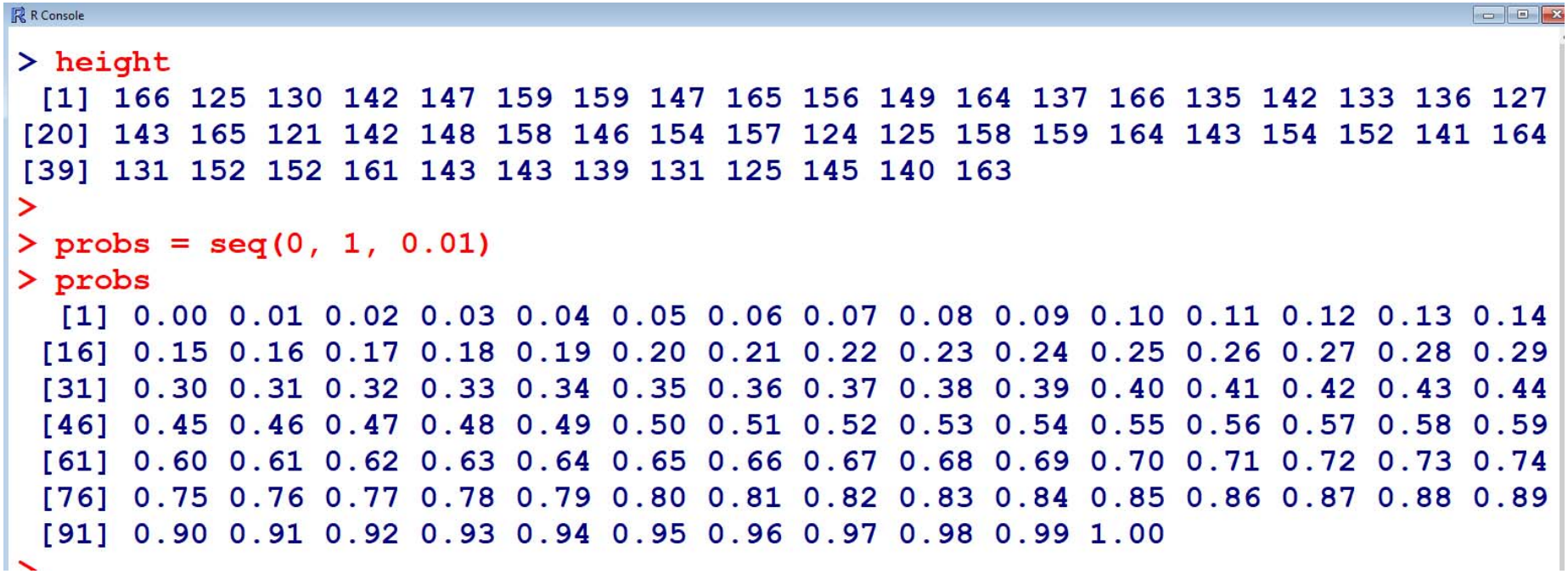
Quantiles

Example: Percentiles P_1, P_2,..., P_{10}

```
> quantile(height, probs = seq(0, 1, 0.01))
    0%     1%     2%     3%     4%     5%     6%     7%     8%     9%    10%
121.00 122.47 123.94 124.47 124.96 125.00 125.00 125.00 125.00 125.82 126.80
   11%    12%    13%    14%    15%    16%    17%    18%    19%    20%    21%
128.17 129.64 130.37 130.86 131.00 131.00 131.66 132.64 133.62 134.60 135.29
   22%    23%    24%    25%    26%    27%    28%    29%    30%    31%    32%
135.78 136.27 136.76 137.50 138.48 139.23 139.72 140.21 140.70 141.19 141.68
   33%    34%    35%    36%    37%    38%    39%    40%    41%    42%    43%
142.00 142.00 142.00 142.00 142.13 142.62 143.00 143.00 143.00 143.00 143.00
   44%    45%    46%    47%    48%    49%    50%    51%    52%    53%    54%
143.00 143.10 144.08 145.03 145.52 146.01 146.50 146.99 147.00 147.00 147.46
   55%    56%    57%    58%    59%    60%    61%    62%    63%    64%    65%
147.95 148.44 148.93 150.26 151.73 152.00 152.00 152.00 152.00 152.72 153.70
   66%    67%    68%    69%    70%    71%    72%    73%    74%    75%    76%
154.00 154.00 154.64 155.62 156.30 156.79 157.28 157.77 158.00 158.00 158.24
   77%    78%    79%    80%    81%    82%    83%    84%    85%    86%    87%
158.73 159.00 159.00 159.00 159.00 159.36 160.34 161.32 162.30 163.14 163.63
   88%    89%    90%    91%    92%    93%    94%    95%    96%    97%    98%
164.00 164.00 164.00 164.00 164.08 164.57 165.00 165.00 165.04 165.53 166.00
   99%   100%
166.00 166.00
```

Quantiles

Example: Percentiles P_1, P_2,..., P_{10}



Quantiles

Example: Percentiles P_1, P_2,..., P_{10}

R Console

```
> quantile(height, probs = seq(0, 1, 0.01))
    0%     1%     2%     3%     4%     5%     6%     7%     8%     9%    10%
121.00 122.47 123.94 124.47 124.96 125.00 125.00 125.00 125.00 125.82 126.80
   11%    12%    13%    14%    15%    16%    17%    18%    19%    20%    21%
128.17 129.64 130.37 130.86 131.00 131.00 131.66 132.64 133.62 134.60 135.29
   22%    23%    24%    25%    26%    27%    28%    29%    30%    31%    32%
135.78 136.27 136.76 137.50 138.48 139.23 139.72 140.21 140.70 141.19 141.68
   33%    34%    35%    36%    37%    38%    39%    40%    41%    42%    43%
142.00 142.00 142.00 142.00 142.13 142.62 143.00 143.00 143.00 143.00 143.00
   44%    45%    46%    47%    48%    49%    50%    51%    52%    53%    54%
143.00 143.10 144.08 145.03 145.52 146.01 146.50 146.99 147.00 147.00 147.46
   55%    56%    57%    58%    59%    60%    61%    62%    63%    64%    65%
147.95 148.44 148.93 150.26 151.73 152.00 152.00 152.00 152.00 152.72 153.70
   66%    67%    68%    69%    70%    71%    72%    73%    74%    75%    76%
154.00 154.00 154.64 155.62 156.30 156.79 157.28 157.77 158.00 158.00 158.24
   77%    78%    79%    80%    81%    82%    83%    84%    85%    86%    87%
158.73 159.00 159.00 159.00 159.00 159.36 160.34 161.32 162.30 163.14 163.63
   88%    89%    90%    91%    92%    93%    94%    95%    96%    97%    98%
164.00 164.00 164.00 164.00 164.08 164.57 165.00 165.00 165.04 165.53 166.00
   99%   100%
166.00 166.00
```

Quantiles with Missing Data

Example

Height of 50 persons in centimetres are recorded and two values are missing as follows:

NA,NA,130,142,147,159,159,147,165,156,149,164,137,166,135,142,

133,136,127,143,165,121,142,148,158,146,154,157,124,125,158,159,

164,143,154,152,141,164,131,152,152,161,143,143,139,131,125,145,

140,163

```
> height.na = c(NA,NA,130,142,147,159,159,147,
165,156,149,164,137,166,135,142,133,136,127,143,
165,121,142,148,158,146,154,157,124,125,158,159,
164,143,154,152,141,164,131,152,152,161,143,143,
139,131,125,145,140,163)
```

Quantiles

Example: Quantiles- Quartiles

```
> quantile(height.na)
Error in quantile.default(height.na) :
  missing values and NaN's not allowed if
'na.rm' is FALSE


> quantile(height.na, na.rm=TRUE)
   0%   25%   50%   75%  100%
121.0 138.5 146.5 158.0 166.0
```

Example: Quantiles- Deciles

```
> quantile(height.na, na.rm=TRUE, probs=seq(0,1,0.1)) # Deciles
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
121.0 129.1 135.4 141.1 143.0 146.5 152.0 155.8 159.0 164.0 166.0
```

Quantiles

Example: Quantiles

R Console

```
> height.na
 [1]  NA  NA 130 142 147 159 159 147 165 156 149 164 137 166 135 142
[17] 133 136 127 143 165 121 142 148 158 146 154 157 124 125 158 159
[33] 164 143 154 152 141 164 131 152 152 161 143 143 139 131 125 145
[49] 140 163
> quantile(height.na)
Error in quantile.default(height.na) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
> quantile(height.na, na.rm=TRUE)
   0%   25%   50%   75%  100%
121.0 138.5 146.5 158.0 166.0
>
> quantile(height.na, na.rm=TRUE, probs=seq(0,1,0.1)) # Deciles
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
121.0 129.1 135.4 141.1 143.0 146.5 152.0 155.8 159.0 164.0 166.0
>
```