Topic 5

Exploratory Data Analysis with R: Descriptive Statistics

Sergey Mastitsky ©
Klaipeda, 28-30 September 2011

A number of functions are available to calculate descriptive statistics

- `mean()`
- `median()`
- `var()`
- `sd()`
- `range()`
- `min()`
- `max()`
- `quantile()`
- `IQR()`, etc.

A few examples…

```
# Using the LWdata dataset, calculate
> mean(Length)
[1] 19.16753
> sd(Length)
[1] 6.238585
> quantile(Length)
  0%   25%   50%   75%   100%
 7.93 13.71 18.19 24.48 34.04
```

Quick statistics for several groups

- The function `tapply()` allows one to apply a function to a set of values grouped by the levels of certain factors
- General syntax:

```
tapply(X, INDEX, FUN = …)
```

```
X - a numerical vector
INDEX - a list of factors
FUN - function to apply
```

Examples...

```
> tapply(Weight, Treatment, mean)
Control         A         B         C
0.3255    0.3071    0.2495    0.2580

> tapply(Weight, Treatment, sd)
Control         A         B         C
 0.2960    0.2691    0.2498    0.2642
```

The `summary()` function

- Generic function

- Can be applied both to a single variable

```
> summary(Length)
> summary(Treatment)
```

- …and to the entire data frame

```
> summary(LWdata)
```

Outputs of the `summary()` function

```
> summary(Length)
Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
7.93   13.71   18.19   19.17   24.48   34.04
```

```
> summary(Treatment)
Control       A       B       C
     90      89      90      92
```

Summary on a data frame

```
> summary(LWdata)
   Treatment       Barrel        Length          Weight      
 Control:90   C3     : 36   Min.   : 7.93   Min.   :0.012  
 A      :89   A2     : 30   1st Qu.:13.71   1st Qu.:0.070  
 B      :90   A3     : 30   Median :18.19   Median :0.182  
 C      :92   B1     : 30   Mean   :19.17   Mean   :0.285  
              B2     : 30   3rd Qu.:24.48   3rd Qu.:0.436  
              B3     : 30   Max.   :34.04   Max.   :1.166  
              (Other):175                                  
```


Author: Sergey Mastitsky