Descriptive Statistics

- Representative value
  - Mean, median, mode, range, min, max, variation, standard deviation, interquartile range, outlier
- Visualization
  - Histogram
  - Boxplot
  - Scatterdiagram

Data for task

- Install packages
  - “Rling”
    - Download “Rling_1.0.tar.gz”

    > install.packages(“C:/yourpath/Rling_1.0.tar.gz”, repos=NULL, type=“source”) # or

    > install.packages(file.choose(), repos=NULL, type=“source”)
- Load packages

  > library(Rling) #install first
- Load data

  > data(ldt)
- **Load “ldt.Rdata” directly**
- Check data “ldt” with head() and str()

Measures of central tendency

- Mean
  - aka, the average

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

- Median
  - The middle value when values are ranked in ascending or descending order
  - (n+1)/2th value, where n='number of values'
  - Useful when there are extreme values that seriously affect the mean
    - Example: consider salaries of a company's 6 employees: $17,200, 18,500, 21,200, 23,000, 24,200 while the company CEO's salary is $225,000. The company's mean salary will be $54,850 while teh median will be: $22,100
- Mode
  - Most frequently occurring value
  - Useful for finding most common, popular item/characteristic of a data set
  - Possible to have more than one mode
    - Example : in the data set [2, 3, 5, 5, 6, 7, 7, 8, 10] modes are: 5 and 7 (bimodal)
- In a symmetrical distribution, mean, median and mode will be identical

Measures of central tendency in R (2)

- Check the mean of word length
  - mean(ldt$Length)
- Check the median of word length
  - Median: the value in the middle of the ordered values
    - Sort word length values in ascending order
    - Check the 50^{th} and 51^{st} element
    - median (odd-numbered items): the middle value
    - median (even-numbered items): average of the two middle values
  - measure the median of word length using function median()

Measures of central tendency in R (3)

- Check the mode of the word length
  - No built-in function for ‘mode’
  - Using table() and names()

```
> myt = table(ldt$Length)
> names(myt)[myt == max(myt)]
```

- Task (mymode)
  - Write a function ‘mymode’ that inputs a numeric vector and outputs its mode value(s).
  - Test your function with “ldt$Length”
  - Test your function with “var= c(2, 3, 5, 5, 6, 7, 7, 8, 10)”

Measures of dispersion

- Range
  - Representing min and max values
  - Use range(), min(), max()
- Variance
  - The average of the squared deviations from the mean
  - Use var()
- Stadndard deviation
  - The square root of the variance
  - Use sd() or sqrt(var())
- Interquartile range
  - The difference between the third (75%) and the first (25%) quartiles.
  - A robust measure as impact of outliers are weakened
  - Useful for non-normal distribution
  - Use IQR()

Why squared?

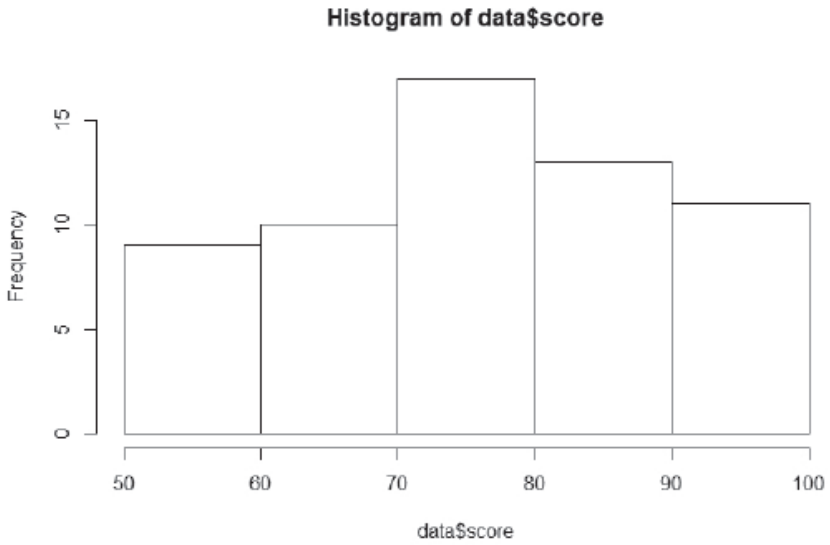
	population	sample
variance	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$	$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$
standard deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

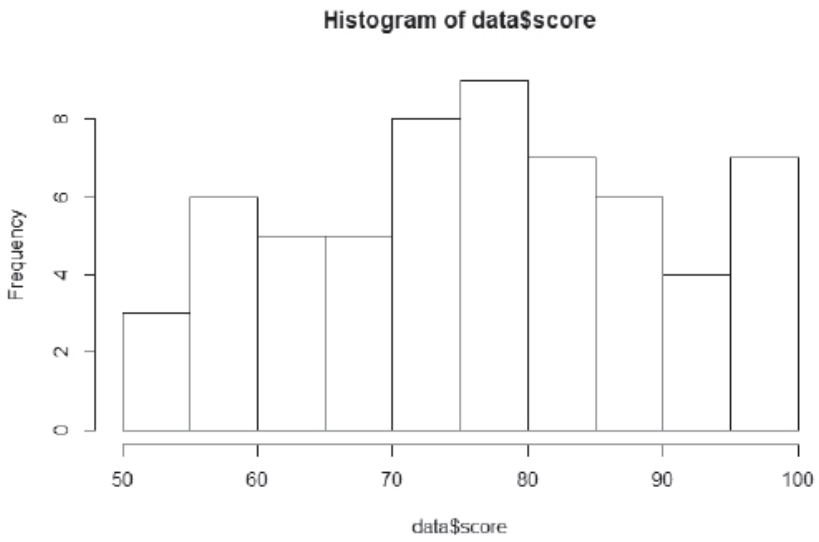
Outliers

- A data point or observation whose value is quite different from the others in the data set being analyzed
- Sometimes they represent data entry errors
- To check if the data was entered correctly
- To investigate whether the cases in question actually belong to the same population as the other cases

Histogram (1)

- Histogram
  - Numerical data (continuous, discrete)
  - Cf. bar chart (categorical data)





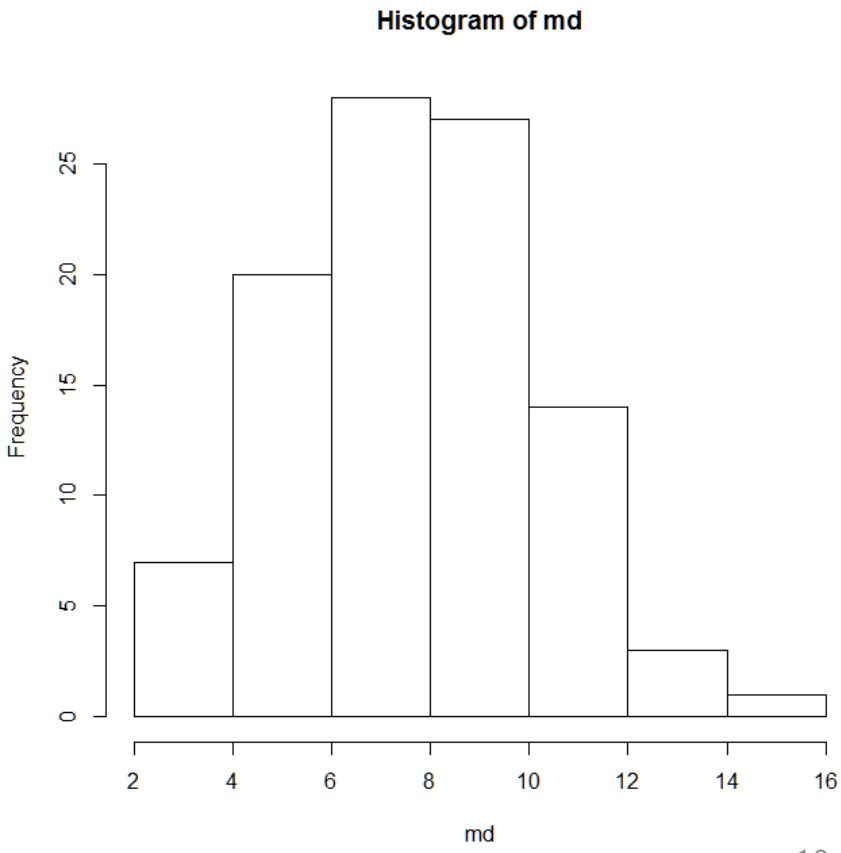
Histogram (2)

- Basic histogram

  > md = ldt$Length

  > hist(md)

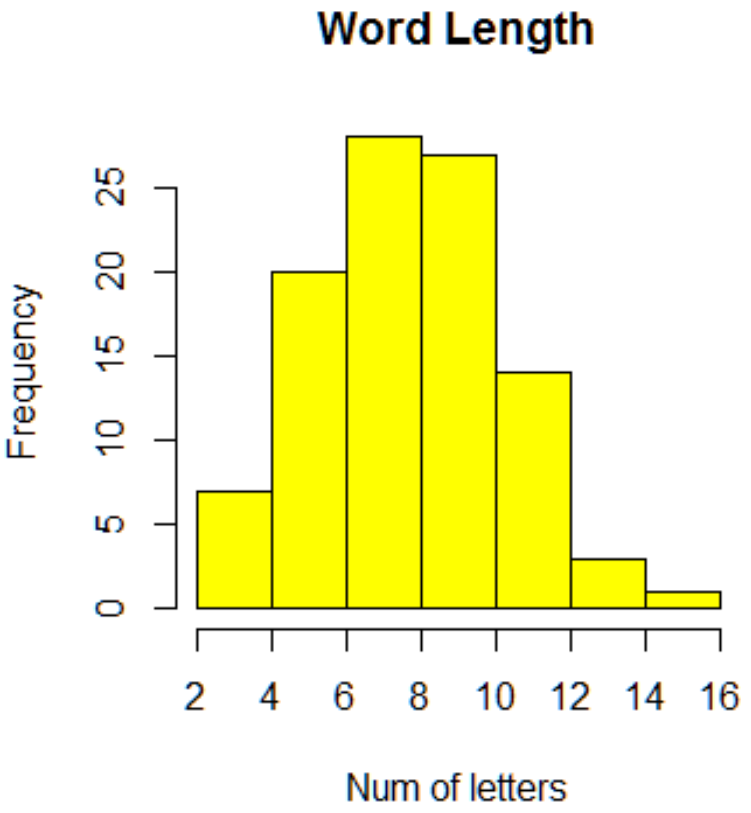
  (or, hist(ldt$Length)



Histogram (3)

- Adjust graph with arguments

  > hist(md, main="Word Length"
  xlab="Num of letters",
  col="yellow")

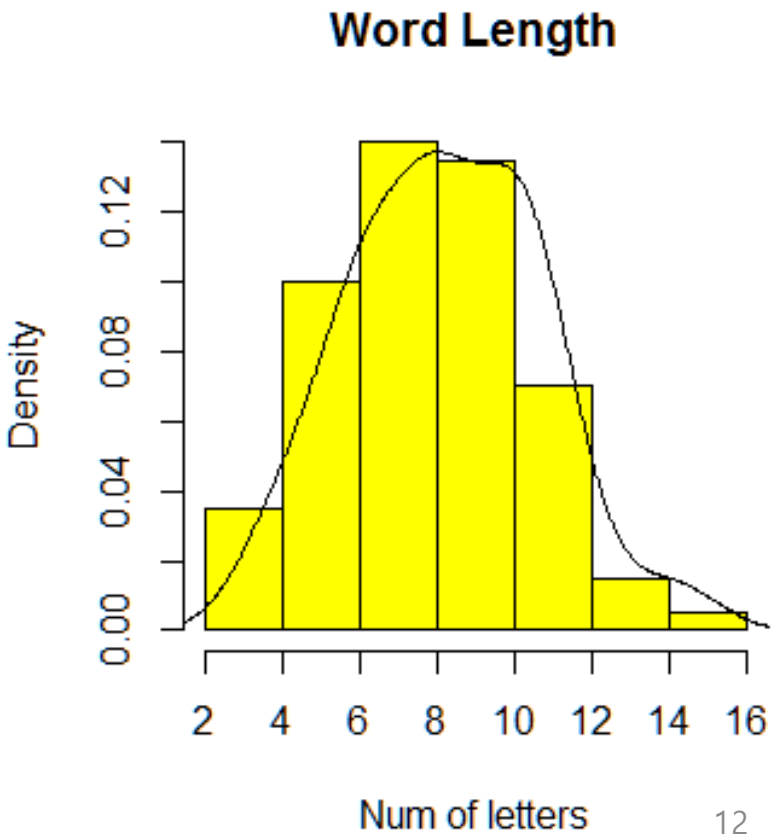


Histogram (4)

- Density plot

  > hist(md, main="Word Length", xlab="Num of letters", col="yellow", prob=TRUE)

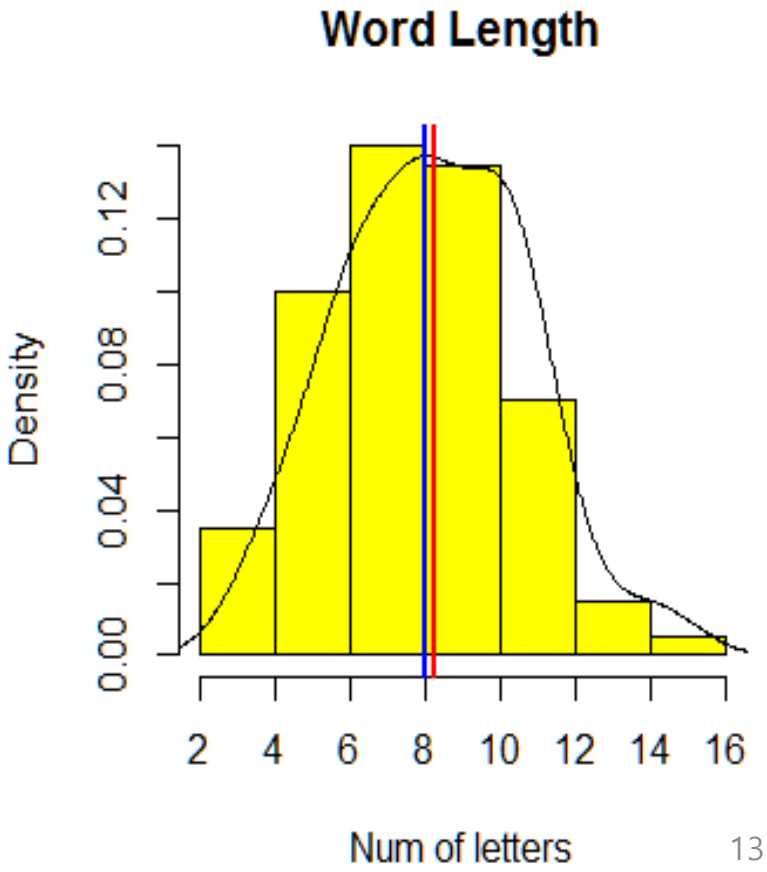
  > lines(density(md), lwd=1.5)



Histogram (5)

- Show mean and median

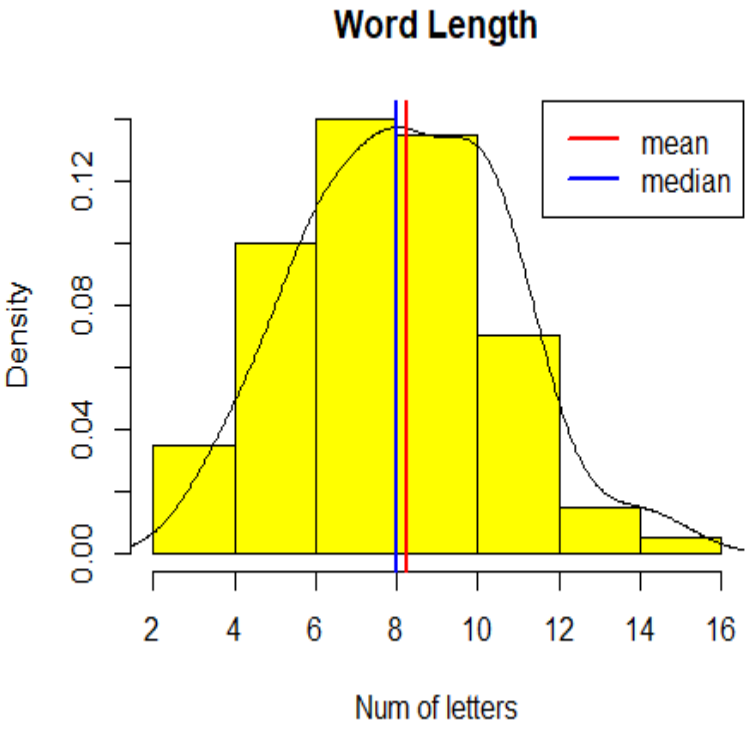
```
>hist(md, main="Word Length",
xlab="Num of letters",
col="yellow", prob=TRUE)
>lines(density(md), lwd=1.5)
>abline(v=mean(md), col="red",
lwd=2)
>abline(v=median(md), col="blue",
lwd=2)
```



Histogram (6)

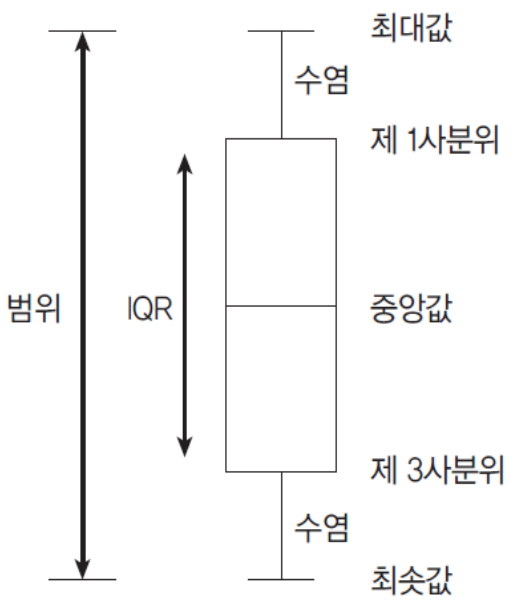
- Add legend

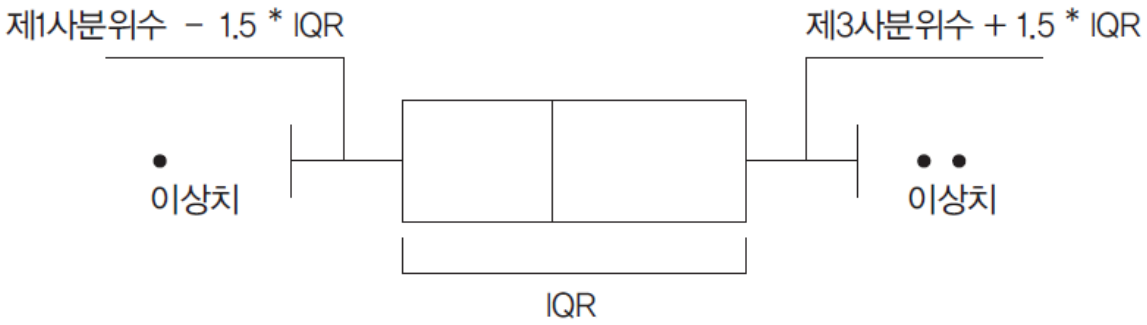
```
>hist(md, main="Word Length",
xlab="Num of letters", col="yellow",
prob=TRUE)
>lines(density(md), lwd=1.5)
>abline(v=mean(md), col="red", lwd=2)
>abline(v=median(md), col="blue",
lwd=2)
> legend(x="topright", c("mean",
"median"), col=c("red", "blue"),
lwd=c(2,2))
```



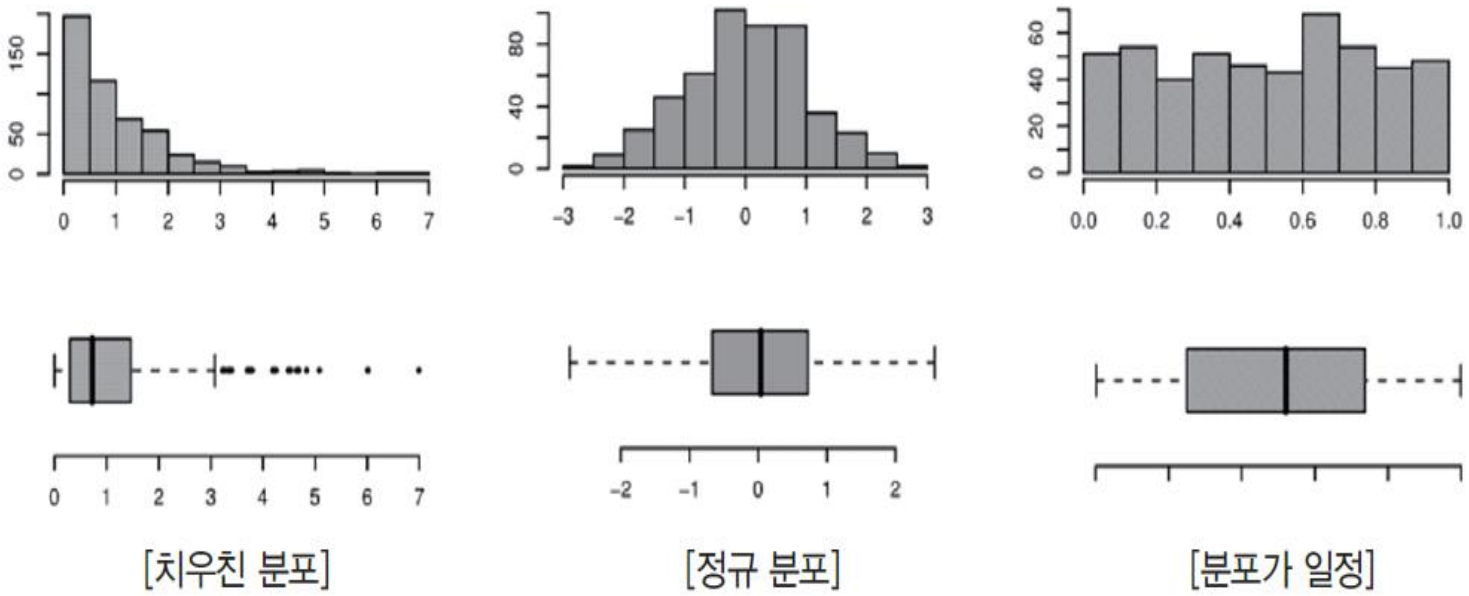
Box plot (1)

- Shows a variable's spread, symmetry, skewness and outliers





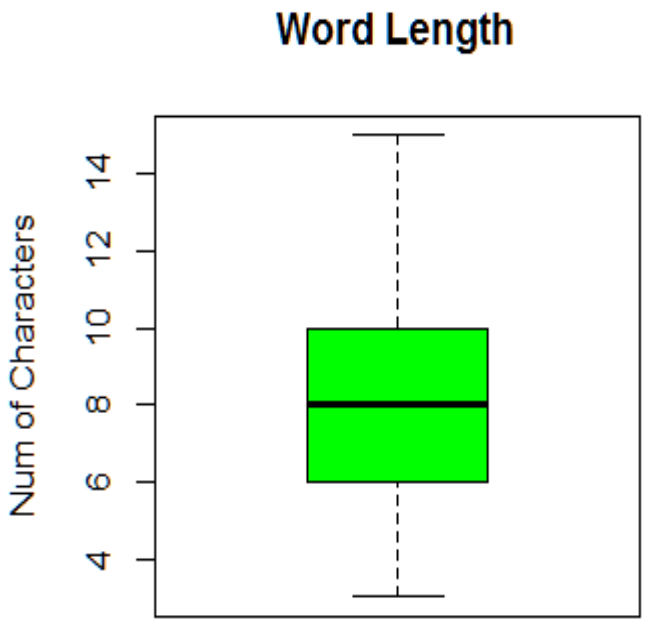
Box plot (2)


0 50 150
0 1 2 3 4 5 6 7
0 40 80
−3 −2 −1 0 1 2 3
0 20 40 60
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5 6 7
−2 −1 0 1 2
[치우친 분포]
[정규 분포]
[분포가 일정]

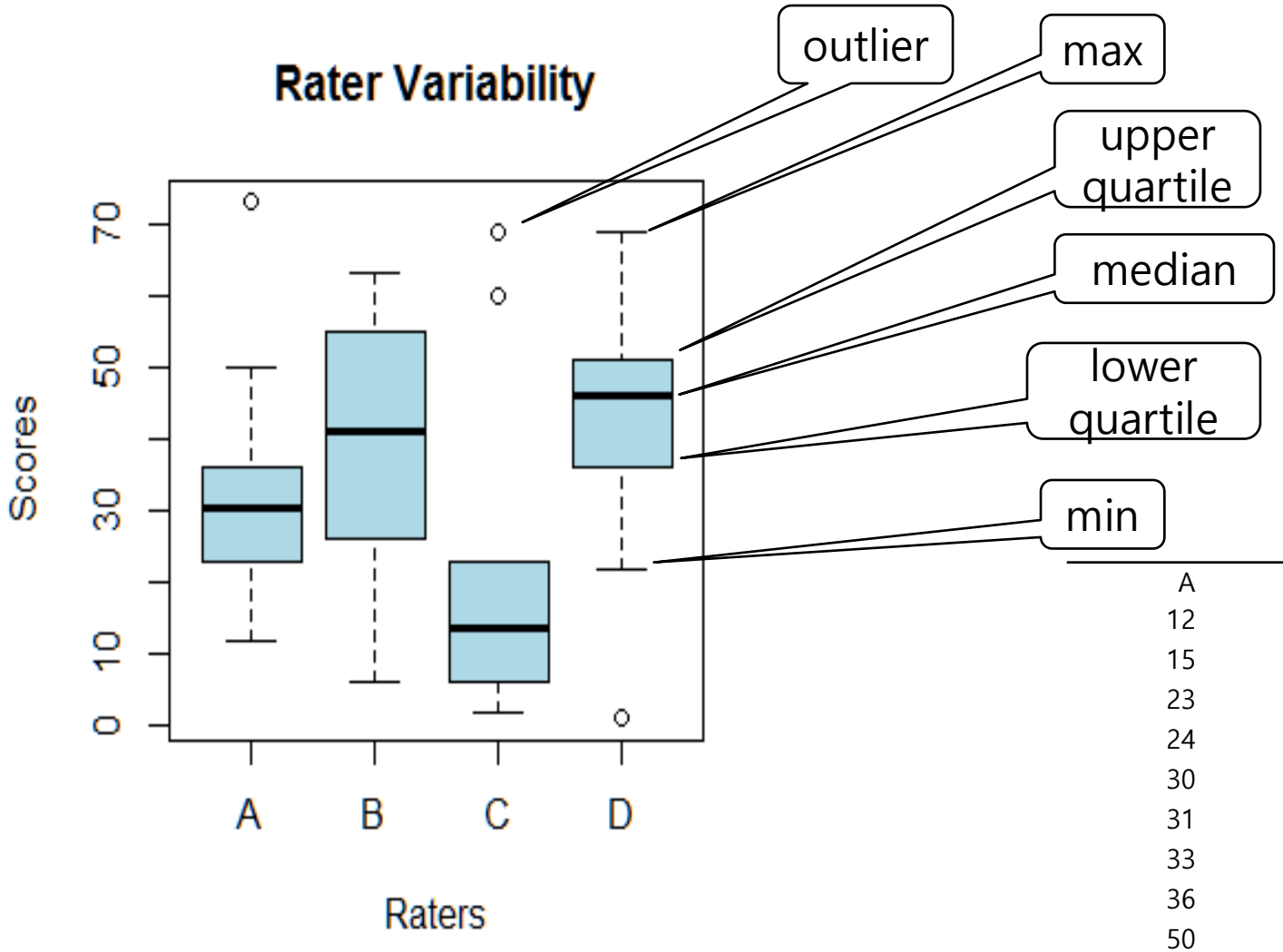
Box plot (3)

- Box plot

```
>boxplot(md,
main="Word Length",
ylab="Num of
Characters", col="green")
```



Box plot (4)



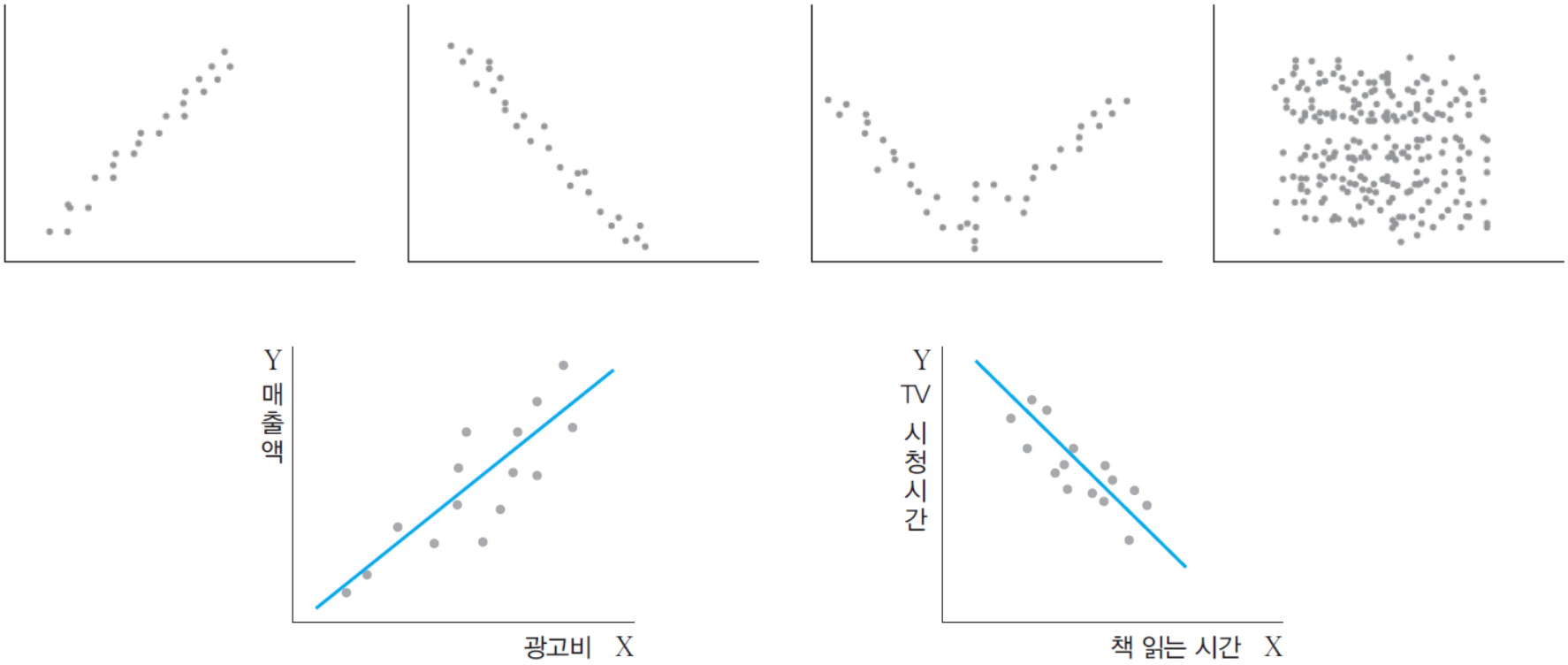
A	B	C	D
12	6	2	1
15	22	3	22
23	26	6	36
24	33	8	37
30	35	13	45
31	47	14	47
33	54	19	48
36	55	23	51
50	62	60	52
73	63	69	69

Task

- Using the data below, draw histograms and a box plot and compare them.

	A	B	C	D
Score 1	12	6	2	1
Score 2	15	22	3	22
Score 3	23	26	6	36
Score 4	24	33	8	37
Score 5	30	35	13	45
Score 6	31	47	14	47
Score 7	33	54	19	48
Score 8	36	55	23	51
Score 9	50	62	60	52
Score 10	73	63	69	69

Scatter diagram



- Correlation between two variables…
- Cover it more in the correlation chapter…

Graph into file

- Save a graph into a file
  - Use functions: jpeg(), bmp(), png(), tiff()
  - Example:

```
> jpeg(“filename.jpg”)
> hist(mydata)
> dev.off()
```