Statistika Deskriptif

`length()`,`min()`, `max()`, `sum()`, `prod()` dan `sort()`

```
lagu <- c(5.3,3.6,5.5,4.7,6.7,4.3,6.2,4.3,4.9,5.1,5.8,4.4)
lagu
```

1. 5.3
2. 3.6
3. 5.5
4. 4.7
5. 6.7
6. 4.3
7. 6.2
8. 4.3
9. 4.9
10. 5.1
11. 5.8
12. 4.4

```
length(lagu) # banyaknya elemen
```

12

```
max(lagu) # nilai terbesar
```

6.7

```
min(lagu) # nilai terkecil
```

3.6

```
sum(lagu) # total penjumlahan elemen di dalam vektor
```

60.8

```
prod(lagu) # total perkalian elemen di dalam vektor
```

241595726.162817

```
# mengurutkan vektor dari kecil ke besar
sort(lagu)
```

1. 3.6
2. 4.3
3. 4.3
4. 4.4
5. 4.7
6. 4.9
7. 5.1
8. 5.3
9. 5.5
10. 5.8
11. 6.2
12. 6.7

```
# mengurutkan vektor dari besar ke kecil
sort(lagu, decreasing=T)
```

1. 6.7
2. 6.2
3. 5.8
4. 5.5
5. 5.3
6. 5.1
7. 4.9
8. 4.7
9. 4.4
10. 4.3
11. 4.3
12. 3.6

Rata - rata

Jenis rata - rata:

- Rata - rata aritmatika (rata - rata)
- Rata - rata geometri
- Rata - rata harmonik

Rata - rata aritmatika

Rata - rata aritmatika dari suatu sampel adalah penjumlahan dari seluruh sampel, dibagi dengan ukuran sampel.

$$\bar{x} = \frac{1}{n}\sum*{i=1}^{n}x*{i}$$

```
lagu
```

1. 5.3
2. 3.6
3. 5.5
4. 4.7
5. 6.7
6. 4.3
7. 6.2
8. 4.3
9. 4.9
10. 5.1
11. 5.8
12. 4.4

```
rata2 <- sum(lagu) / length(lagu)
rata2
```

5.06666666666667

```
# pakai fungsi built-in
rata2 <- mean(lagu)
rata2
```

5.06666666666667

Rata - rata geometri

Rata - rata geometri didefinisikan sebagai akar ke-n dari perkalian seluruh sampel.

$$RG = \sqrt[n]{\prod*{i=1}^{n}x*{i}}$$

```
lagu
```

1. 5.3
2. 3.6
3. 5.5
4. 4.7
5. 6.7
6. 4.3
7. 6.2
8. 4.3
9. 4.9

10. 5.1
11. 5.8
12. 4.4

```
rata2geom <- prod(lagu)^(1/length(lagu))
rata2geom
```

4.99563581610903

```
# Cara yang lebih efisien:
rata2geom <- exp(mean(log(lagu)))
rata2geom
```

4.99563581610903

Aplikasi rata - rata geometri

Umum digunakan di dunia bisnis, misalnya:

- Perhitungan laju pertumbuhan.
- Perhitungan pengembalian portofolio keamanan.

Perhitungan *compounded annual growth rate*. Misalkan ada saham sebuah perusahaan dengan:

- Pertumbuhan sebesar 10 % pada tahun pertama (misalkan awalnya harga saham: $$100$):
  $$100 + 10 = 10$$
- Penurunan sebesar 20 % di tahun kedua:
  $$100 - 20 = 80$$
- Pada tahun ketiga pertumbuhan sebesar 30%:
  $$100 + 30 = 130$$

Laju pertumbuhan: $$\sqrt[3]{110 \times 80 \times 130} = 104,586$$

Karena dalam persen, maka:
104.586 - 100 = 4.58 % (laju pertumbuhannya)

```
saham <- c(100 + 10, 100 - 20, 100 + 30)
rg <- exp(mean(log(saham)))
rg
```

104.58643063512

```
rg - 100
```

4.58643063511965

Rata - rata harmonik

Rata - rata harmonik merupakan kebalikan dari rata - rata terbalik dari sampel:

$$\frac{1}{H} = \frac{1}{n}\sum*{i=1}^{n}\frac{1}{x*{i}}\text{ dengan } x_{i} > 0$$

```
rata2harm <- 1 / mean(1/lagu)
rata2harm
```

4.92500029031758

- Rata - rata harmonik digunakan untuk mencari hubungan perkalian atau pembagian antar pecahan.
- Rata - rata harmonik banyak digunakan untuk merata - ratakan suatu kelajuan.
- Di bidang finansial banyak digunakan untuk menghitung *price-earnings ratio*.

Median dan modus

Median: nilai tengah suatu sampel yang telah diurutkan.

- Pada sampel ganjil:
  $$x_{\frac{n + 1}{2}}$$
- Pada sampel genap:
  $$\frac{x*{\frac{n}{2}} + x*{\frac{n}{2}+1}}{2}$$

Modus merupakan nilai yang paling sering muncul di dalam suatu sampel.

- Suatu sampel dapat mempunyai sebuah modus, lebih dari satu modus, atau tidak mempunyai modus sama sekali.
- Modus dapat mempunyai nilai yang sama dengan rata - rata dan median.

```
sort(lagu)
```

1. 3.6
2. 4.3
3. 4.3
4. 4.4
5. 4.7
6. 4.9
7. 5.1
8. 5.3
9. 5.5
10. 5.8
11. 6.2
12. 6.7

```
median(lagu) # cara menghitung median
```

5

R tidak mempunyai fungsi untuk menghitung modus karena modus jarang digunakan untuk analisis statistik

Pencilan

```
gaji <- c(12,14,18,90,16,19,21) # gaji bulanan dalam juta rupiah
gaji
```

1. 12
2. 14
3. 18
4. 90
5. 16
6. 19
7. 21

```
mean(gaji)
```

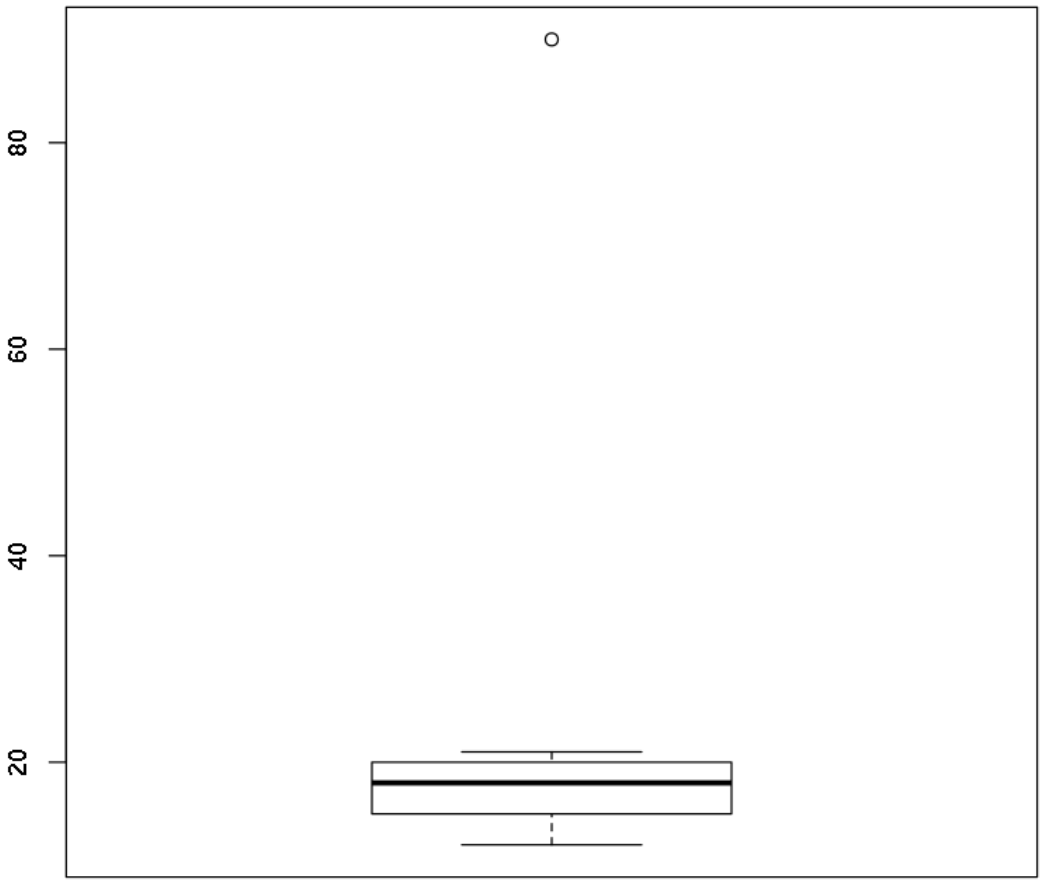
27.1428571428571

```
median(gaji)
```

18

Terdapat perbedaan yang sangat jauh antara rata - rata dan median.

```
boxplot(gaji)
```




```
mean(gaji, trim=0.1)
```

27.1428571428571

```
mean(gaji, trim=0.5)
```

18

```
mean(gaji, trim=0.2)
```

17.6

Kuartil dan kuantil

Kuartil merupakan istilah statistik yang digunakan untuk mendeskripsikan pembagian selang data ke dalam empat interval.

- Kuartil memisahkan data ke dalam tiga titik, yakni kuartil bawah, median, dan kuartil atas.
- Kuartil digunakan untuk menghitung jangkauan antar kuartil (IQR = Q1 - Q3) guna menghitung variabilitas di sekitar median.
- Setiap kuartil memuat 25% dari total data.

Untuk mencari letak kuartil, gunakan persamaan:
$$Q_{i} = \frac{i(n+1)}{4}\text{ dengan } i = 1,2,3$$

```
lagu
```

1. 5.3
2. 3.6
3. 5.5
4. 4.7
5. 6.7
6. 4.3
7. 6.2
8. 4.3
9. 4.9
10. 5.1
11. 5.8
12. 4.4

```
mean(lagu)
```

5.06666666666667

```
median(lagu)
```

5

```
summary(lagu) # digunakan untuk mencari sari data (termasuk kuartil)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.600   4.375   5.000   5.067   5.575   6.700 
```

```
# cara menghitung kuantil:
# sintaks: quantile(data, c(probabilitas1,probabilitas2))
quantile(lagu, c(.25, .75))
```

25%
: 4.375

75%

5.575

Varian dan standar deviasi

Digunakan untuk mengukur variabilitas dari suatu data atau akurasi dari parameter - parameter statistik.

- Varian: merupakan ukuran sebaran antar elemen di dalam sampel.
- Standar deviasi: merupakan akar kuadrat dari varian.

$$s^2 = \frac{\sum*{i=1}^{n} (\bar{x} - x{i})^{2}}{n-1}$$

```
lagu
```

1. 5.3
2. 3.6
3. 5.5
4. 4.7
5. 6.7
6. 4.3
7. 6.2
8. 4.3
9. 4.9
10. 5.1
11. 5.8
12. 4.4

```
var(lagu) # varian
```

0.787878787878788

```
sd(lagu) # std
```

0.887625364598595

Varian dan standar deviasi pada harga saham

```
library(dplyr)
```

```
Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

```
gedata <- read.csv("../data/GEStock.csv")
geprice <- select(gedata,Price)
```

```
ibmdata <- read.csv("../data/IBMStock.csv")
ibmprice <- select(ibmdata, Price)
```

```
var(geprice)
```

	Price
Price	575.6425

```
var(ibmprice) # lebih volatil ketimbang general electrics
```

	Price
Price	7712.717

```
sd(as.vector(geprice$Price))
```

23.992551305301

```
sd(as.vector(ibmprice$Price))
```

87.822078211186

Korelasi dan kovarian

- Korelasi merupakan metode statistik yang digunakan untuk mengukur derajat relasi antar dua variabel.

- Di dalam dunia finansial, korelasi dapat mengukur pergerakan harga saham.

Korelasi:
$$r = \frac{\sum (x-\bar{x}) (y-\bar{y})}{\sqrt{\sum (x-\bar{x})^{2}} \sqrt{(y-\bar{y})^{2}}}$$

Hasil korelai selalu berada di antara -1 hingga +1

- Korelasi positif menunjukkan bahwa tren kedua data berada pada arah yang sama.
- Korelasi negatif menunjukkan bahwa tren kedua data berada pada arah yang berbeda.
- Korelasi nol menunjukkan bahwa tidak ada hubungan antar tren kedua data.

```
x <- seq(10,50, by=10)
y <- x
```

```
cor(x,y) # karena data sama maka r = 1
```

1

```
y <- c(50,40,30,20,10)
```

```
cor(x,y) # terbalik r = -1
```

-1

```
x <- c(41,19,23,40,55)
y <- c(94,60,74,71,82)
cor(x,y)
```

0.64810840039477

```
gedates <- select(gedata, Date)
geprice <- select(gedata, Price)
ibmdates <- select(ibmdata, Date)
ibmprice <- select(ibmdata, Price)
```

```
cor(geprice,ibmprice) # secara default menggunakan korelasi pearson
```

	Price
Price	0.1098373

```
cor(geprice,ibmprice, use='complete.obs') # guna menangani nilai NaN
```

	Price
Price	0.1098373

```
cor(geprice, ibmprice, method = 'spearman') # korelasi spearman
```

	Price
Price	0.1665118

```
geprice_vec <- as.vector(geprice$Price)
ibmprice_vec <- as.vector(ibmprice$Price)
```

```
cor.test(geprice_vec,ibmprice_vec, method='pearson')
```

```
    Pearson's product-moment correlation

data:  geprice_vec and ibmprice_vec
t = 2.416, df = 478, p-value = 0.01607
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02053871 0.19739721
sample estimates:
      cor
0.1098373
```

```
cor.test(geprice_vec,ibmprice_vec, method='spearman')
```

```
    Spearman's rank correlation rho

data:  geprice_vec and ibmprice_vec
S = 15362788, p-value = 0.0002528
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1665118
```

```
cor.test(geprice_vec,ibmprice_vec, method='kendall')
```

```
    Kendall's rank correlation tau

data:  geprice_vec and ibmprice_vec
z = 3.9796, p-value = 6.902e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1215379
```

- Kovarian merupakan metode statistik yang digunakan untuk mengukur hubungan langsung antara kedua variabel.
- Ketika dua harga saham mempunyai perbedaan varian yang besar, maka kovariannya positif, pun begitu sebaliknya (kebalikan dari korelasi).
- Banyak digunakan pada teori potofolio modern di bidang finansial.

```
cov(geprice, ibmprice)
```

	Price
Price	231.4354

Contoh kasus perbandingan harga saham

```
cor(geprice,ibmprice)
```

	Price
Price	0.1098373

```
cov(geprice,ibmprice)
```

	Price
Price	231.4354

```
# mengimpor data saham cocacola
cocadata <- read.csv("../data/CocaColaStock.csv")
cocadates <- select(cocadata, Date)
cocaprice <- select(cocadata, Price)
```

```
cor(geprice, cocaprice)
```

	Price
Price	0.1775435

```
cov(geprice, cocaprice)
```

	Price
Price	107.2014

Dapat dikatakan harga saham di GE dan CocaCola secara komparatif lebih terhubung daripada GE dengan IBM.