Gema P Mindara

sebuah catatan kecil minggu ke-3...

P P D S

Prodi Manajemen Informatika, Sekolah Vokasi IPB

Covid-Bogor, September 2020

Daftar Isi

3 Statistik Deskriptif **1**
3.1 Ukuran Pemusatan 1
3.1.1 Ukuran Penyebaran 4
3.1.2 Distribusi Data 5
3.2 Visualisasi Data 5
3.2.1 Fungsi Plot 5
3.2.2 Histogram 6
3.2.3 Density 7
3.3 Fungsi Subset Data 7

Daftar Tabel

3.1 Fungsi Sederhana Ukuran Sentral Tendency 2

Daftar Gambar

BAB 3 | Statistik Deskriptif

> Statistik Deskriptif dan Visualisasi atau ekplorasi data secara visual dibutuhkan untuk memperoleh gambaran secara lengkap dan mendalam mengenai data yang dimiliki tanpa mebuat kesimpulan secara umum. Statistik deskriptif menggunakan prosedur numerik dan grafis dalam meringkas gugus data dengan cara yang jelas dan dapat dimengerti.Terdapat 3 (tiga karakteritik) utama untuk variabel tunggal, yakni: (1) Distribusi Data (Distribusi Frekuensi), (2) Ukuran Pemusatan (Central Tendency), (3)Ukuran Penyebaran (Dispersion). ...

R adalah bahasa pemrograman untuk analisis statistik dan grafik yang didistribusikan dibawah lisensi GNU Genaral Public License. R memberikan fleksibilitas dan kekuatan, konsisten yang mengintegasikan tool-tool untuk manipulasi data, analisis dan menampilkannya. Software R dapat didownload secara gratis di CRAN

Sebelum melakukan analisa data yang lebih mendalam, statistik deskriptif dan ekplorasi data secara visual merupakan langkah awal yang sangat penting untuk mendapat gambaran yang lebih lengkap dan mendalam tentang data yang kita miliki.

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Pengklasifikasian menjadi statistika deskriptif dan statistika inferensi dilakukan berdasarkan aktivitas yang dilakukan.[wiki].

Penyajian data dalam bentuk numerik terdiri dari beberapa bentuk antara lain:

1. Ukuran pemusatan data (Central Tendency),
2. Ukuran penyebaran data,
3. Tingkat kemiringan data (Skewness) dan
4. Tingkat keruncingan data** (Kurtosis).

3.1 Ukuran Pemusatan

Salah satu aspek yang paling penting untuk menggambarkan distribusi data adalah nilai pusat data pengamatan (Ukuran Pemusatan, Tendensi Sentral). Ukuran Pemusatan Data adalah Nilai yang digunakan untuk menjelaskan pusat dari sekelompok data yang dapat mewakili data secara keseluruhan.

Setiap pengukuran aritmatika yang ditujukan untuk menggambarkan suatu nilai yang mewakili nilai pusat atau nilai sentral dari suatu gugus data (himpunan pengamatan) dikenal sebagai ukuran tendensi sentral (Nilai tunggal yang dapat mewakili seluruh data). Terdapat tiga ukuran pemusatan data yang sering digunakan, yaitu:

1. **Mean** (Rata-rata hitung/rata-rata aritmetika),
2. **Median** (Nilai Tengah), dan
3. **Mode /Modus** (Nilai yg paling sering muncul).

Selain angka-angka statistik di atas ada pula nilai statistik lainnya, yaitu minimum, kuartil pertama, nilai tengah (median), kuartil ke tiga dan maksimum. Nilai-nilai tersebut ditampilkan dengan menjalankan *fungsi summary()*. Fungi-fungsi di R untuk analisa data yang berkaitan dengan sentral tendency adalah sebagai berikut:

Tabel 3.1: Fungsi Sederhana Ukuran Sentral Tendency

Fungsi	Kegunaan
sum(x)	Jumlah dari elemen x
max(x)	Nilai Maksimum
min(x)	Nilai Minimum
which.min(x)	Urutan data terkecil
which.max(x)	Urutan data terbesar
range(x)	Rentang(max - min)
length(x)	Banyaknya elemen x
mean(x)	Rata-rata elemen x
var(x)	Variasi dari elemen x
cor(x,y)	Korelasi antara x, y
summary(x)	Lima angka statistik

Latihan Lab-1:

Instruksi :

1. Jalankan R
2. Ketikkan di console perintah berikut ini. Perhatikan, catat dan berikan penjelasan secukupnya.

```
>data<c(8,2,7,1,2,9,8,2,10,9)
>hist(data)
>boxplot(data)
>sum(data)/ length(data
>mean(data))
>median(data)
>sort(data)
>table(data)
>y<-table(data)
>names(y)[which(y==max())] //mode
>names(table(data))[which(table(data)==max(table(data))]
```

Latihan Lab-2:

Latihan berikut, kita menggunakan dataset *mtcars* yang tersedia di R. Kemudian dari data ini kita mencoba menghitung ukuran pemusatan dari data-data tersebut.
Instruksi :

1. Jalankan R
2. Ketikkan di console perintah berikut ini. Perhatikan, catat dan berikan penjelasan secukupnya.

```
>data(mtcars) #load dataset mtcars
>dim(mtcars)
>names(mtcars)
>head(mtcars)
>head(mtcars,11)
>x1<-(mtcars$wt)
>mean(x1)
>median(x1)
>x11<-table(x1)
>x11
# mencari modus:
>names(x11)[which(x11==max(x11))]
atau:
>names(table(x1))[table(x1)==max(table(x1))]
```

Latihan Lab-3:

Latihan berikut, kita menggunakan dataset *airquality* yang tersedia di R. Kemudian dari data ini kita mencoba menghitung ukuran pemusatan dari data-data tersebut.
Instruksi :

1. Jalankan R
2. Ketikkan di console perintah berikut ini. Perhatikan, catat dan berikan penjelasan secukupnya.

```
>data(airquality)
>dim(airquality)
>names(airquality)
>str(airquality)
>airquality$Ozone
>airquality$Solar.R
>y<-airquality$Solar.R
>table(is.na(y))
>mean(y)
>mean(y,na.rm=TRUE)
>median(y)
>median(y,na.rm=TRUE)
MODUS
>sort(table(y))
names(table(y))[table(y)==max(table(y))]
```

3.1.1 Ukuran Penyebaran

Ukuran penyebaran data adalah suatu ukuran yang menyatakan seberapa besar nilai nilai data berbeda atau bervariasi dengan nilai ukuran pusatnya atau seberapa besar penyimpangan nilai nilai data dengan nilai pusatnya.

1. Range (Rentang)
   Adalah selisih antara nilai maksimum dan nilai inimum yang terdapat dalam data. Range merupakan ukuran yang paling sederhana ketika kita mencoba menjelaskan data.
2. Interquartil Range adalah selisih dari kuartil 3 dan 1 yang dirumuskan:

$$IQR = Q_3 - Q_1$$

3. Ragam atau Varians yang dirumuskan dengan :

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

4. Standar Deviasi (Simpangan Baku) $s = \sqrt{s^2}$. Ragam dan simpangan baku adalah ukuran penyebara/ disperse yang paling sering digunakan karena memiliki persamaan matematika yang penting.
5. Koefisien variasi, KV. KV digunakan untuk membandingkan variasi data, apabila satuan pengukuran dari variabel-variabel yang diukur berbeda satu sama lain (misalnya tinggi badan dalam cm dan berat badan dalam kg). KV dirumuskan :

$$KV = \frac{s}{\bar{x}} \times 100\%$$

Bagaimana dengan R apakah menyediakan fungsi untuk ukuran ukuran di atas? R secara default telah menyediakan berbagai penyebaran seperti rentang, standard deviasi, dan juga varians.

Latihan Lab-4:

Latihan berikut, kita menggunakan dataset *state* yang tersedia di R. Kemudian dari data ini kita mencoba menghitung ukuran pemusatan dari data-data tersebut.
Instruksi :

1. Jalankan R
2. Ketikkan di console perintah berikut ini. Perhatikan, catat dan berikan penjelasan secukupnya.

```
>data(mtcars)
>head(mtcars)
>tail(mtcars)
>dim(mtcars)
>colnames(mtcars)
>median(mtcars$mpg)
>quantile(mtcars$mpg,c(0.25,0.50,0.75))
>quantile(mtcars$mpg,c(0.20))  #desil ke-2
```

```
>quantile(mtcars$mpg,c(0.99)) #persentil ke-99
>sd(mtcars$mpg)
>var(mtcars$mpg)
>max(mtcars$mpg)
>range(mtcars$mpg)
>summary(mtcars$mpg)
```

3.1.2 Distribusi Data

Aspek penting dari "deskripsi" suatu variabel adalah bentuk distribusinya, yang menunjukkan frekuensi dari berbagai selang nilai variabel. Statistik deskriptif sederhana dapat memberikan beberapa informasi yang relevan dengan masalah ini. Sebagai contoh, jika **skewness** (kemiringan), yang mengukur kesimetrisan distribusi data, tidak sama dengan 0, maka distribusi dikatakan tidak simetris (a simetris), dan apabila skewness bernilai 0 berarti data tersebut berdistribusi normal (simetris).

Ukuran kemiringan atau skewness merupakan suatu nilai yang mengukur ketidaksimetrisan distribusi data. Suatu data dikatakan berdistribusi simetris sempurna bila nilai *rata-rata, median, dan modus* dalam data adalah sama.*(Distribusi normal merupakan distribusi yang simetris dan nilai skewness adalah 0)*.

1. Skewness yang bernilai positif menunjukkan ujung dari kecondongan menjulur ke arah nilai positif (ekor kurva sebelah kanan lebih panjang, nilai kemiringan > 0).
2. Skewness yang bernilai negatif menunjukkan ujung dari kecondongan menjulur ke arah nilai negatif (ekor kurva sebelah kiri lebih panjang, nilai kemiringan < 0.

Kurtosis menggambarkan keruncingan (peakedness atau kerataan flatness) suatu distibusi data dibandingkan dengan distribusi normal.

1. Pada distribusi normal, nilai kurtosis sama dengan 0.
2. Nilai kurtosis yang positif menunjukkan distribusi yang relatif runcing, sedangkan nilai kurtosis yang negatif menunjukkan distribusi yang relatif rata.

Disimpulkan bahwa: *Skewness dan kurtosis merupakan dua alat ukur dalam menelusuri distribusi data yang diperbandingkan dengan distribusi normal.*

3.2 Visualisasi Data

3.2.1 Fungsi Plot

Fungsi plot() untuk membuat titik dari nilai sumbu x dan sumbu y. Sintaks umum fungsi plot(). Syntax: *plot(x,y)*,
Parameter x dan y berisi nilai-nilai numerik.

Bentuk grafik sangat standart, sehingga perlu dilengkapi dengan beberapa keterangan tambahan untuk memperjelas dan mempermudah dalam melakukan interpretasi grafik. Hal ini dapat dilakukan dengan menambahkan fitur warna atau simbol dalam tampilan grafik. Untuk hal tersebut, R mempunyai fasilitas pewarnaan (yaitu dengan

argumen col), simbol (dengan argumen pch), ukuran (dengan argumen cex), label/nama sumbu kordinat (dengan argumen xlab dan ylab), judul grafik (dengan argumen main)

Contoh:

```
>x<-mtcars$mpg
>y<-mtcars$hp
>plot(x,y)
>plot(x,y,xlab="Miles/(US) gallon",ylab="Gross horsepower")
>plot(x,y,xlab="Miles/(US) gallon",ylab="Gross horsepower",\
  main="Grafik MPG vs HP")
```

catatan type simbol untuk plot:

1. "p" - points(default)
2. "l" - lines
3. "b" - both points and lines
4. "c" - empty points joined by lines
5. "o" - overplotted points and lines
6. "s" and "S" - stair steps
7. "h" - histogram-like vertical lines
8. "n" - does not produce any points or lines

```
>x<- seq(-pi,pi,0.1)
>plot(x, sin(x))
>plot(x, sin(x),main="Grafik Fungsi Sinus",ylab="sin(x)", xlab="x")
>plot(x, sin(x),main="Grafik Fungsi Sinus",ylab="sin(x)", \
   xlab="x",type="l",col="blue")

*dua grafik overlap*

>plot(x, sin(x),main="Overlaying Graphs", \
  ylab="",type="l",col="blue")
>lines(x,cos(x), col="red")
>legend("topleft",c("sin(x)","cos(x)"),fill=c("blue","red"))
```

3.2.2 Histogram

Selain plot, bentuk representasi grafis lainnya yang paling mudah digunakan untuk menggambarkan sebaran data adalah histogram. R menyediakan fasilitas fungsi histogram yang digunakan untuk mengetahui sebaran sampel suatu data. Sebagai catatan: *histogram ataupun boxplot, digunakan untuk satu variable*. Grafik ini dapat memberikan informasi frekuensi distribusi data. Sintaks fungsi hist() adalah: **hist(objek)**

```
>head(mtcars)
>hist(mtcars[,1])
```

```
>hist(mtcars$mpg)
>hist(mtcars$mpg,ylab="Frekuensi",xlab="galon per mile")
>hist(mtcars$mpg,ylab="Frekuensi",xlab="galon per mile", \
  main="Histogram Variabel MPG")
```

3.2.3 Density

Fungsi density() berguna untuk melakukan estimasi kernel density. Syntax:**plot(density(objek))**. Contoh :

```
>plot(density(mtcars$mpg))
```

3.3 Fungsi Subset Data

Dalam matematika, terutama teori himpunan, suatu himpunan A adalah himpunan bagian atau subset dari himpunan B bila A "termuat" di dalam B. A dan B boleh jadi merupakan himpunan yang sama. Hubungan suatu himpunan yang menjadi himpunan bagian yang lain disebut sebagai "termasuk ke dalam" atau kadang-kadang "pemuatan". Himpunan B adalah superhimpunan atau superset dari A karena semua elemen A juga adalah elemen B.wiki.

Apa kaitannya dengan R yang kita bahas saat ini?. Pada saat kita mempunyai data yang cukup banyak dan kadangkala kita hanya menginginkan sebagian saja yang dianalisis. Platform R telah memiliki lebih 100 dataset (*built-in-dataset*) yang dapat digunakan untuk latihan. Fungsi data() digunakan untuk melihat daftar dataset tersebut. Beberapa diantara paket tersebut masih memerlukan pemanggilan secara eksplisit dengan menggunakan fungsi data().Fungsi data() ini memuat satu data frame dengan menspesifikasikan di argumen. Sebagai contoh dataset "mtcars" pada paket data *built-in-dataset* yang terdiri dari 32 baris x 11 kolom.

Platform R telah memiliki lebih 100 dataset yang dapat digunakan untuk latihan. Fungsi data() digunakan untuk melihat daftar dataset tersebut. Ketik fungsi data()

Langkah Praktikum 1:

Sebagai langkah awal jalankan paket R Anda dan ketikkan di prompt sebagai berikut :

```
>data()  \\load dataset
>data(mtcars)  \\load dataset mtcars di workspace
>data()
>help(mtcars)
>str(mtcars)
>nrow(mtcars)
>ncol(mtcars)
>dim(mtcars)
```

Ketarangan: 1. Fungsi nrow(objek) digunakan untuk menghitung jumlah baris yang dimiliki oleh obyek yang menyimpan data. 2. Fungsi ncol(objek) digunakan untuk menghitung jumlah kolom yang dimiliki oleh obyek yang menyimpan data. 3. Fungsi dim(objek) digunakan untuk melihat dimensi atau ukuran dimensi data (banyaknya baris dan banaknya kolom)

Tugas Anda:

1. Catat keluaran di layar monitor anda.
2. Diskusikan hasilnya kemudian berikan penjelasan diskriptif singkat.

Mari kita dalami dataset ini lebih lanjut:

Langkah Praktikum 2:

```
>colnames(mtcars)
>rownames(mtcars)
>rownames(mtcars)[]
>names(mtcars)
>head(mtcars)
>tail(mtcars)
>head(mtcars,5)
>tail(mtcars,5)
```

Kesinpulan:

1. Fungsi colnames() atau names() untuk menampilkan nama-nama kolom atau variabel pada data frame atau dataset.
2. Fungsi rownames() untuk menampilkan nama baris pada data frame atau dataset.
3. Sedangkan untuk melihat nama pada suatu baris tertentu digunakan rownames(objek)[indek]
4. Fungsi head() dan tail () untuk menampilkan beberapa data teratas dan beberapa data terbawah. Biasanya menggunakan syntax sbb:
   head(objek,jumlBaris) dan tail(objek,jumlBaris).

Tugas Anda:

1. Catat keluaran di layar monitor anda.
2. Diskusikan hasilnya kemudian berikan penjelasan diskriptif singkat.

Bagaimana jika saya hanya ingin melihat satu atau beberapa variabel saja (tidak semua) ?

Langkah Praktikum 3:

```
>mtcars$mpg
>mtcars$hp
>mtcars$gear
>mtcars[2:4]
>mtcars[c(2,4)]
>mtcars[-1]
>mtcars[-1,]
>mtcars$mpg>20
```

Tugas Anda: 1. Catat hasilnya dan berikan analisa singkat.

Langkah Praktikum 4:

```
>mtcars$mpg > 20
```

Apa yang Anda dapatkan ketika mengetikkan *"mtcars$mpg>20?"* tentunya tidak sesuai dengan Anda harapkan bukan ?
Ekspresi logika! Hasilnya hanya memberikan dua kemungkinan yakni 0 atau 1 (TRUE atau FALSE). Jika dibaca secara sederhana perintah di R :
"Apakah elemen elemen dari variabel mpg pada dataset mtcars lebih besar dari 20?"

```
>mtcars[mtcars$mpg>20,]
```

//memberitahu R untuk mendapatkan semua baris dari mtcars dimana nilai variabel mpg>20. "Cari dan tampilkan elemen elemen dataset mtcars yang nilai mpgnya lebih besar dari 20" Pertanyaan: *Keluaran data diatas terlalu banyak. Saya hanya menginginkan untuk kolom tertentu saja, misalnya variabel mpg dan hp.*

```
>mtcars[mtcars$mpg>20,c("mpg","hp")]
>mtcars[mtcars$mpg>20,c(1,4)]
>mtcars[which(mtcars$mpg>20),]
```

Keterangan:

1. Untuk memfilter data berdasarkan suatu nilai pada sebuah kolom dapat digunakan fungsi which(). Syntax: objek[which()]

langkah-langkah diatas merupakan langkah **filtering** data atau mencari subset dari sekumpulan data. Adakah cara lain untuk ini? R memberikan alternatif untuk penyaringan data, yakni dengan fungsi **subset()**. Fungsi ini dapat digunakan baik untuk struktur data vector, list maupun dataframe. Fromat fungsi subset adalah sebagai berikut:

$$***\text{subset}dataset, syaratlogikbaris, kolomPilihan***$$

```
>subset(mtcars,mpg>20,c("mpg","hp"))
>subset(mtcars,mpg==max(mpg))
>subset(mtcars,mpg==max(mpg),mpg)
>mtcars_a1<-subset(mtcars,mpg>30)
>mtcars_a2<-subset(mtcars,mpg>30,c("mpg","cyl","gear"))
>mtcars_a3<-subset(mtcars,pg>30,select=c(mpg,cyl,gear))
```

Mudah bukan ?