

Materi Kuliah Sebelum UTS

Konsep Dasar Analisis Data Eksploratif

Diagram Stem & Leaf untuk penyederhanaan Data

Ukuran Pemusatan Data :

Mean, Modus, Median, Trirata, Rata-Rata Tengah

Ukuran Persebaran Data :

Range, Sebaran Tengah, Simpangan Rata-rata, Variansi & Standar Deviasi

Diagram Boxplot & Standardisasi

Transformasi Angkatan

Materi Kuliah Setelah UTS

Distribusi Normal, t dan F

Statistika Konfirmasi : Uji Hipotesis

Regresi Eksplorasi & Konfirmasi :

estimasi koefisien regresi, smoothing, koefisien korelasi r^2

Analisis Data Kategorik :

Data cacah table kategorik (Uji Homogenitas & Independensi)

Capaian Pembelajaran

CP MK	CAPAIAN PEMBELAJARAN	BOBOT PENILAIAN
M ₁	Mahasiswa mampu menjelaskan konsep analisis data eksploratif , membuat ringkasan numerik, standarisasi & transformasi data	20
M ₂	Mahasiswa mampu menerapkan analisis eksploratif & konfirmasi untuk single batch, multiple batch dan regresi	20
M ₃	Mahasiswa mampu menerapkan konsep analisis data eksploratif dalam bidang sains data	20
UTS	CP M ₁	20
UAS	CP M ₂ & M ₃	20

Pendahuluan

Exploratory *Data Analysis* (EDA) atau dikenal pula dengan analisis data eksploratif merupakan pendekatan analisis untuk suatu data guna membuat gambaran keseluruhan (*summary*) data sehingga mudah untuk dipahami.



Pendahuluan

Metode analisis ini menyediakan berbagai alat untuk meringkas dan memperoleh wawasan tentang sekumpulan data dengan cepat menggunakan grafik sebagai bentuk visualisasi data, tanpa menggunakan model statistik, atau formulasi hipotesis.



Pendahuluan

- Analisis data eksploratif sangat penting dalam menelaah dan menemukan karakteristik data yang selanjutnya dapat berguna dalam pemilihan model statistika yang tepat.
- *Exploratory Data Analysis* pertama kali diperkenalkan oleh John Tukey pada tahun 1977. Tukey menyarankan membuat visualisasi untuk pemeriksaan data sebelum membuat formulasi hipotesis yang dapat diuji pada data set tersebut atau pada data set yang akan dikoleksi selanjutnya



Contoh Eksplorasi Data

Tabel 1. Angka Kematian
karena Penyakit Jantung per
100.000 orang Tahun 1971
di Jogja & Jateng

Kabupaten	JK	Usia				
		20-29	30-39	40-49	50-59	60-69
Bantul	L	23,7	30,1	39,4	45,5	56,2
	P	9,7	14,4	22,6	26,6	26,9
Sleman	L	7,2	7,8	19,0	28,5	37,5
	P	7,4	4,2	10,5	13,0	22,1
Brebes	L	20,3	19,2	26,3	38,8	71,4
	P	13,2	11,2	37,1	58,2	61,4
Tegal	L	29,2	39,4	44,0	51,8	54,4
	P	9,2	15,6	30,2	42,9	49,0
Kendal	L	15,1	26,7	33,8	40,3	54,5
	P	7,5	9,8	14,0	20,5	23,8
Batang	L	27,4	35,3	49,5	54,1	58,2
	P	13,2	14,2	19,8	23,8	32,7
Cilacap	L	42,7	66,4	82,3	82,6	89,9
	P	13,5	17,9	32,9	45,0	46,2

- Pada Tabel 1 diperoleh informasi bahwa Tingkat Kematian karena penyakit jantung untuk jenis kelamin laki-laki lebih besar dari perempuan
- Untuk mengetahui pengaruh usia, angka kematian di DIY dan Jateng akan lebih baik jika melihat satu kelompok jenis kelamin dulu, misal laki-laki.
- Dari kelompok laki-laki ini, dapat diketahui apakah usia tua mempunyai tingkat kematian karena penyakit jantung lebih tinggi dari yang usia muda

Menyederhanakan Angka & Daftar Tally

- Dengan melakukan pemilihan Sebagian angka dan memusatkan perhatian terhadapnya, analisis data eksploratif telah dimulai.
- **Data/Angkatan (batch)** adalah kumpulan angka yang sejenis atau saling berhubungan.
- Contoh:
 - a. tingkat kematian karena penyakit jantung laki-laki kelompok usia 20-29 tahun untuk semua kota (1 Angkatan)
 - b. Bila dipandang tingkat kematian karena penyakit jantung laki-laki kelompok usia 20-29 tahun dan 30 – 39 tahun untuk semua kota (2 Angkatan)

Kabupaten	Laki-laki 20-29 Tahun		Laki-laki 20-29 Tahun	
	Asli	Pembulatan	Asli	Pembulatan
Bantul	23,7	24	30,1	30
Sleman	7,2	7	7,8	8
Brebes	20,3	20	19,2	19
Tegal	29,2	29	39,4	39
Kendal	15,1	15	26,7	27
Batang	27,4	27	35,3	35
Cilacap	42,7	43	66,4	66

- Aturan penyederhanaan:
 - a. Angka decimal $< 0,5$
dibulatkan ke bawah
 - b. Angka decimal $\geq 0,5$
dibulatkan ke atas

- Data disamping adalah angka kematian karena penyakit jantung per 100000 orang laki-laki untuk kelompok usia tertentu dari berbagai kota.
- Artinya, kota disebut sebagai satuan pengamatan (sesuatu yang diamati untuk mendapatkan angka tersebut)

Kabupaten	Laki-laki 20-29 Tahun		Laki-laki 20-29 Tahun	
	Asli	Pembulatan	Asli	Pembulatan
Bantul	23,7	2	30,1	3
Sleman	7,2	1	7,8	1
Brebes	20,3	2	19,2	1
Tegal	29,2	2	39,4	3
Kendal	15,1	1	26,7	2
Batang	27,4	2	35,3	3
Cilacap	42,7	4	66,4	6

Aturan penyederhanaan:

mengubah angka kematian menjadi 10000 orang maka 23,7 dibulatkan menjadi 2; 7,2 dibulatkan menjadi 1.

- Pengambilan penyederhanaan yang berlebihan menyebabkan banyak angka menjadi sama besar.
- Pembulatan tergantung pada pertimbangan yang diambil oleh penganalisis data berdasarkan ketelitian angka semua, sebaran data dalam Angkatan dan fungsi dari angka-angka tersebut.

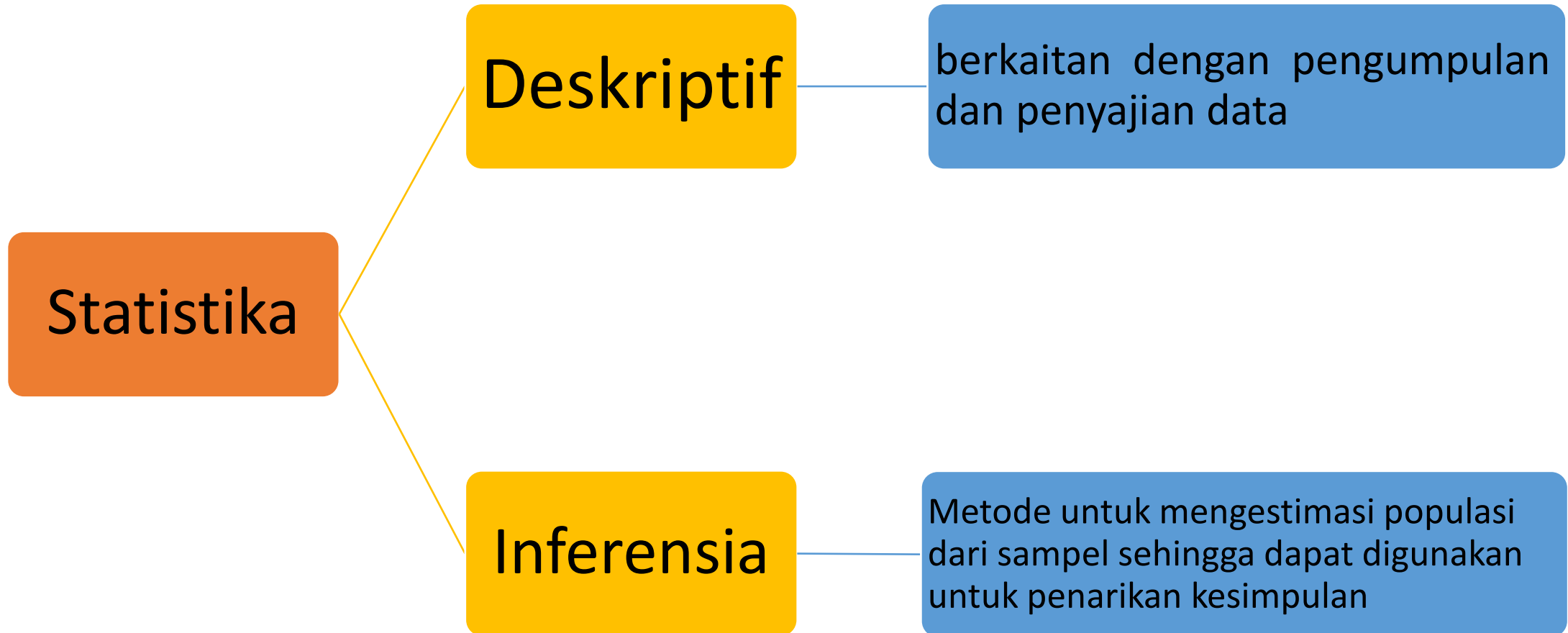


TALLY/ TURUS

Laki-laki 20-29 Tahun		Laki-laki 20-29 Tahun	
Rentang	Tally/Turus	Rentang	Tally/Turus
0-9	I	0-9	I
10-19	I	10-19	I
20-29	IIII	20-29	I
30-39		30-39	III
40-49	I	40-49	
50-59		50-59	
60-69		60-69	I

Konsep Dasar Dalam Statistika

- Statistika umumnya bekerja dengan data numerik yang berupa hasil cacahan ataupun hasil pengukuran, atau dengan data kategorik yang diklasifikasikan menurut kriteria tertentu.
- Informasi yang tercatat dan terkumpul, baik numerik dan kategorik disebut pengamatan.
- Metode statistika yaitu prosedur yang dipakai dalam pengumpulan, penyajian, analisis, dan penafsiran data



Statistika Deskriptif

Bar chart, pie chart, stem & leaf, boxplot, histogram, dsb

Ukuran pemusatan data, ukuran sebaran data, skewness, keruncingan kurva, dsb

Statistika Inferensia

Probabilitas, distribusi peluang, sampling & distribusi sampling, dsb

Uji hipotesis, Anava, Analisis korelasi, analisis regresi, dsb

Data

Data adalah sekumpulan informasi atau fakta yang dapat diolah untuk analisis lebih lanjut.

Contoh: Data Tingkat Kematian di Jawa Tengah, Data Gambar, Data Suara, hasil eksperimen di Lab, dsb.

Jenis Data

- Berdasarkan sifatnya, Data dibagi menjadi 2 yaitu:

❑ Data Kualitatif

- data yang tidak dapat dinyatakan dalam angka/numerik
- ex. Agama, alamat tinggal, jenjang pendidikan, dsb

❑ Data Kuantitatif

- data yang dapat dinyatakan dalam angka/numerik
- ex. Tinggi badan, penghasilan, berat badan, dsb

DATA KUANTITATIF

```
graph LR; A[DATA KUANTITATIF] --> B[DISKRIT]; A --> C[KONTINU]; B --> D["a. Jam kerja dalam sehari,  
b. banyak telur ayam yang dihasilkan ayam,  
c. banyak hari libur dalam 1 bulan,dst."]; C --> E["Suhu udara di berbagai tempat berada pada kisaran -20°C sampai dengan 30°C"];
```

DISKRIT

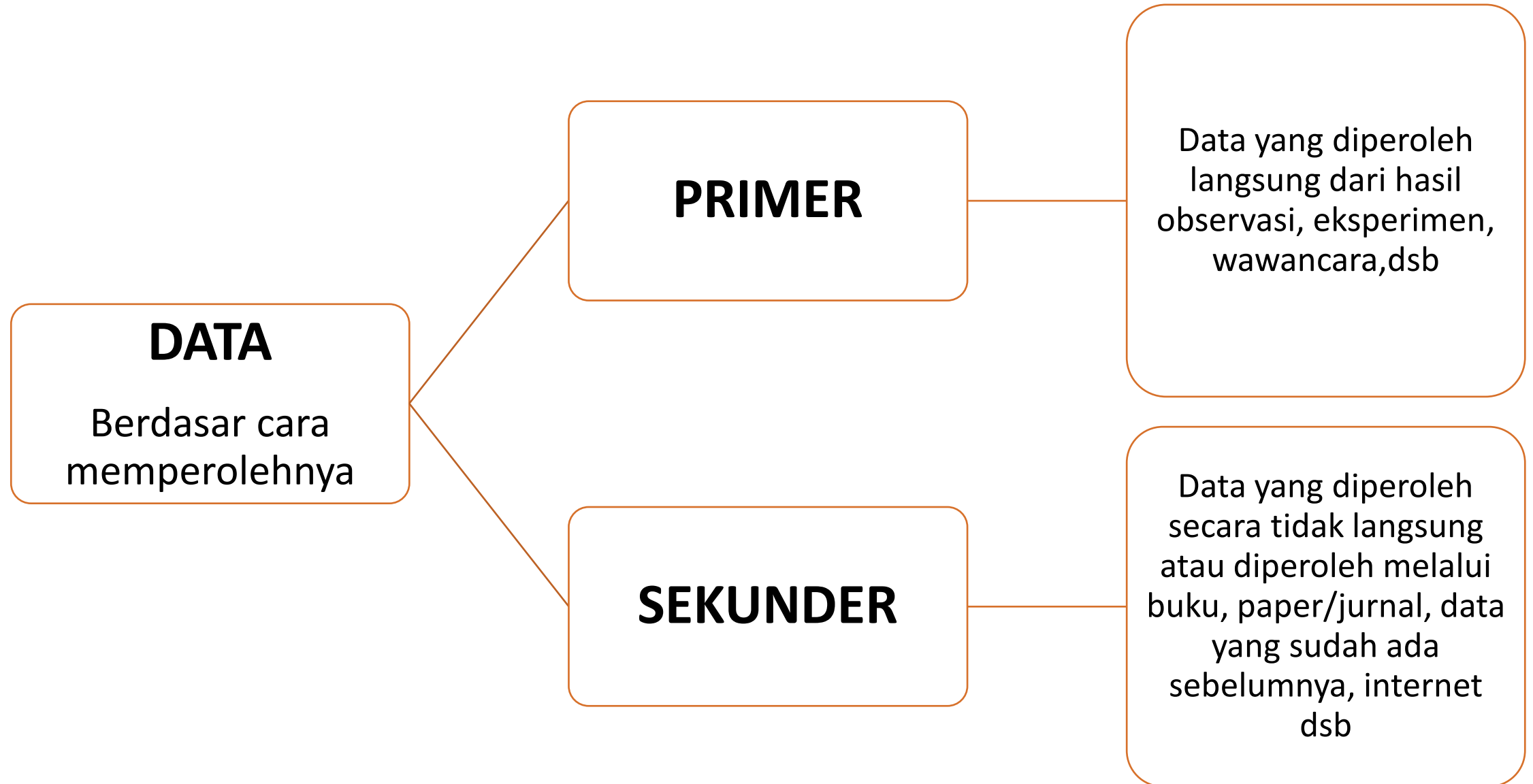
Data yang banyaknya berhingga atau terhitung

- a. Jam kerja dalam sehari,
- b. banyak telur ayam yang dihasilkan ayam,
- c. banyak hari libur dalam 1 bulan,dst.

KONTINU

Data yang banyaknya tak berhingga

Suhu udara di berbagai tempat berada pada kisaran -20°C sampai dengan 30°C



Skala Pengukuran/ Tipe Data

SKALA NOMINAL

Data yang bersifat kategorik dan tidak dapat di rangking.

Jenis Kelamin, pilihan jawaban ya/tidak pada kuesioner,dll.

SKALA ORDINAL

Data yang dapat dirangking , tetapi selisihnya tidak bermakna apapun.

nilai statistika mahasiswa UTY adalah A-, A, B, B-, C, C+, D, A

SKALA INTERVAL

Data yang dapat dirangking tetapi tidak mempunyai "titik nol" yang tetap sebagai awal.

suhu tubuh : suhu antara 36°C dan 37°C.

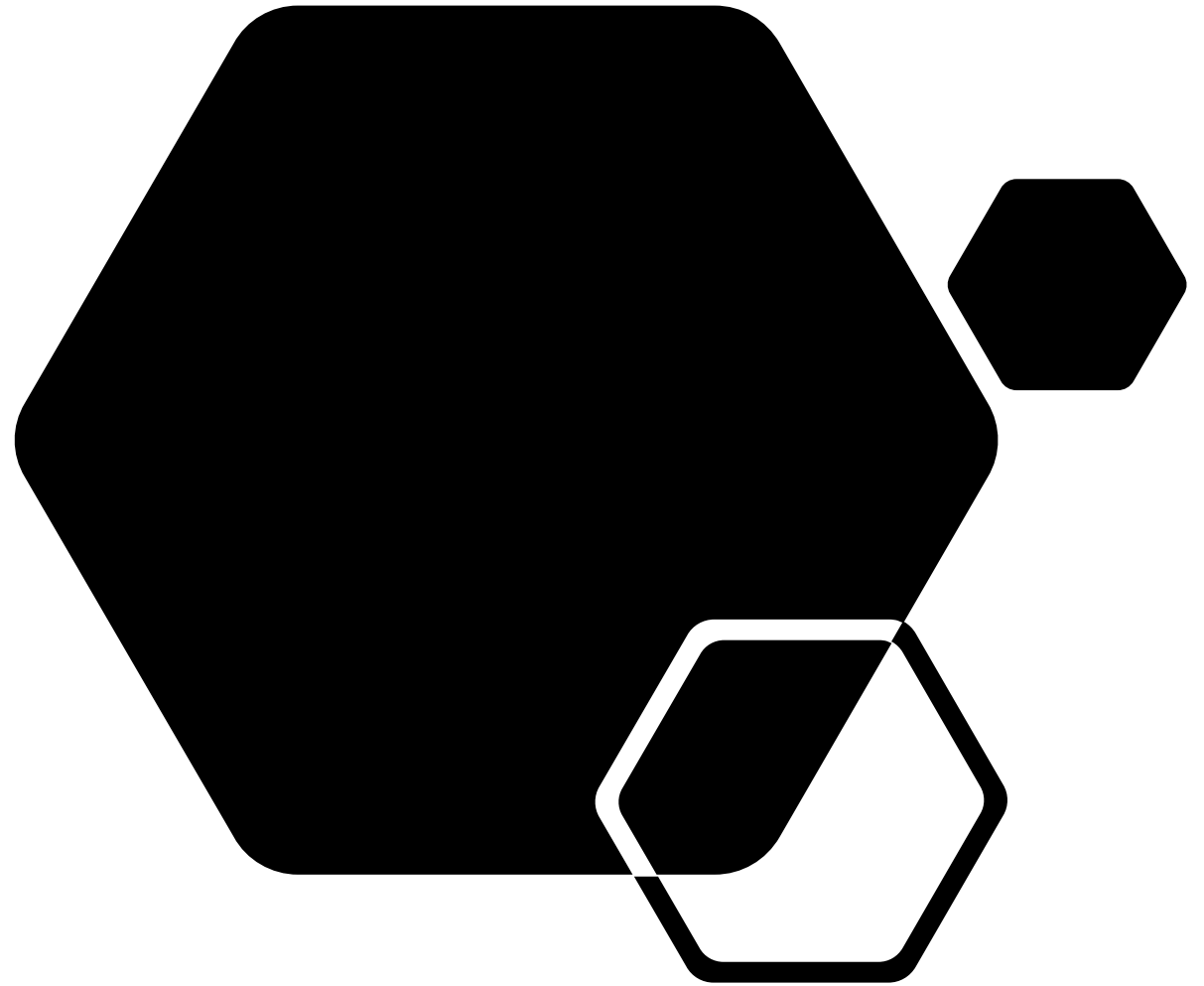
SKALA RASIO

Sama seperti skala interval tetapi tipe data ini sudah mempunyai titik nol yang tetap.

Harga barang, Berat badan, Tinggi Badan, Ipk,dsb.

Ringkasan Numerik

Ukuran Pemusatan Data



Tabel Distribusi Frekuensi

- Untuk menunjukkan frekuensi kemunculan data.
- Untuk meringkas data numerik maupun kategorik

Nilai UTS	Frekuensi	Frekuensi Kumulatif
30	4	4
40	12	16
50	5	21
60	8	29
70	9	38
80	15	53
90	20	73

Tabel Frekuensi Data Berkelompok

1. Urutkan data dari yang terkecil
2. Hitung Range / jangkauan kelas

$$R = \text{Data terbesar} - \text{Data terkecil}$$

3. Hitung jumlah kelas (K) dengan rumus Sturges :

$$K = 1 + 3,3 \log n$$

4. Hitung Panjang Kelas interval (P)

$$P = \frac{\text{Range}}{\text{Jumlah Kelas}} = \frac{R}{K}$$

5. Buat batas bawah data pertama
6. Buat table dengan dihitung satu per satu untuk interval kelas

Contoh

Diberikan data penghasilan buruh bangunan di satu kota (dalam ribuan rupiah)

58, 72, 64, 65, 67, 92, 55, 51, 69, 73,

64, 59, 65, 55, 75, 56, 89, 60, 84, 68,

74, 67, 55, 68, 74, 43, 67, 71, 72, 66,

62, 63, 83, 64, 51, 63, 49, 78, 65, 75

Buatlah table distribusi frekuensi berkelompok dari data tersebut!

Penyelesaian

□ Langkah :

1. Nilai terendah = 43 ; Nilai tertinggi = 92

2. Range = $92 - 43 = 49$

3. $K = 1 + 3,3 \log (40) = 6,287 \approx 7$

$$4. P = \frac{R}{K} = \frac{49}{6,287} = 7,794 \approx 8$$

5. Batas bawah data pertama : dipilih 43

Kelas	Frekuensi	Frekuensi Kumulatif
43 – 50	2	2
51 – 58	7	9
59 – 66	12	21
67 – 74	12	33
75 – 82	3	36
83 – 90	3	39
91 – 98	1	40

Mean/Rata-Rata

Diberikan data random sebagai berikut : $x_1, x_2, x_3, \dots, x_n$, rata – rata hitung (mean) dari data tersebut didefinisikan sebagai

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

Apabila data sudah dikelompokkan dan diketahui frekuensi masing-masing data maka rata-rata hitung tersebut dapat ditulis sebagai

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

Data (x_i)	Frekuensi (f_i)
x_1	f_1
x_2	f_2
x_3	f_3
\vdots	\vdots
x_n	f_n

Contoh

1. Pegawai di sebuah Pabrik X memberikan sumbangan (dalam \$) pada United Fund :

10, 40, 25, 5, 20, 10, 25, 50, 30, 10, 5, 15, 25, 50, 10, 30, 5, 25, 45, 15

Tentukan berapa rata-rata Pegawai di Pabrik X menyumbang pada United Fund?

2. Diberikan data hasil UTS Statistika Kelas A sbb :

70, 75, 80, 80, 70, 65, 75, 75, 80, 90, 80, 85, 85, 80, 60, 60, 70, 75, 60, 100

Tentukan mean-nya!

❖ MEAN DATA BERKELOMPOK

Mean Dengan Titik Tengah Kelas ke- i

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

dengan x_i adalah titik tengah kelas ke- i

Contoh

Tentukan rata-rata hitung dari data berikut.

Interval Kelas	Frekuensi
9-21	3
22-34	4
35-47	4
48-60	8
61-73	12
74-86	23
87-99	6
	$\Sigma f = 60$

Interval Kelas	Frekuensi (f_i)	x_i	$f_i x_i$
9-21	3	15	45
22-34	4	28	112
35-47	4	41	164
48-60	8	54	432
61-73	12	67	804
74-86	23	80	1840
87-99	6	93	558
Σ	60		3955

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{3955}{60} = 65,92$$

Sifat Rata-Rata

1. *Jumlahan dari penyimpangan setiap data terhadap rata-ratanya adalah 0.*

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Bukti:

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x}) &= \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = \sum_{i=1}^n x_i - n\bar{x} \\ &= \frac{n \sum_{i=1}^n x_i}{n} - n\bar{x} = n \left(\frac{\sum_{i=1}^n x_i}{n} \right) - n\bar{x} = n\bar{x} - n\bar{x} = 0\end{aligned}$$

Sifat...

2. Jumlah kuadrat penyimpangan setiap data terhadap rata-ratanya lebih kecil dari jumlah kuadrat simpangan data terhadap nilai yang lain

$$\sum_{i=1}^n (x_i - \bar{x})^2 < \sum_{i=1}^n (x_i - a)^2$$

Dengan $a \neq \bar{x}$, a sebarang konstanta.

Sifat...

3. Jika $d_i = x_i \pm a$ maka $\bar{d} = \bar{x} \pm a$ atau $\bar{x} = \bar{d} \mp a$

4. Jika $d_i = ax_i$ maka $\bar{d} = a\bar{x}$

5. Jika dibentuk transformasi $d_i = \frac{x_i - a}{c}$, dengan a, c sebarang

konstanta maka $\bar{d} = \frac{\bar{x} - a}{c}$

Contoh

Diketahui bahwa dari 5 pengamatan x_i , dengan $i = 1, 2, 3, 4, 5$, yaitu

1001, 1002, 1003, 1004, 1005.

Dipilih transformasi $d_i = x_i - 1000$, sehingga diperoleh d_i , berturut-turut $d_1 = 1$, $d_2 = 2$, $d_3 = 3$, $d_4 = 4$ dan $d_5 = 5$.

Akibatnya didapat $\bar{d} = 3$.

Jadi, $\bar{x} = \bar{d} + 1000 = 1003$

Keterangan

- Rata-rata mudah dihitung tetapi kurang tepat apabila digunakan untuk eksplorasi data.
- Hal ini dikarenakan untuk mendapatkannya diperlukan lebih banyak perhitungan dibandingkan dengan ukuran pemusatan yang lain.
- Rata-rata sangat terpengaruh (tidak robust) oleh adanya outlier.

Median

- Median diperoleh dengan mengurutkan data mulai dari yang terkecil hingga terbesar, kemudian dipilih data yang berada di tengah.
- **Data Ganjil**

$$\text{median} = \text{data ke } \left(\frac{n + 1}{2} \right)$$

- **Data Genap**

$$\text{median} = \frac{\text{data ke } \left(\frac{n}{2} \right) + \text{data ke } \left(\frac{n}{2} + 1 \right)}{2}$$

- Median tidak terpengaruh oleh ekstrim.

Contoh

1. Pegawai di sebuah Pabrik X memberikan sumbangan (dalam \$) pada United Fund :

10, 40, 25, 5, 20, 10, 25, 50, 30, 10, 5, 15, 25, 50, 10, 30, 5, 25, 45, 15

Tentukan median dari data sumbangan Pegawai di Pabrik X menyumbang pada United Fund?

2. Diberikan data hasil UTS Statistika Kelas A sbb :

70, 75, 80, 80, 70, 65, 75, 75, 80, 90, 80, 85, 85, 80, 60, 60, 70, 75, 60, 100

Tentukan median-nya!

Median Data Berkelompok

$$\text{Med} = L_0 + c \left(\frac{\frac{n}{2} - F}{f} \right)$$

L_0 = batas bawah kelas median

F = jumlah frekuensi semua kelas sebelum
kelas yang mengandung median

f = frekuensi kelas median

Contoh

Tentukan median dari data berkelompok berikut:

Interval Kelas	Frekuensi
9-21	3
22-34	4
35-47	4
48-60	8
61-73	12
74-86	23
87-99	6
	$\Sigma f = 60$

Letak Median : data ke- $(n/2)$ = data ke-30

$$\begin{aligned} Me &= L_o + \left(\frac{\frac{n}{2} - F}{f} \right) c \\ &= 60,5 + \left(\frac{30 - 19}{12} \right) 13 \\ &= 60,5 + 11,92 \\ &= 72,42 \end{aligned}$$

Modus

- Observasi dalam angkatan yang paling sering muncul.
- Oleh karena, suatu saat bisa jadi semua observasi merupakan modus atau tidak ada modus maka ringkasan numerik dalam hal ini tidak dapat didapatkan.
- Contoh:
 1. Angkatan terdiri atas observasi: 5,2,1,3,7,11,7,9,8,6 memiliki modus 7 (unimodal)
 2. Angkatan yang terdiri atas observasi: 5,5,1,3,7,11,7,9,3,6 memiliki modus 3,5,7 (multimodal)

Modus Data Berkelompok

$$\text{Mod} = L_0 + c \left(\frac{b_1}{b_1 + b_2} \right)$$

L_0 = batas bawah kelas modus

b_1 = selisih antara frekuensi kelas modus dengan
frekuensi tepat satu kelas sebelum kelas modus

b_2 = selisih antara frekuensi kelas modus dengan
frekuensi tepat satu kelas sesudah kelas modus

CONTOH

Tentukan Modus dari data berkelompok berikut !

Interval Kelas	Frekuensi
9-21	3
22-34	4
35-47	4
48-60	8
61-73	12
74-86	23
87-99	6
	$\Sigma f = 60$

$$b_1 = 23 - 12 = 11$$

$$b_2 = 23 - 6 = 17$$

maka

$$\begin{aligned} Mo &= Lo + \left(\frac{b_1}{b_1 + b_2} \right) c \\ &= 73,5 + \left(\frac{11}{11 + 17} \right) 13 \\ &= 73,5 + 5,107 \\ &= 78,607 \end{aligned}$$

Kuartil

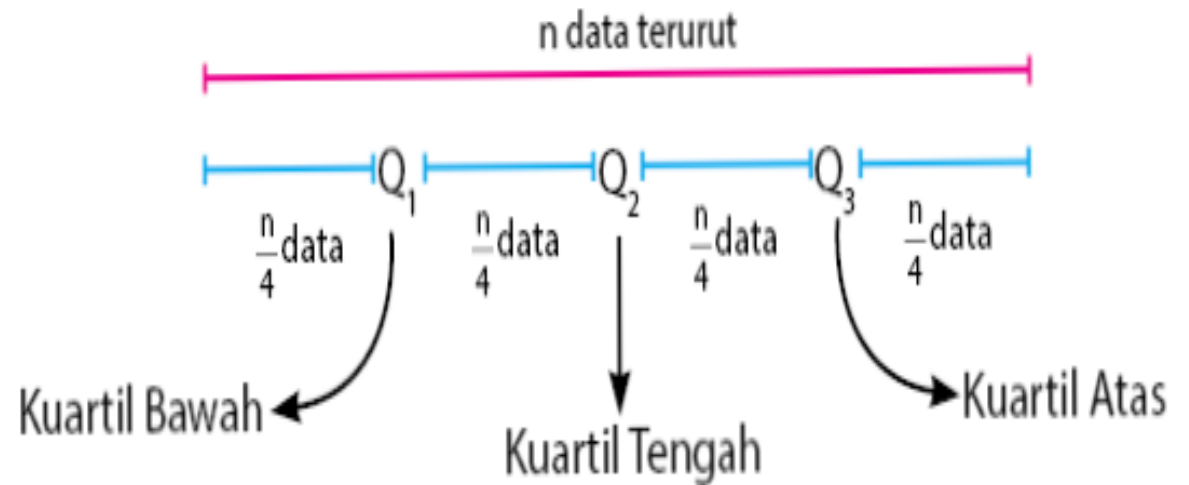
Misal diberikan data terurut sbb :

$$x_1, x_2, x_3, \dots, x_n$$

Maka kuartil ke- i (Q_i) didefinisikan sbb

- Nilai Q_1 disebut sebagai kuartil bawah
- Nilai Q_2 disebut sebagai Median
- Nilai Q_3 disebut sebagai kuartil atas

$$Q_i = \text{Data ke } \left(\frac{i(n+1)}{4} \right)$$



Contoh

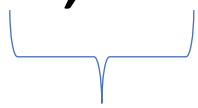
1. Diketahui data suatu penjualan dari perusahaan XYZ (dalam Jutaan Rupiah) adalah


TAHUN	2009	2010	2011	2012	2013	2014	2015	2016
HASIL	47	39	45	49	56	60	78	70

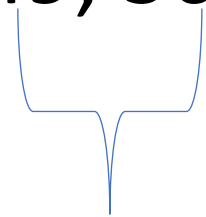
Tentukan median, Q_1 dan Q_3 !

Jawab :

Data Terurut : 39, 45, 47, 49, 56, 60, 70, 78


$$Q_1 = \frac{45 + 47}{2} = 46$$


$$Q_3 = \frac{60 + 70}{2} = 65$$


$$Q_2 = \frac{49 + 56}{2} = \frac{105}{2} = 52,5$$

2. Perhatikan data berikut :

Data	frekuensi
50	2
55	5
60	5
65	10
70	11
75	10
Jumlah	43

Tentukan Q_1, Q_2, Q_3 !

$$Q_1 = \text{data ke} - \frac{1(43 + 1)}{4} = \text{data ke} - 11 = 60$$

$$Q_2 = \text{data ke} - \frac{2(43 + 1)}{4} = \text{data ke} - 22 = 65$$

$$Q_3 = \text{data ke} - \frac{3(43 + 1)}{4} = \text{data ke} - 33 = 70$$

Quartil Data Berkelompok

$$Q_i = L_i + \left[\frac{\frac{in}{4} - \sum F}{f_i} \right] c$$

KETERANGAN :

- $\sum F$ = Jumlah Frekuensi kelas sebelum kelas kuartil, desil dan persentil
- f_i = Frekuensi kelas kuartil, desil dan persentil ke- i
- L_i = batas bawah kelas kuartil
- c = Panjang kelas
- n = banyak data

Contoh

TENTUKAN Q_1 , Q_3 DARI DATA BERKELOMPOK BERIKUT :

Nilai	Frekuensi
60-64	2
65-69	6
70-74	15
75-79	20
80-84	16
85-89	7
90-94	4

$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - \sum F}{f_1} \right) c$$

$$= 69,5 + \left(\frac{17,5 - 8}{15} \right) 5 = 72,67$$

$$Q_3 = L_3 + \left(\frac{\frac{3n}{4} - \sum F}{f_3} \right) c$$

$$= 79,5 + \left(\frac{52,5 - 43}{16} \right) 5 = 82,47$$

TRIRATA

- Rata-rata terboboti dari Quartil bawah (Q_1), Median dan Quartil Atas (Q_3).

$$TRI(x) = \frac{Q_1 + 2Med + Q_3}{4}$$

- Sebagai ukuran pusat, trirata termasuk ukuran yang tidak dipengaruhi oleh nilai ekstrim.

Contoh

Diberikan data dalam table berikut.

No	x_i	f_i
1	2	10
2	4	20
3	6	25
4	7	15
	Total	70

- Median = $\frac{\text{data ke } 35 + \text{data ke } 36}{2} = 6$

- $Q_1 = \text{data ke } 18 = 4$

- $Q_3 = \text{data ke } 53 = 6$

Sehingga didapat

$$TRI(x) = \frac{4 + 6 + 6}{4} = 4$$