KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 6:

Analisis Data dengan Metode Statistika Deskriptif

Sub CPMK

- Mahasiswa menemukan struktur dan karakteristik data menggunakan metode statistika deskriptif.

Pokok Bahasan

- Mendefinisikan dan menghitung ukuran pemusatan data serta ukuran penyebaran data
- Membuat plot untuk identifikasi nilai ukuran pemusatan dan persebaran data

Pengantar

- Statistika deskriptif dapat digunakan untuk mengidentifikasi properti / karakteristik data.
- Dalam statistika deskriptif kita mengenal ukuran pemusatan dan penyebaran data.
- Ukuran pemusatan data membahas mengenai ukuran mean, median, modus, dan midrange.
- Ukuran dpersebaran data yang akan dibahas diantaranya range, kuartil dan interquartile range, varians dan standar deviasi.
- Selain ukuran tersebut, statistika deskriptif juga dapat dimuat secara visual.
- Secara visual kita akan belajar mengenai diagram batang, diagram lingkaran, dan grafik garis, histogram, diagram pencar, dan lainnya.

Ukuran Pemusatan Data: Mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Contoh

Besar gaji (dalam juta rupiah) karyawan selama setahun adalah sebagai berikut:

5, 8, 9, 10, 8, 9, 12, 12, 10, 7, 13, 12

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{5 + 8 + 9 + 10 + 8 + 9 + 12 + 12 + 10 + 7 + 13 + 12}{12}$$

$$\bar{x} = \frac{115}{12} = 9{,}583$$

Sehingga kita dapatkan informasi bahwa rata-rata gaji pegawai tersebut selama setahun sebesar Rp 9,583 juta.

Ukuran Pemusatan Data: Median

Kita dapat memperkirakan median dari seluruh kumpulan data dengan interpolasi menggunakan rumus:

$$median = L_1 + \left(\frac{\frac{N}{2} - (\sum freq)l}{freq_{median}} \right) width$$

Contoh pada kasus gaji karyawan. Data kita urutkan terlebih dahulu dalam urutan yang meningkat, seperti ini:
5,7,8,8,9,9,10,10,12,12,12,13.

Sehingga nilai mediannya berada diantara data urutan ke-6 dan ke-7:

$$\frac{9 + 10}{2} = 9{,}5$$

Jadi, median dari gaji karyawan adalah Rp 9,5 juta

Ukuran Pemusatan Data: Modus

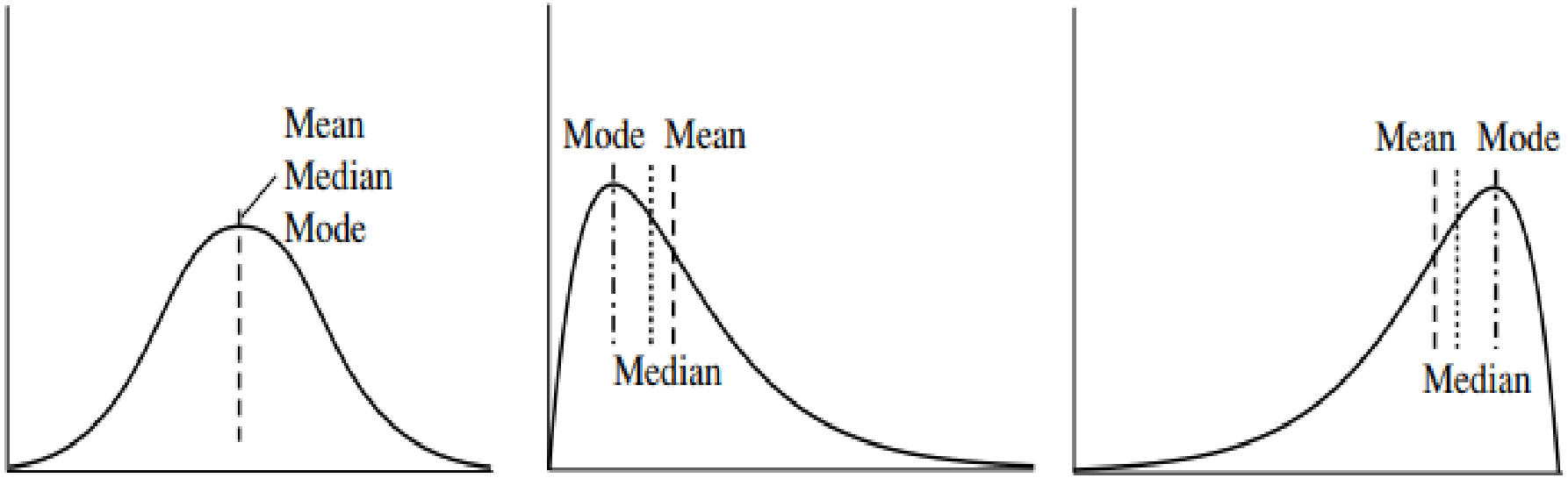
- Modus untuk sekumpulan data adalah nilai yang paling sering muncul dalam sekumpulan data tersebut.
- Frekuensi terbesar dimungkinkan untuk sama besarnya dengan beberapa nilai yang berbeda, yang menghasilkan lebih dari satu modus.
- Kumpulan data dengan satu, dua, atau tiga modus masing-masing disebut unimodal, bimodal, dan trimodal. Secara umum, kumpulan data dengan dua atau lebih modus adalah multimodal.
- Di sisi lain, jika setiap nilai data hanya muncul sekali, maka tidak ada modus.
- Contoh untuk kasus besarnya gaji karyawan sebelumnya adalah unimodal dengan modus sebesar Rp 12 juta.

Ukuran Pemusatan Data: Midrange

- Midrange adalah rata-rata dari nilai terbesar dan terkecil dalam himpunan.
- Ukuran ini mudah dihitung menggunakan fungsi max() dan min().
- Pada contoh gaji karyawan, nilai midrange nya adalah

$$\frac{5 + 13}{2} = Rp\ 9\ juta$$

Ukuran Pemusatan Data



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

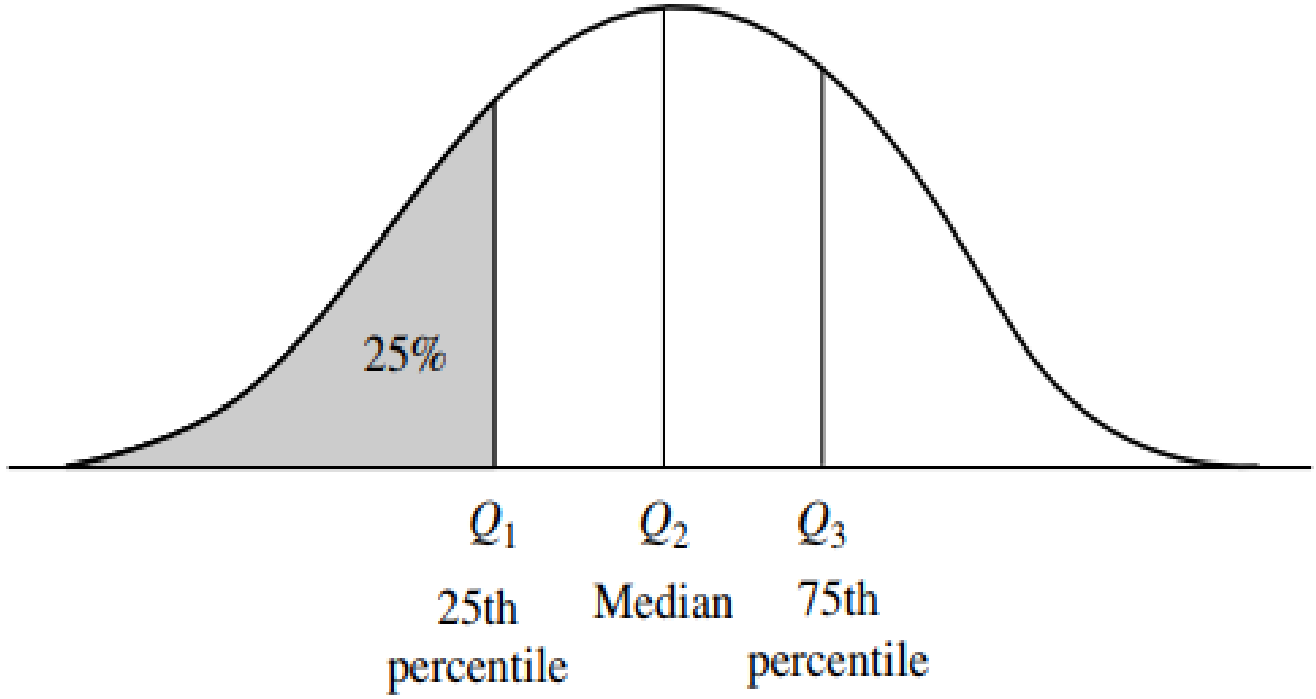
Mean, median, modus untuk data simetris vs skew positif vs skew negatif

Ukuran Penyebaran Data: Range

- Misalkan $x_1, x_2, \ldots, x_N$ adalah himpunan pengamatan untuk beberapa atribut numerik, X.
- Jangkauan (*range*) dari himpunan adalah selisih antara nilai terbesar maks() dan terkecil min().

Ukuran Penyebaran Data: Kuantil

- Misalkan data untuk atribut X diurutkan dalam urutan numerik yang meningkat.
- Kemudian dipilih titik data tertentu untuk membagi distribusi data menjadi ukuran yang sama dalam bagian yang berurutan, seperti pada Gambar di samping.
- Titik data ini disebut kuantil (*quantiles*).



Ukuran Penyebaran Data: Kuantil

- 2-Kuantil adalah titik data yang membagi bagian bawah dan atas dari distribusi data. Hal ini sesuai dengan median.
- 4-Kuantil adalah tiga titik data yang membagi distribusi data menjadi empat bagian yang sama, setiap bagian mewakili seperempat dari distribusi data. Hal ini lebih sering disebut sebagai kuartil.
- 100-Kuantil lebih sering disebut sebagai persentil, dimana membagi distribusi data menjadi 100 bagian berurutan berukuran sama.
- Median, kuartil, dan persentil adalah bentuk kuantil yang paling banyak digunakan.

Ukuran Penyebaran Data: Kuartil

- Kuartil memberikan indikasi pusat distribusi, penyebaran, dan bentuk distribusi.
- Kuartil pertama, dilambangkan dengan Q_1, adalah persentil ke-25. Q_1 ini memotong 25% terendah dari data.
- Kuartil ketiga, dilambangkan dengan Q_3, adalah persentil ke-75—memotong 75% terendah (atau 25% tertinggi) dari data.
- Kuartil kedua adalah persentil ke-50. Sebagai median, Q_2 memberikan pusat distribusi data.
- Jarak antara kuartil pertama dan ketiga (jangkauan interkuartil (IQR)) adalah ukuran penyebaran sederhana yang memberikan rentang yang dicakup oleh bagian tengah data.

$$IQR = Q_3 - Q_1$$

- Kuartil untuk contoh gaji karyawan (5,7,8,8,9,9,10,10,12,12,12,13) yaitu, Q_1 Rp 8 juta dan Q_3 adalah Rp 12 juta.
- Jangkauan interkuartilnya adalah

$$IQR = 12 - 8 = Rp\ 4\ juta$$

Ukuran Penyebaran Data: Varians & Deviasi standar

- Varians dari N observasi, $x_1, x_2, \ldots, x_N$, untuk atribut numerik X adalah

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}{x_i}^2\right) - \bar{x}^2$$

- Simpangan baku atau deviasi standar, σ, dari pengamatan adalah akar kuadrat dari varians, σ^2.
- Pada contoh besar gaji karyawan, $\bar{x} = 9{,}583$, $N = 12$.
- Maka varians dan simpangan bakunya adalah

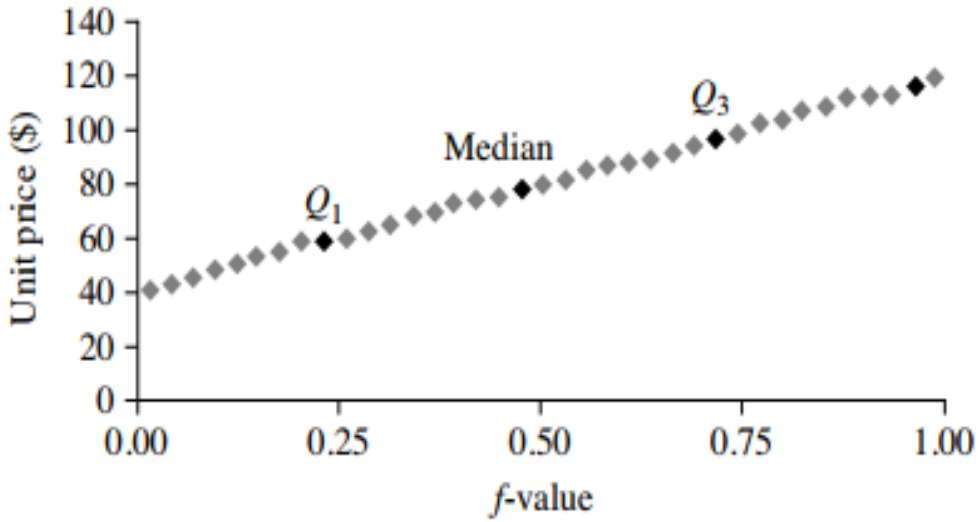
$$\sigma^2 = \frac{1}{12}(5^2 + 7^2 + \cdots + 13^2) - 9{,}583^2 = 5{,}7197$$

$$\sigma = \sqrt{5{,}7197} = 2{,}392$$

Plot untuk statistika deskriptif: Plot kuantil

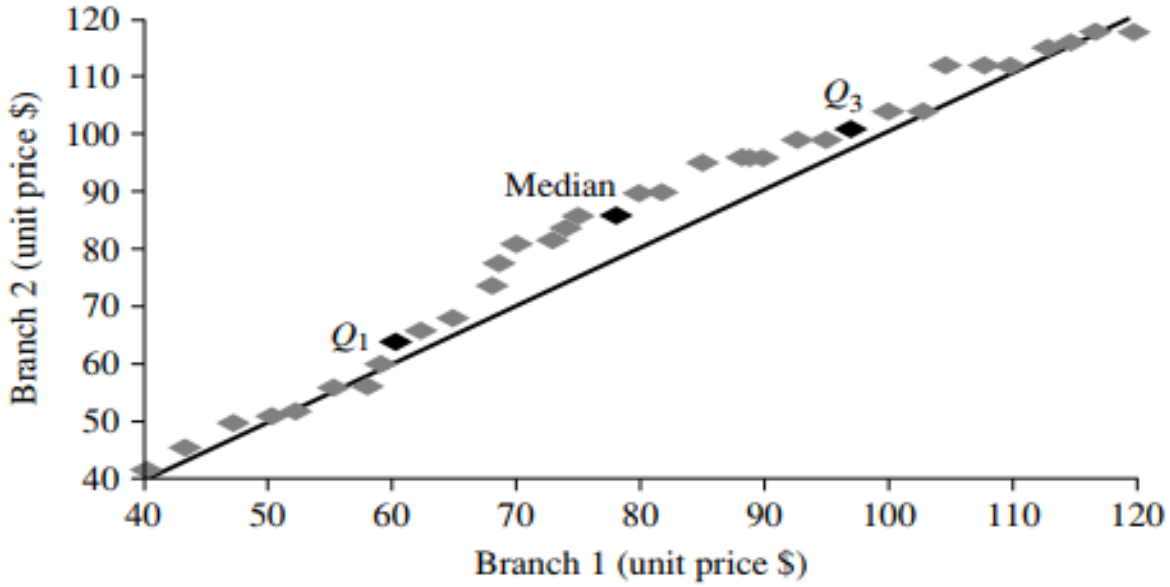
Plot kuantil adalah cara sederhana dan efektif untuk melihat distribusi data univariat untuk pertama kali.

Unit price ($)	Count of items sold
40	275
43	300
47	250
-	-
74	360
75	515
78	540
-	-
115	320
117	270
120	350



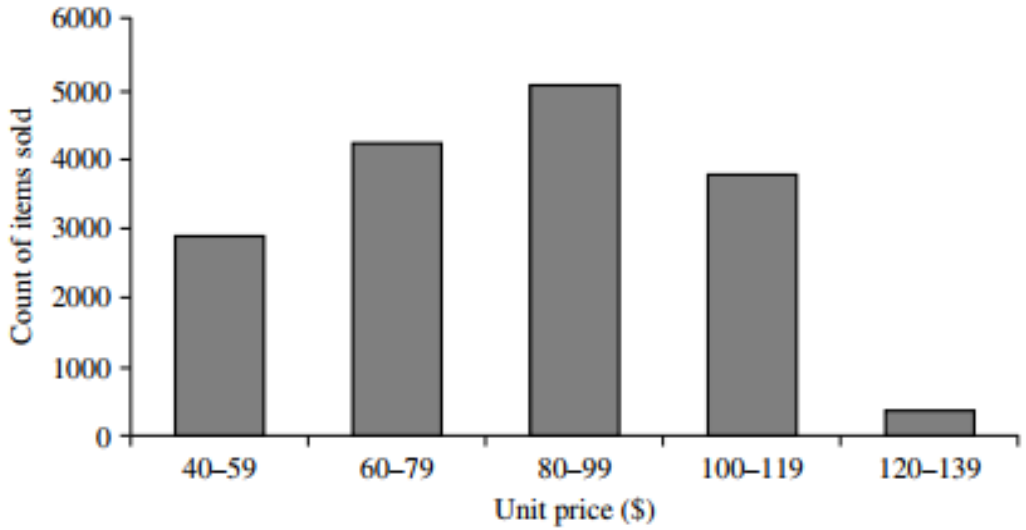
Plot untuk statistika deskriptif: Plot kuantil-kuantil

- Plot kuantil-kuantil atau q-q plot, adalah grafik kuantil dari satu distribusi univariat terhadap kuantil yang sesuai dari yang lain.
- Plot ini adalah alat visualisasi yang kuat karena memungkinkan pengguna untuk melihat apakah ada pergeseran dari satu distribusi ke distribusi lainnya



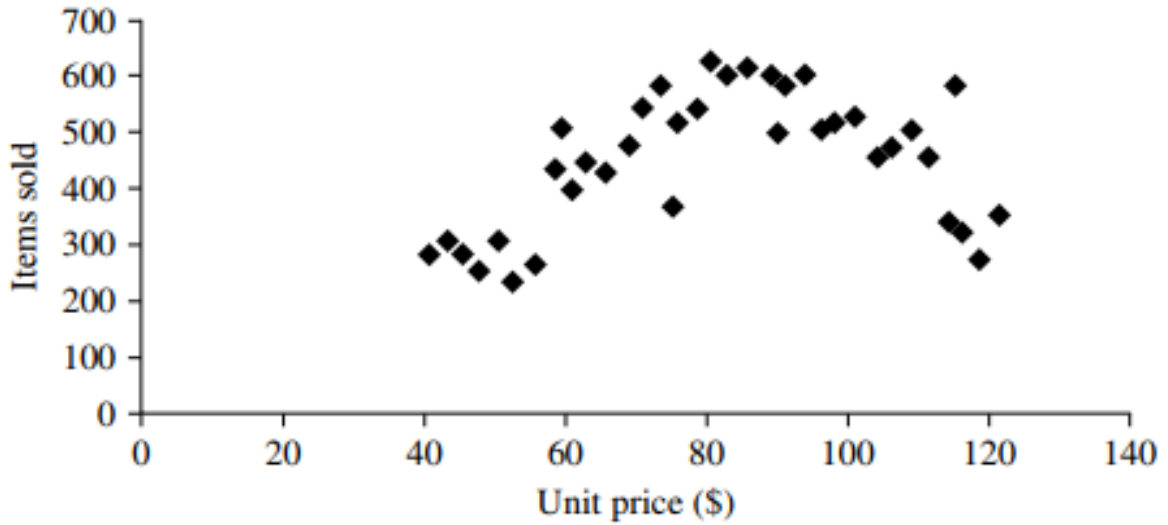
Plot untuk statistika deskriptif: Histogram

- Hstogram berasal dari kata "Hstos" berarti tiang dan "gram" berarti bagan (*chart*).
- Jadi histogram adalah bagan dari tiang.
- Jika nilai dari data pengamatan berskala nominal, maka grafik yang dihasilkan disebut diagaram batang (bar chart).



Plot untuk statistika deskriptif: Scatter plot

- Scatter plot adalah salah satu metode grafis yang paling efektif untuk menentukan apakah ada hubungan, pola, atau tren antara dua atribut numerik. Untuk membuat scatter plot, setiap pasangan nilai diperlakukan sebagai pasangan koordinat dan diplot sebagai titik.



Praktikum

Menggunakan Bahasa pemrograman R, informasi awal terkait data, khususnya statistika deskriptif dapat dihitung sebagai berikut.

- ✓ Ukuran data
- ✓ Ringkasan data
- ✓ Pengurutan data berdasarkan variabel
- ✓ Memahami distribusi data
  a) Box plot
  b) Hstogram
  c) Skewness dan Kurtosis

Tugas

1. Merangkum metode statistika yang digunakan untuk analisis data.
2. Melakukan analisis deskriptif pada dataset baru yang diunduh dari Kaggle atau UCI Machine Learning Repository di RStudio.
3. Lengkapi dengan plot yang sesuai dengan variabel pada dataset Anda. Banyak nya plot, minimal 6 jenis plot, diantaranya mencakup: boxplot, histogram, density plot, scatter plot, quantile plot, dan quantile-quantile (q-q) plot.
4. Berikan interpretasi untuk setiap hasil yang Anda dapatkan.
   - Tugas dikerjakan berkelompok.
   - Tugas dikumpulkan paling lambat pukul 23.59 WIB di LMS.
   - Beri nama file tugas: Tugas 06_Kelompok XX. (Contoh: Tugas 06_Kelompok 01)

FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN

PERTUKARAN MAHASISWA merdeka

Kampus Merdeka INDONESIA JAYA

Terima Kasih

SMART UNIVERSITY

TOP #465 WORLD UNIVERSITY RANKINGS UNAIR UNGGUL

ZONA INTEGRITAS menuju wilayah bebas korupsi