







Modul 2

Introduction to everything

Bagian

Data analytics with python- Basic statistics



Kompus Merdeko Indonesia Jaya







Learning Objectives

Di akhir modul ini, Anda akan dapat:

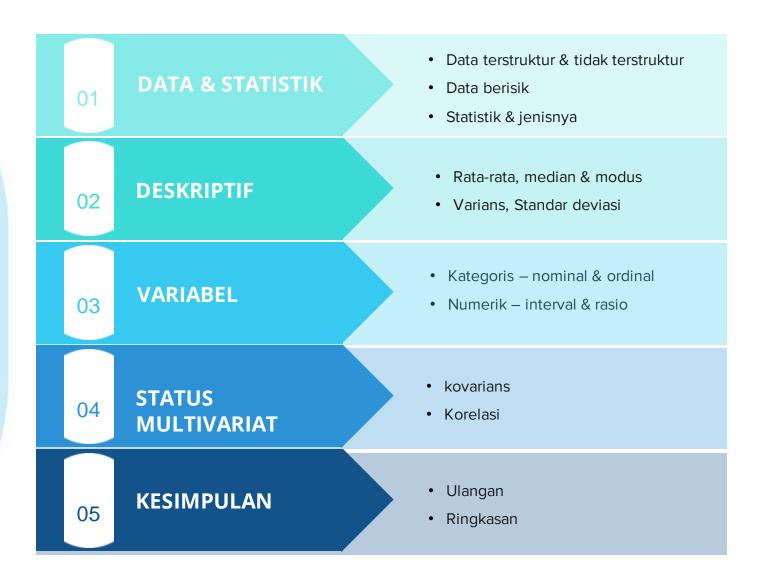
- Sebutkan macam-macam data.
- Menjelaskan konsep dasar statistika beserta klasifikasinya: statistik deskriptif dan inferensial.
- Memahami ukuran skala: skala nominal, ordinal, interval, dan rasio.
- Jelaskan ukuran tendensi sentral: mean, median & modus.
- Hitung ukuran variabilitas: jangkauan, kuartil dan jangkauan antarkuartil, varians dan standar deviasi.
- Jelaskan ukuran-ukuran bentuk: kemiringan dan kurtosis.
- Memahami pengenalan dan definisi penting dari Teori Probabilitas.
- Jelaskan kemungkinan bersyarat
- Memahami konsep Kovarians dan Korelasi.



















DATA & STATISTIK

- Data terstruktur & tidak terstruktur
- Noisy Data
- Statistik & jenisnya

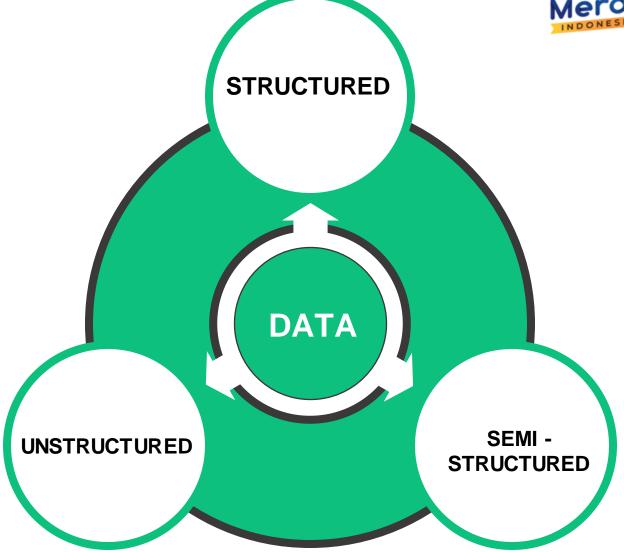




Jenis Data









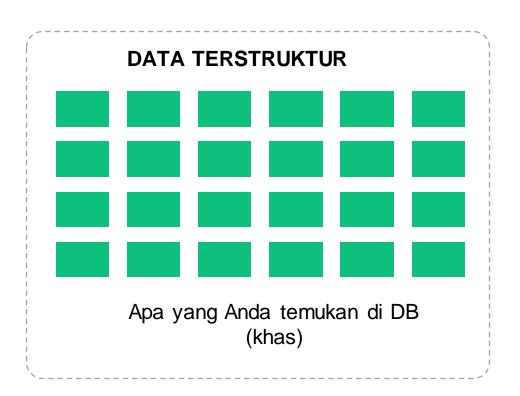




Data Terstruktur

Data Terstruktur:

- Sangat terorganisir dan mudah dicari dengan kueri atau algoritme.
- Dapat dengan cepat dikonsolidasikan menjadi fakta.
- Mengikuti skema yang telah ditentukan sebelumnya.
- Biasanya berada di bidang tetap.
- Contoh tipikal adalah Relational Database Management System (RDBMS).
- Skema ditentukan sebelum konten dibuat dan data diisi.



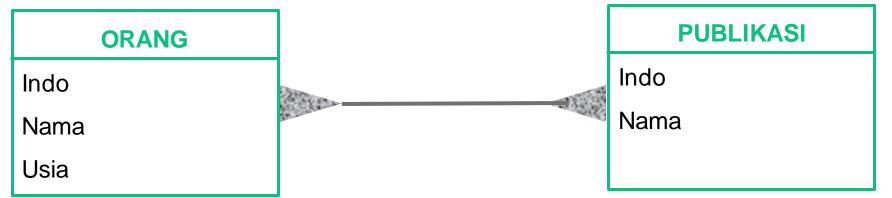
Organisasi Data Terstruktur





Diagram dan tabel berikut menggambarkan data terstruktur.





Model Skema (diagram ER, DDL)

Tabel: Orang		
Indo	Nama	Usia
1	Rolf	28
2	Steffi	22

1	1
1	2
2	2

Tabel: Publikasi		
Indo	Nama	
1	Penandaan Sosial	
2	Klasifikasi	

Tabel data

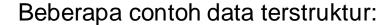
Contoh tabel dalam sistem database relasional

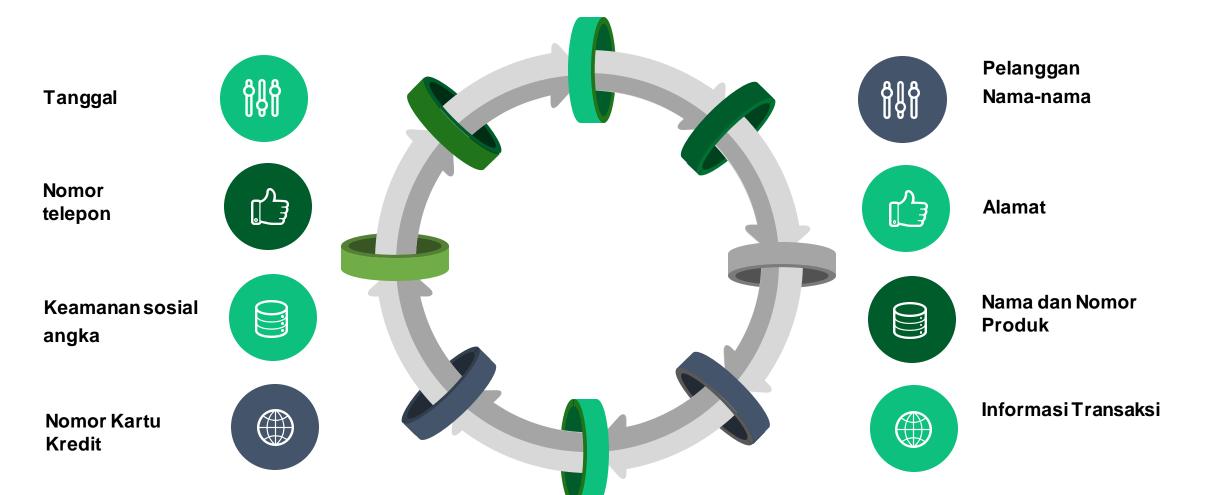
Contoh Data Terstruktur

Kampus Merdeka



Artificial Intelligence
Mastery Program





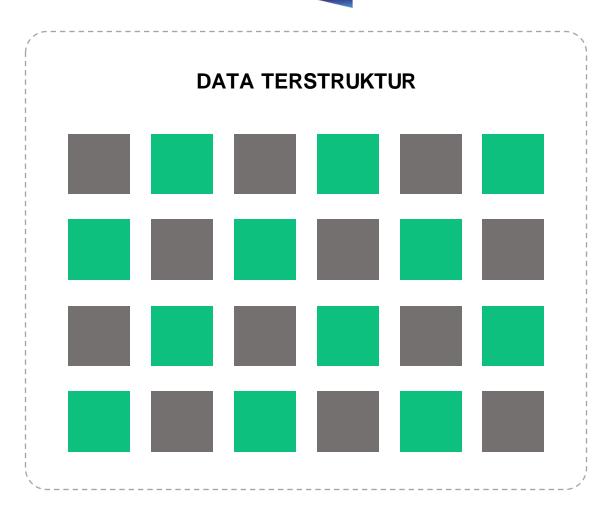
Aplikasi Data Terstruktur





Artificial Intelligence
Mastery Program

- Aplikasi database relasional dengan Data Terstruktur:
- Sistem reservasi maskapai penerbangan
- Kontrol inventaris
- Transaksi penjualan
- aktivitas ATM



Apa itu Data Tidak Terstruktur?

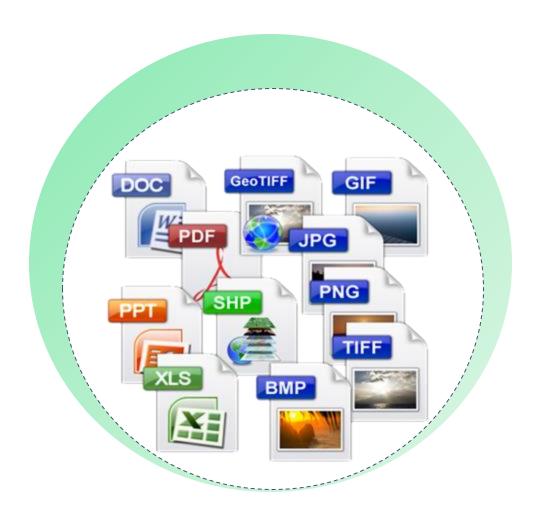




Artificial Intelligence
Mastery Program

Data Tidak Terstruktur:

- Tidak ada struktur yang dapat diidentifikasi untuk jenis data ini.
- Memiliki struktur internal, tetapi tidak memiliki skema yang telah ditentukan sebelumnya.
- Tidak dapat disimpan dalam baris dan kolom seperti database relasional.
- Tidak ada model data tetap, tidak terorganisir
- Memerlukan lebih banyak ruang penyimpanan daripada data terstruktur.



Contoh Data Tidak Terstruktur



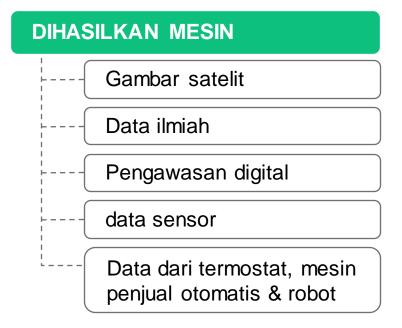


Artificial Intelligence
Mastery Program

Data Tidak Terstruktur yang dibuat oleh manusia mencakup hal-hal berikut:



Data Tidak Terstruktur yang dihasilkan mesin mencakup hal-hal berikut:



Apa itu Data Semi Terstruktur?

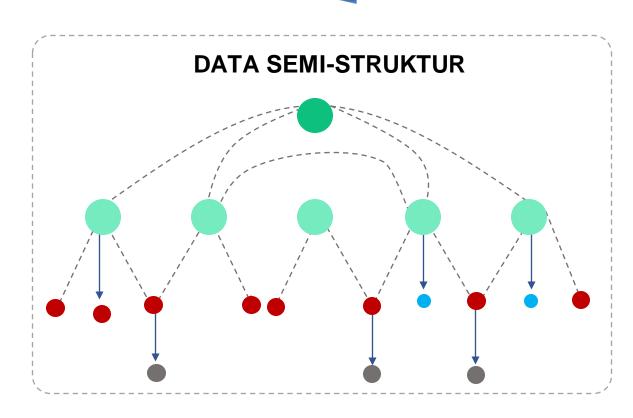




Artificial Intelligence
Mastery Program

Data semi terstruktur:

- Sering dijelaskan sebagai tanpa skema atau menggambarkan diri sendiri.
- Berisi tag semantik, tetapi tidak sesuai dengan standar atau struktur database relasional biasa.
- Tidak ada deskripsi terpisah tentang jenis atau struktur data.
- Tidak memerlukan definisi skema, definisi tersebut bukan tidak mungkin, tetapi bersifat opsional.
- Data dapat memiliki atribut yang berbeda, dan atribut baru dapat ditambahkan kapan saja.









Contoh Data Semi Terstruktur

Contoh Data Semi-terstruktur meliputi:

- Bahasa markup XML seperangkat aturan pengkodean dokumen yang menentukan format yang dapat dibaca manusia dan mesin.
- Open Standard JSON format pertukaran data ringan, teks biasa, berdasarkan subset dari bahasa pemrograman JavaScript.
- NoSQL beberapa database noSQL berisi data semi-terstruktur.







Aplikasi Data Semi-terstruktur

Aplikasi yang berisi data semi terstruktur antara lain:

- Infrastruktur Data Besar.
- Aplikasi web
- LinkedIn
- Tenaga penjualan
- Rekomendasi pembaca di Amazon

Noisy Data



Noisy Data:

- Berisi sejumlah informasi yang tidak terlalu dibutuhkan, menyimpang, atau tidak diketahui kebenarannya
- Sebuah anomali dan dianggap sebagai kesalahan. Menurut Zhu et al. (2004), noise dalam data berdampak negatif pada hasil data mining.
- Beberapa sumber Noisy Data:
- a) Random Noise, misalnya suara kebisingan yang terekam saat meeting online
- b) Human Error, misalnya kesalahan input data
- c) Outlier, suatu data yang sangat berbeda dari data umumnya







Pengantar Statistika

Statistika:

- Ilmu yang berkaitan dengan pengumpulan, penataan, penyajian, analisis, dan interpretasi data menjadi informasi untuk membantu pengambilan keputusan yang efektif.
- Disiplin ilmu yang menggunakan sampel data untuk mendukung klaim tentang populasi.



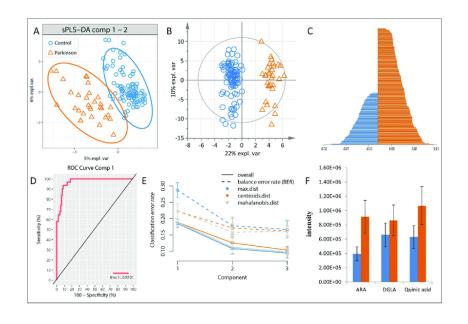




Pengantar Statistika

Statistik:

- Pengukuran dari data, seperti mean, median, modus, dll.
- Penyajian data statistik dapat berbentuk tabel, grafik, diagram, deretan angka dan visualisasi angka



Beberapa istilah penting







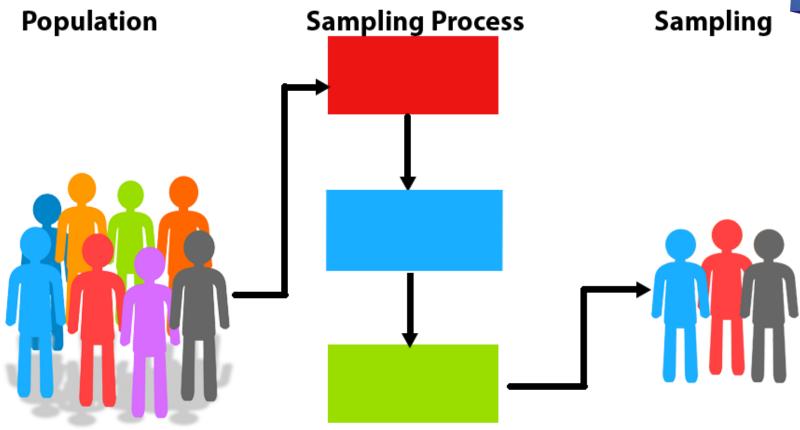
- Populasi: Himpunan objek yang akan dianalisa.
 Misalnya: Semua Mobil, Semua karyawan Microsoft
- Sensus: Ketika peneliti mengumpulkan data dari seluruh populasi.
 Misalnya Sensus Penduduk AS diambil setiap 10 tahun
- Sampel: Bagian dari populasi yang dianggap sudah cukup mewakili populasi.
 - Misalnya: Dalam eksperimen mobil diambil sampel 1 mobil dari tiap brand
- Survey : Ketika peneliti mengumpulkan data dari Sampel saja

Beberapa istilah penting





Artificial Intelligence
Mastery Program



Sensus

Survey



Pembagian Statistik

Deskriptif

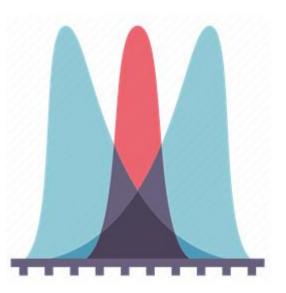


Mendeskripsikan dan menganalisis sampel data tanpa melakukan proses penarikan kesimpulan





Inferensial



Menganalisa sampel untuk mengestimasi dan membuat kesimpulan mengenai populasi







Statistik deskriptif

- Bertujuan untuk mendapatkan gambaran karakteristik dari sampel.
- Meringkas dan melihat representasi dari data menggunakan grafik, bagan, dan tabel.
- Beberapa metodenya:
- **□**Ukuran Pemusatan
- □ Ukuran penyebaran
- ■Visualisasi Data
- □Eksplorasi hubungan data







Statistik Inferensial

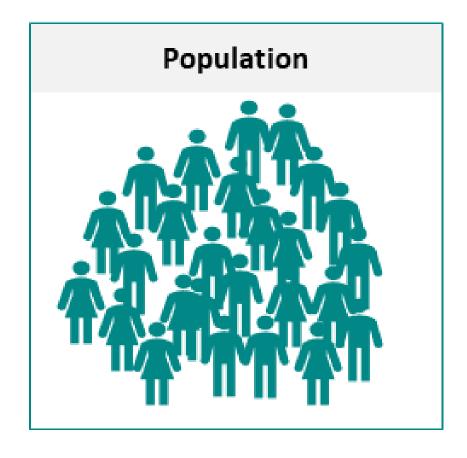
- Bertujuan untuk menguji suatu hipotesis dan menarik kesimpulan tentang suatu populasi berdasarkan sampel.
- Hasil analisis digeneralisasikan dari sampel ke populasi yang lebih besar.
- Yang berhubungan dengan metode ini:
- **□**Uji Hipotesis
- ☐P-value
- ☐ Selang kepercayaan

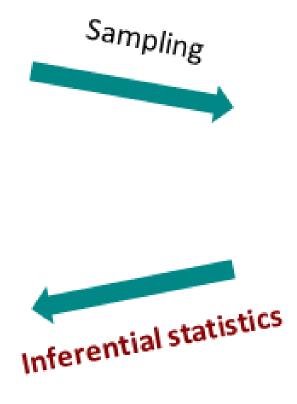
Inferential Statistics Vs **Descriptive Statistics**

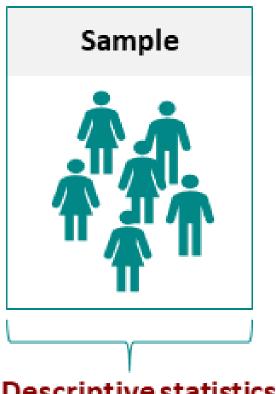




Artificial Intelligence
Mastery Program







Descriptive statistics









02

STATISTIK DESKRIPTIF

- Rata-rata, median & modus
- Varians, Standar deviasi
- Kemiringan & Kurtosis









Ukuran Pemusatan

- Ukuran pemusatan adalah metode yang umum digunakan untuk mengidentifikasi pusat kumpulan data.
- Digunakan untuk menggambarkan kumpulan data dengan mengidentifikasi posisi sentral dalam kumpulan data tersebut.

Ukuran pemusatan:

- Mean
- median
- Mode

Mean







Mean:

- Juga disebut sebagai nilai rata-rata dalam kumpulan data.
- Didefinisikan sebagai jumlah semua nilai dalam kumpulan data, dibagi dengan jumlah total nilai dalam kumpulan data.
- Mudah terpengaruh outlier
- Contoh: Misalkan x1, x2, x3, Xn menjadi nilai dalam kumpulan data, dan ada 'n' jumlah nilai dalam kumpulan data:

Mean =
$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \text{ or } \frac{\sum x}{n}$$

Median







median:

- Nilai tengah dalam suatu kumpulan data, setelah disusun menurut urutan besarnya.
- Nilai diurutkan terlebih dahulu dari yang terkecil sampai yang terbesar.
- Median adalah nilai tengah dari kumpulan data jika jumlah nilai adalah bilangan ganjil; rata-rata dari dua nilai tengah, jika jumlah nilai adalah bilangan genap.
- Median = Nilai ke-[(n+1)/2, jika ada bilangan ganjil.
- Median = rata-rata dari nilai ke-[n/2] dan [(n/2)+1], jika ada bilangan genap.

Mode

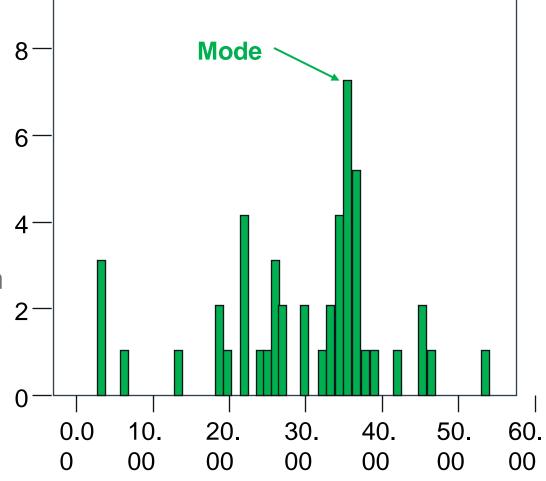




Artificial Intelligence
Mastery Program

Mode:

- Nilai yang paling sering muncul dalam kumpulan data, yaitu nilai sampel acak yang terjadi dengan frekuensi terbesar.
- Digunakan secara kualitatif.
- Berlaku untuk semua tingkat pengukuran data, yaitu skala nominal, ordinal, interval dan rasio.
- Bimodal jika nilai paling sering muncul, dua mode.
- Multimodal Kumpulan data dengan lebih dari dua mode.



Latihan

Kompus
Merdeko
INDONESIA JAYA

Artificial Intelligence
Mastery Program

Misalkan kita punya 2 data berikut:

Data A: 1, 1, 2, 2, 3, 3, 4, 4, 5, 50

Data B: 1, 1, 2, 2, 3, 3, 4, 4, 5, 5

Coba tentukan mean dan median kedua data!

Latihan

Misalkan kita punya 2 data berikut:

Data A: 1, 1, 2, 2, 3, 3, 4, 4, 5, 50

Data B: 1, 1, 2, 2, 3, 3, 4, 4, 5, 5



Jawaban:

Data A dan B sama-sama memiliki median 3, tetapi mean data A 7.5, sementara mean data B 3.

Terlihat mean dari A kurang mampu merepresentasikan data A yang didominasi angka 1-5, ini karena A memiliki *outlier* dan mean mudah terpengaruh oleh *outlier*. Sementara median tidak terpengaruh sehingga mampu memberi kita gambaran (representasi) dari data keseluruhannya.

Biasanya, jika mean >> median, maka outliemya nilai ekstrim besar (seperti contoh diatas), sementara jika mean << median, maka outliemya nilai ekstrim kecil







Ukuran Penyebaran

- Digunakan untuk mengetahui bagaimana data tersebar di sekitar mean
- Digunakan dalam data kuantitatif, karena variabel dapat disusun dalam urutan yang logis, dengan nilai rendah dan tinggi.

Langkah-langkah penyebaran:

- Range
- Kuartil dan Inter-quartile range
- Varians
- Standar deviasi







Range

- Ukuran penyebaran yang paling sederhana.
- Jarak antara nilai terkecil dan nilai terbesar dalam sebuah dataset.
- Dipengaruhi oleh outlier, karena varians mungkin terlalu rendah atau terlalu tinggi karena outlier.





Kuartil dan Inter-quartile range

- Bagilah kumpulan data yang diurutkan menjadi empat bagian yang sama, dan lihat nilai titik di antara kuarter:
- Kuartil bawah (Q1) adalah titik antara 25% nilai terendah dan 75% nilai tertinggi. Ini juga disebut persentil ke-25.
- Kuartil kedua (Q2) adalah bagian tengah dari kumpulan data. Ini juga disebut persentil ke-50, atau median.
- Kuartil atas (Q3) adalah titik antara nilai 75% terendah dan 25% tertinggi. Ini juga disebut persentil ke-75.
- Rentang interkuartil (IQR) adalah perbedaan antara kuartil atas (Q3) dan bawah (Q1) dan menggambarkan 50% nilai tengah ketika diurutkan dari terendah ke tertinggi.

Varians



Perbedaan:

- Varians adalah ukuran seberapa menyebar set data.
- Dihitung sebagai deviasi kuadrat rata-rata setiap angka dari rata-rata kumpulan data.
- Rumus untuk menghitung varians (S2) dari kumpulan data adalah:

$$s^2 = \frac{\sum (x - \overline{x})^2}{n}$$

• (di mana x adalah nilai individual, $\overline{\mathbf{x}}$ adalah rata-rata dari kumpulan sampel, n adalah jumlah nilai dalam distribusi.)

Standar Deviasi (SD)







SD:

- Jarak rata-rata semua nilai dari rata-rata keseluruhan.
- Akar kuadrat dari varians.
- SD(s) dari kumpulan data sampel dapat dihitung menggunakan rumus:

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

• (di mana x adalah nilai individua xadalah rata-rata dari kumpulan sampel, n adalah jumlah nilai dalam distribusi.)

Latihan



Misalkan ada data nilai di 2 kelas:

C: 60, 70, 70, 70, 80, 80, 80, 80, 90, 100

D: 75, 75, 75, 75, 80, 80, 80, 80, 80, 80

Coba hitung mean, range, dan varians, dan standar deviasinya!

Latihan

C: 60, 70, 70, 70, 80, 80, 80, 80, 90, 100

D: 75, 75, 75, 75, 80, 80, 80, 80, 80, 80



*Penjabaran:

$$78 - 11.4 = 66.6$$
, $78 + 11.4 = 89.4$
 $78 - 2.6 = 75.4$, $78 + 2.6 = 80.6$

Mean dari kedua kelas sama-sama 78. Namun Range C 40, dan Range D 5. Varians C 128.9 (std=11.4) dan Varians D 6.7 (std=2.6).

Meski rata-ratanya sama, tapi dari range dan varians kita mengetahui bahwa nilai di kelas C cenderung beragam sementara di D hampir seragam. Biasanya data paling banyak **tersebar di mean ± standar deviasi**. Sehingga, tanpa melihat datanya 1 per 1 kamu bisa memperkirakan nilai kelas C paling banyak tersebar di 66 hingga 89 sementara nilai kelas D tersebar di 75 hingga 80.









03 Skala Pengukuran Data

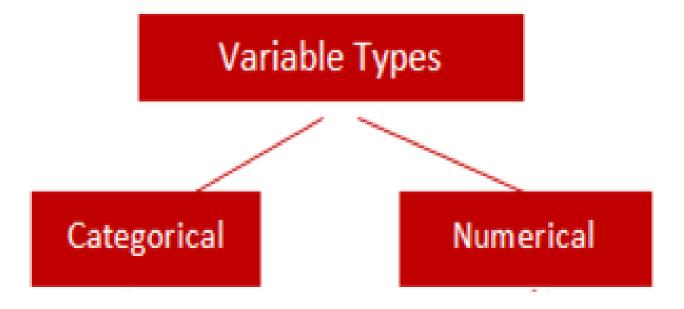
- Kategoris nominal & ordinal
- Numerik interval & rasio





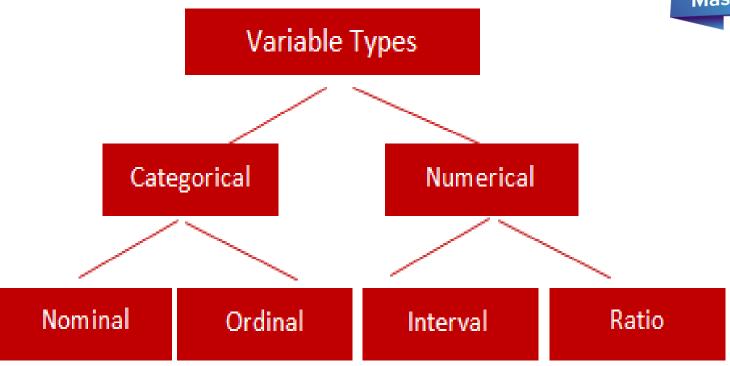


Jenis variabel / skala pengukuran



Jenis variabel / skala pengukuran





Mengapa penting dipelajari? Karena perbedaan skala pengukuran juga menentukan pemilihan statistik yang tepat untuk analisa datanya.

Skala Nominal



Kategorik tanpa urutan

- Data berupa kategorik, namun tidak berurutan
- Misalnya Data jenis kelamin, misalnya kita mengkodekan 1 = laki-laki, 2 perempuan, karena 2 > 1 bukan berarti perempuan itu diatas laki-laki
- Contoh lainnya adalah data provinsi, data suku, dll
- Pada data ini, statistik yang cocok adalah frekuensi dan modus

Skala Ordinal







Kategorik berurutan

- Data bersifat kategorik, namun memiliki makna urutan
- Misalkan kita mengkodekan data tingkat dengan : 4 = PT, 3 = SMA sederajat, 2 = SMP sederajat, 1 = SD sederajat
- 4 > 3 > 2 > 1 menunjukkan lulusan PT tingkat pendidikannya memang diatas lulusan SMA, lulusan SMA diatas lulusan SMP, begitu seterusnya
- Contoh lainnya adalah data jabatan di perusahaan, data skala kepuasan (Sangat Puas, Puas, Tidak Puas, Sangat Tidak Puas), dll
- · Pada data ini, statistik yang cocok adalah frekuensi, modus, dan median

Skala Interval







Numerik, 0 tidak mutlak

- Datanya berupa angka
- 0 pada data interval tidak menunjukkan 0 yang sesungguhnya
- Contohnya data suhu, 0°C bukan berarti tidak ada suhunya, 0 disini bukan 0 sebenarnya
- Statistik yang cocok untuk analisa datanya adalah frekuensi, mean, median, modus, varians

Skala Rasio







Numerik, 0 mutlak

- Data berupa angka
- O disini bernilai sebenarnya (O mutlak)
- Misalnya data kecepatan, saat kecepatannya 0 artinya benda tersebut tidak bergerak
- Statistik yang cocok untuk analisa datanya adalah frekuensi, mean, median, modus, varians







Ringkasan

Skala	Penjelasan	Contoh	Statistik deskriptif	grafik
Nominal	data kategorik tanpa urutan	Laki-laki dan perempuan	frekuensi, modus	lebih utama pie chart, bisa bar chart, tidak disarankan line chart
Ordinal	data kategorik dengan urutan	tingkat pendidikan	frekuensi, modus, median	bisa pie chart, lebih utama bar chart, tidak disarankan line chart
Interval	data numerik tanpa 0 mutlak	suhu di beberapa wilayah		tidak disarankan pie chart, untuk bar chart harus diubah
Rasio	data numerik dengan 0 mutlak	jumlah orang di suatu ruangan	frekuensi, mean, median, modus, varians	jadi interval tertentu dulu (atau biasanya disebut histogram), bisa juga line chart









04 STATUS MULTIVARIAT

- kovarians
- Korelasi









kovarians

kovarians:

- Memberikan ukuran yang menunjukkan sejauh mana dua variabel acak berubah bersama-sama.
- Meski mengukur hubungan antara dua variabel dalam kumpulan data yang diberikan, namun nilai hubungan ini masih bergantung pada varians data, sehingga belum terstandarisasi (data bervarians tinggi kovariansnya bisa tinggi pula, begitu pun sebaliknya).
- Nilainya bisa dari -∞ hingga ∞ (karena belum terstandarisasi)
- Misalkan E(x) adalah nilai ekspektasi dari suatu variabel x, dan E(y) adalah nilai ekspektasi dari variabel y, maka kovarians antara x dan y diberikan oleh:

cov(x, y) = E[xy] E[x] E[y]







Sifat Kovarians

Diberikan konstanta 'a' dan variabel acak X, Y, dan Z, sifat-sifat berikut adalah:

$$Cov(X, X) = Var(X) \ge 0$$

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX, Y) = aCov(X, Y)$$

$$Cov(aX, Y) = aCov(X, Y)$$

 $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$.







Korelasi

Korelasi:

- Ukuran yang menunjukkan hubungan antar 2 variabel acak
- Nilainya terletak pada kisaran -1 dan +1. (sudah terstandarisasi sehingga lebih mudah diinterpretasikan daripada kovarians)
- Semakin tinggi nilai koefisien korelasinya (tanpa melihat atau +) maka hubungan antar 2 variabel semakin erat
- - atau + menunjukkan bagaimana hubungan kedua variabel, misalnya :
- ☐ Hubungan A dan B –0.9, ini menunjukkan hubungan A dan B sangat kuat, namun ketika nilai A semakin besar, nilai B akan semakin kecil
- ☐ Hubungan C dan D 0.2, ini menunjukkan hubungan C dan D cenderung kecil, ketika nilai A semakin besar, nilai B cenderung makin besar pula
- Rumus : $cor(x,y) = cov(x,y)/\sigma x y$

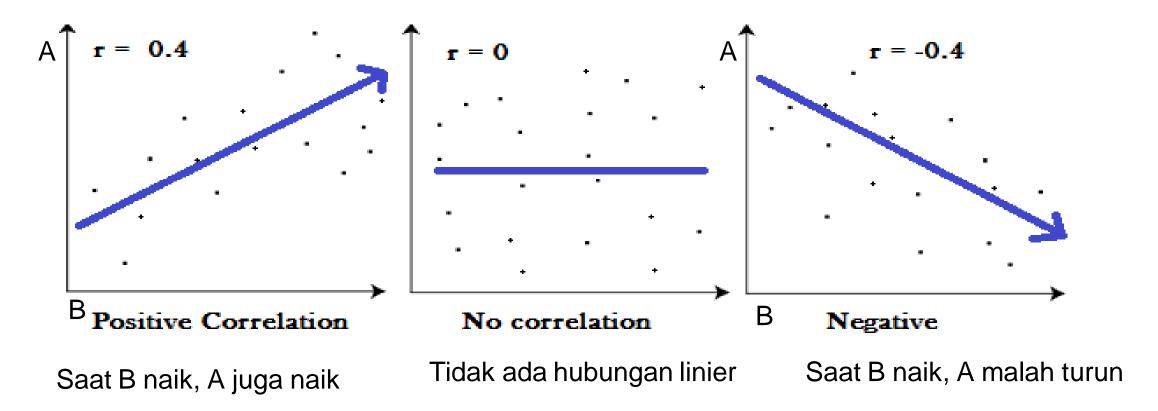




Artificial Intelligence Mastery Program

Koefisien Korelasi

Koefisien korelasi adalah cara untuk memberi nilai pada hubungan. Grafik berikut menunjukkan korelasi -1, 0 dan 1.



Besar keeratan hubungan





Artificial Intelligence
Mastery Program

 Menurut Guilford (1956) keeratan hubungan antar kedua variabel (positif/negatif hanya bentuk hubungannya seperti di slide sebelumnya):

ABSOLUTE VALUE OF R	INTERPRETATION		
< 0.19	Slight; almost no relationship		
0.20-0.39	Low correlation; definite but small relationship		
0.40-0.69	Moderate correlation; substantial relationship		
0.70-0.89	High correlation; strong relationship		
0.90-1.00	Very high correlation; very dependable relationship		
≥ 0.30	Practically significant relationship		

Korelasi ≠ Kausalitas



- Punya korelasi yang kuat bukan berarti berlaku hukum kausalitas.
- Jika 2 variabel saling berhubungan, misalnya A berhubungan erat dengan B, bukan berarti A mempengaruhi B.
- ☐ Jika A dan B berkorelasi bisa saja :
- A mempengaruhi B atau B mempengaruhi A (bisa keduanya, atau hanya salah 1)
- A mempengaruhi C, lalu C mempengaruhi A, makanya keduanya terlihat berhubungan. Misalnya Cuaca panas mempengaruhi rasa haus, rasa haus mempengaruhi intensitas minum, jadi pengaruh cuaca ke intensitas minum secara tidak langsung.
- C mempengaruhi A dan B, misalnya musim hujan mempengaruhi intensitas minum kopi dan masak mie (bukan karena habis minum kopi jadi pengen masak mie).
- Coincidential: kebetulan kejadiannya bersamaan.









05 KESIMPULAN

- Ulangan
- Ringkasan

Ringkasan

- Tipe data Data Terstruktur, Tidak Terstruktur, dan Semi Terstruktur.
- Data Terstruktur, organisasinya, contoh, kelebihan dan keterbatasannya.
- Data Tidak Terstruktur, contoh, kelebihan dan keterbatasan.
- Data semi terstruktur dan contohnya.
- Pengantar statistik beserta klasifikasinya: statistik deskriptif dan inferensial.
- Empat ukuran skala: skala nominal, ordinal, interval dan rasio.
- Ukuran Tendensi Sentral: mean, median dan modus.
- Ukuran variabilitas: jangkauan, kuartil dan jangkauan interkuartil, varians dan standar deviasi.
- Konsep Kovarians dan Korelasi.





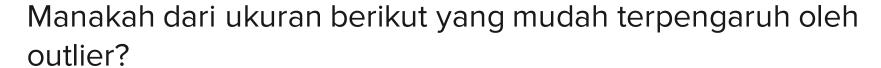






Ulangan

Pertanyaan

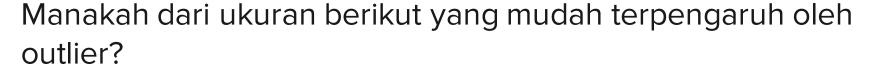


- A. Mean
- B. median
- C. Modus
- D. Kecondongan



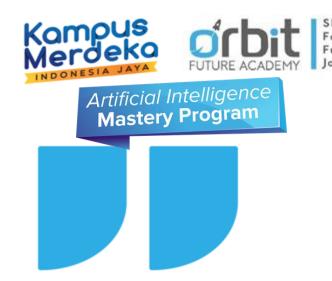
Ulangan

Pertanyaan



- A. Mean
- B. median
- C. Modus
- D. Kecondongan

Jawaban: A





Pertanyaan



- A. Mean
- B. median
- C. Modus
- D. Kecondongan





Pertanyaan

Manakah dari ukuran berikut yang cocok untuk data nominal?

- A. Mean
- B. median
- C. Modus
- D. Kecondongan

Jawaban: C









Artificial Intelligence
Mastery Program

TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia Center of Excellence (Jakarta Selatan) Gedung Veteran RI, Lt.15 Unit Z15-002, Plaza Semanggi JI. Jenderal Sudirman Kav.50, Jakarta 12930, Indonesia

- ☐ Jakarta Selatan/Pusat
- ☐ Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

Hubungi Kami

Director of Sales & Partnership ira@orbitventura.com +62 858-9187-7388

Social Media





@OrbitFutureAcademyIn1



iemy

Orbit Future Academy