Bayu Widodo

2025-08-29

Daftar Isi

Daftar Gambar

Sekolah Vokasi
School of Vocational Studies

1 Pengantar Data Mining

> Dunia digital semakin bergerak maju, data menjadi sesuatu yang sangat penting sebagai dasar mengambil keputusan bisnis. Salah satu profesi yang bergerak di bidang olah data, adalah Data Scientist yang menuntut kemampuan analisis statistik".

Statistika adalah ilmu yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan organisasi data. Dalam berbagai bidang kehidupan, statistika memainkan peran penting dalam membantu kita memahami dan menginterpretasikan informasi yang kompleks. Dengan menggunakan teknik statistika, kita dapat membuat keputusan yang lebih baik berdasarkan data dan fakta yang ada.

Istilah *Statistika* berbeda dengan *Statistik*. Statistik adalah data itu sendiri, informasi-nya, atau hasil penerapan algoritme statistika pada data tersebut. [id.wikipedia]
Pondasi dasar dari statistika adalah teori peluang. Semua teknik statistika dibangun di atas teori peluang yang merupakan bahasa matematika untuk mengukur derajat ketidakpastian.

1.1 Statistika & Statistik

Untuk mengenal lebih dekat apa itu statistik dan statistika, maka kita perlu mengetahui definisi keduanya.

1. **Statistik** sering diartikan sebagai kumpulan angka dalam bentuk tabel dan gambar, diagram atau grafik, mengenai suatu hal pada suatu waktu tertentu. Oleh karena itu statistik sering digunakan untuk menerangkan bidang-bidang tertentu, misalnya statistik pendidikan, statistik pertanian, statistik penduduk dsb. [id.wikipedia]
2. **Statistika** berkenaan dengan metode ilmiah untuk mengumpulkan, mengorganisasi, meringkas, menyajikan, menganalisa data termasuk menarik kesimpulan yang sah, dan membuat keputusan beralasan berdasarkan analisis tertentu. Pondasi dasar dari statistika adalah teori peluang. Semua teknik statistika dibangun di atas teori peluang yang merupakan bahasa matematika untuk mengukur derajat ketidakpastian.

Statistik lebih dimaknai sebagai data (informasi) yang digunakan untuk memahami fenomena atau menyelesaikan masalah tertentu. Dengan mengumpulkan, menganalisis, dan menginterpretasikan data, statistik membantu dalam pengambilan keputusan yang lebih baik. Selain itu, statistik dapat digunakan dalam berbagai bidang seperti ekonomi, kesehatan, pendidikan, dan ilmu sosial untuk menghasilkan wawasan (insight) yang lebih mendalam dan akurat.

Manfaat statistika sangat luas dan beragam, mulai dari dunia bisnis hingga ilmu pengetahuan, pemerintahan, kesehatan, dan teknologi. Berikut adalah beberapa manfaat utama dari statistika:

1. Pengambilan Keputusan yang Tepat:
   Statistika membantu dalam pengambilan keputusan yang lebih akurat dan berdasarkan data. Misalnya, dalam bisnis, analisis statistik dapat digunakan untuk memahami tren pasar, perilaku konsumen, dan efektivitas kampanye pemasaran.
2. Perencanaan dan Pengendalian Kualitas:
   Dalam industri manufaktur, statistika digunakan untuk perencanaan dan pengendalian kualitas. Teknik statistik seperti kontrol kualitas statistik (Statistical Quality Control) memungkinkan perusahaan untuk memonitor dan meningkatkan kualitas produk mereka.

3. Penelitian dan Pengembangan:
   Dalam bidang ilmu pengetahuan dan teknologi, statistika digunakan untuk merancang eksperimen, menganalisis data penelitian, dan menguji hipotesis. Ini membantu ilmuwan dan peneliti dalam mengembangkan teori baru dan teknologi inovatif.
4. Epidemiologi dan Kesehatan Masyarakat:
   Dalam bidang kesehatan, statistika digunakan untuk menganalisis data kesehatan masyarakat, memprediksi penyebaran penyakit, dan mengevaluasi efektivitas intervensi kesehatan. Hal ini sangat penting dalam pengendalian penyakit dan perencanaan program kesehatan.
5. Ekonomi dan Keuangan:
   Di bidang ekonomi, statistika digunakan untuk menganalisis data ekonomi, memprediksi tren ekonomi, dan membuat keputusan investasi. Dalam keuangan, statistika membantu dalam manajemen risiko dan analisis portofolio.
6. Pemerintahan dan Kebijakan Publik:
   Pemerintah menggunakan statistika untuk menyusun kebijakan publik yang efektif berdasarkan data sensus, survei, dan data administratif. Statistika membantu dalam perencanaan pembangunan, alokasi sumber daya, dan evaluasi program pemerintah.
7. Sosial dan Demografi:
   Statistika digunakan untuk menganalisis data sosial dan demografi, seperti pola migrasi, tingkat kelahiran, dan kematian. Informasi ini penting untuk perencanaan kebijakan sosial dan ekonomi.
8. Pemasaran dan Penjualan:
   Dalam pemasaran, analisis statistik digunakan untuk memahami perilaku konsumen, segmentasi pasar, dan efektivitas kampanye pemasaran. Ini membantu perusahaan dalam mengembangkan strategi pemasaran yang lebih efektif. 9 Olahraga:
   Di dunia olahraga, statistika digunakan untuk menganalisis kinerja pemain, strategi permainan, dan hasil pertandingan. Data statistik membantu pelatih dan manajer dalam membuat keputusan yang lebih baik terkait tim dan strategi.
9. Pendidikan:
   Dalam bidang pendidikan, statistika digunakan untuk mengevaluasi program pendidikan, menganalisis hasil ujian, dan memahami faktor-faktor yang mempengaruhi keberhasilan siswa. Hal ini penting untuk meningkatkan kualitas pendidikan dan pembelajaran.

Apa manfaat statistika untuk bidang Rekayasa Perangkat Lunak?
Manfaat statistika dalam rekayasa perangkat lunak meliputi peningkatan kualitas dan efisiensi pengembangan. Dengan statistika, pengembang dapat menganalisis performa, mengidentifikasi bug, dan menyelesaikan masalah lebih cepat. Statistika juga membantu dalam perencanaan proyek dan estimasi biaya, serta dalam analisis kebutuhan pengguna dan pengujian A/B, memastikan fitur baru memberikan manfaat yang diharapkan. Ini memungkinkan keputusan berdasarkan data dan peningkatan kualitas produk.

1.1.1 Ruang Lingkup Statistika

Berasarkan metodenya (Tahapan atau tujuan analisisnya), ruang lingkup statistika dikelompokkan dalam: (1) *Statistika Deskriptif*, dan (2) *Statistika Inferensial*.

1. Statistika Deskriptif / Non-Eksperimental
   Statistika deskriptif berkaitan dengan deskripsi data, menggambarkan informasi dari suatu data tersebut misalnya rata-rata, median, modus (mode), standart deviasi dan varian dari sekumpulan data yang dapat dianalisa dan divisualisasikan dengan tabel dan grafik agar mudah dibaca dan lebih bermakna. Statistika deskriptif digunakan untuk memperoleh pengetauan dari data. Penjelasan

atau pengetahuan berbagai karakteristik data melalui:
- Ukuran Pemusatan (Central Tendency): mode, mean, median, dll
- Ukuran Variabilitas/Dispersi: varians, deviasi stkitar, range, dll
- Ukuran Bentuk: skewness, kurtosis, plot boks
- Penyajian tabel dan grafik misalnya :
  - Distribusi Frekuensi
  - Histogram, Pie chart, Box-Plot dsb

2. Statistika Inferensial (Induktif, Probabilitas / Eksperimental.)
   - Statistika Inferensial merupakan cabang ilmu statistik yang berkaitan dengan penerapan metode metode statistik untuk menaksir dan / atau menguji karakteristik populasi yang dihipotesiskan berdasarkan data sampel.
   - Statistik inferensial digunakan untuk melakukan pengujian hipotesis, melakukan prediksi di masa depan dengan regresi, atau membuat klasifikasi suatu data dengan cara membuat model dan biasanya digunakan untuk melakukan pengambilan keputusan berdasarkan analisis data.

Tujuan dari statistika pada dasarnya adalah melakukan deskripsi terhadap data sampel, kemudian melakukan inferensi terhadap populasi data berdasar pada informasi (hasil statistika deskriptif) yang terkandung dalam sampel. Dengan demikian, dalam prakteknya kedua bagian statistika tersebut digunakan bersama-sama, umumnya dimulai dengan statistika deskriptif lalu dilanjutkan dengan berbagai analisis statistika untuk inferensi.

Berdasarkan asumsi distribusi yang digunakan dikenal :

1. **Statistika Parametrik**
   Teknik-teknik pengukuran statistik yang didasarkan pada asumsi tertentu, misalnya data yang diambil dari populasi yang berdistribusi normal. Teknik statistik ini digunakan untuk data yang berskala interval dan rasio.
2. **Statistika non-Parametrik**
   Teknik-teknik statistika yang menggunakan sedikit asumsi (atau bahkan tidak sama sekali) terkadang juga dikenal dengan model statistika yang bebas terhadap distribusi tertentu. Statistika non parametrik ini digunakan untuk menganalisis data berskala nominal dan ordinal.

1.2 Probabilitas

Teori Probabilitas didasarkan pada konsep dari suatu eksperimen **random**. Tiap proses yang menghasilkan data mentah atau proses untuk membangkitkan data mentah di mana keluaran individual tidak pasti tetapi ada distribusi yg regular dari keluaran untuk jumlah pengulangan yang banyak.

Probabilitas didifinisikan sebagai peluang atau kemungkinan terjadinya suatu peristiwa. Probabilitas dinyatakan dalam angka pecahan antara 0 sampai 1 atau dalam persentase. Probabilitas sangat berguna untuk pengambilan keputusan yang tepat, karena kehidupan di dunia tidak ada kepastian, sehingga diperlukan untuk mengetahui berapa besar probabilitas suatu peristiwa akan terjadi.

Kehidupan kita penuh dengan ketidakpastian. Sebagai seorang developer, kita harus bisa merencanakan suatu system yang akan dioperasikan sampai beberapa tahun kedepan. Maka kita harus bisa memperhitungkan segala sesuatu yang akan terjadi di masa depan, meskipun **secara pasti** kita tidak tahu apa yang terjadi di masa depan.

> Meskipun di masa depan banyak hal berubah, tetapi sebagian besar masih ada yang tetap sama atau setidak-tidaknya mirip dengan yang ada sekarang. Masa depan "tertanam" di masa kini

Apa Hubungan Probabilitas dengan Statistika?

Dalam kegiatan statistik, kita mengambil sampel dan kemudian menghasilkan ukuran statistik yang selanjutnya kita generalisasikan pada populasi. Generalisasi sampel terhadap populasi tersebut tentu saja mengandung unsur ketidakpastian. Unsur ketidakpastian tersebut dipelajari melalui ilmu peluang. Oleh karena itu peluang menjadi ilmu dasar dari statistika.

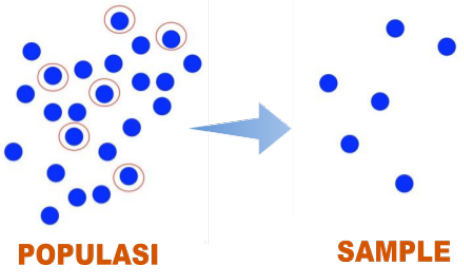
> Inilah keunikan dari Statistika : kemampuan menghitung ketidakpastian dengan tepat dan manjadi dasar untuk membuat pernyataan yang tegas dan lengkap

1.3 Populasi dan Sampel

Populasi adalah keseluruhan dari data secara lengkap dari suatu masalah yang sedang dianalisa. Populasi adalah sekelompok manusia, komunitas, atau subjek lainnya yang diteliti dan memiliki karakteristik tertentu atau kumpulan yang lengkap dari suatu elemen atau unsur yang sejenis, akan tetapi dapat dibedakan satu sama lain karena nilai karakteristiknya berlainan.
Parameter adalah ukuran populasi seperti rata-rata populasi, standar deviasi populasi. Contoh parameter adalah mean (μ), standar deviasi (σ), proporsi (P) dan koefisien korelasi (ρ).

Sampel adalah sebagian dari data yang dianggap mewakili data populasi yang digunakan untuk dianalisa karena sulitnya mendapatkan data jika populasi berjumlah sangat besar. Sampel adalah sebagian subjek yang diteliti yang diambil dari populasi atau bagian dari populasi yang disebut juga contoh yang dapat mewakili obyek yang akan diselidiki.
Statistik (statistic) merupakan karakteristik yang diukur dari sampel. Karakteristik di sini berupa rata-rata($\bar{x}$) sampel, standar deviasi sampel (s). dan proporsi (p). Statistik adalah informasi atau data.

Data yang dimaksudkan di sini biasanya merujuk pada informasi kuantitatif berupa angka yang dikumpulkan melalui kegiatan pengumpulan data.
Nilai statistik merupakan penaksir (estimator) bagi nilai parameter, yang nilai sesungguhnya tidak pernah diketahui besarnya.



Gambar 1: Populasi dan Sample

1.4 Variabel

Variabel berasal dari kata “vary” dan “able”, yang artinya “berubah” dan “dapat”. Dengan demikian, variabel secara harfiah mengacu pada kemampuannya untuk berubah, yang memungkinkan setiap vari-

abel untuk memiliki nilai yang dapat bervariasi. Nilai-nilai ini dapat bersifat kuantitatif (terukur atau terhitung, yang dinyatakan dalam bentuk angka) atau kualitatif (atribut seperti jumlah atau derajat mutu).

Variabel merupakan elemen kunci dalam konteks penelitian. Dalam statistik, variabel didefinisikan sebagai konsep, kualitas, karakteristik, atribut, atau sifat dari suatu objek (baik itu orang, benda, atau tempat) yang nilainya berbeda-beda antara satu objek dengan objek lainnya. Peneliti menetapkan variabel ini untuk dipelajari dan untuk menarik kesimpulan.

Karakteristik adalah ciri khas tertentu dari objek yang sedang diteliti, yang membedakan objek tersebut dari yang lain. Objek yang menjadi fokus pengamatan disebut satuan pengamatan, sedangkan nilai spesifik dari variabel yang diamati disebut variate. Koleksi nilai yang diperoleh dari pengukuran atau penghitungan suatu variabel disebut data.

1.5 Data

Dalam statistika dikenal beberapa jenis data. Data merupakan kumpulan fakta atau angka atau segala sesuatu yang dapat dipercaya kebenarannya, sehingga dapat digunakan sebagai dasar menarik suatu kesimpulan. Data dapat dibedakan menjadi dua jenis, yakni **data kualitatif** (non-metrik, categorical) dan **data kuantitatif** (metrik). Data kualitatif berbentuk pernyataan verbal, simbol atau gambar atau data yang berbentuk kategori atau atribut, tidak dinyatakan dengan angka. Contoh Variabel yang datanya kualitatif, yaitu : warna (putih, merah), jenis kelamin (laki-laki, perempuan), tingkat pendidikan (SD, SMP, SMA), agama (Islam, Kristen, Katholik, Hindu, Budha.

Data kuantitatif adalah data yang berbentuk bilangan atau data kualitatif yang diangkakan. Sebagai contoh variabel agama yang datanya bersifat kualitatif yang diangkakan seperti : 1. Islam, 2. Protestan, 3. Katolik, 4. Hidhu, 5. Budha dan 6. Konghuchu. Angka-angka didepan nama agama hanya sekedar simbol dan tidak mempunyai makna matematis apa apa (Kita tidak bisa mengatakan bahwa $1 + 2$ adalah 3).

Menurut **kontinyuitasnya**, data kuantitatif dapat dibagi menjadi dua yakni data **diskrit dan kontinyu**. Data diskrit adalah data yang angka-angkanya memiliki kemungkinan nilai terbatas dan antara satu angka dengan angka yang lain jelas terpisah (penghitungan). Contohnya adalah banyak mobil di parkiran mall. Tidak mungkin ada 1,5 mobil di parkiran, yang ada hanyalah 7 mobil atau 1 mobil di parkiran (data hasil penghitungan).
Data kontinyu adalah data yang angka-angkanya memiliki kemungkinan nilai tidak terbatas dalam kisaran tertentu (pengukuran). Contohnya berat badan dapat dinyatakan 60,5 kg (data hasil pengukuran). Lihat ringkasan jenis data pada gambar 2.

Menurut waktu pengumpulannya data dikelompokkan atas **data berkala** (time series) dan **data cross section**.
Data time series adalah data yang terkumpul dari waktu ke waktu untuk memberikan gambaran perkembangan suatu kegiatan atau fenomena. Sedangkan data cross section adalah data yang terkumpul dalam suatu waktu tertentu untuk memberikan gambaran perkembangan keadaan atau kegiatan waktu itu.
Contoh data cross section : Data nilai ujian masuk IPB tahun 2020, data Sensus Penduduk tahun 2020 dsb.

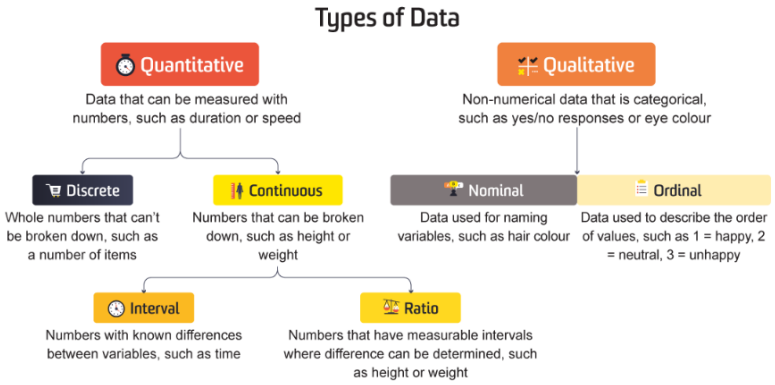
1.6 Skala Pengukuran

Pengukuran adalah dasar dari penyelidikan ilmiah. Segala sesuatu yang kita lakukan dimulai dengan pengukuran objek yang akan kita pelajari. Pengukuran adalah **pemberian angka atau kode pada suatu obyek**. Misalnya, orang (objek orang) dapat digambarkan dari beberapa karakteristik: umur, tingkat pendidikan, jenis kelamin, tingkat pendapatan.

Skala pengukuran merupakan seperangkat aturan yang diperlukan untuk mengkuantitatifkan data dari pengukuran suatu variable. Dalam melakukan analisis statistik, perbedaan jenis data sangat berpengaruh terhadap pemilihan model atau alat uji statistik. Ketidaksesuaian antara skala pengukuran dengan operasi matematik /peralatan statistik yang digunakan akan menghasilkan kesimpulan yang bias dan tidak tepat/relevan.

Skala pengukuran mengkategorikan dan mengukur data dengan cara tertentu. Terdapat empat skala pengukuran utama: **nominal**, **ordinal**, **interval**, dan **rasio**. Skala nominal mengklasifikasikan data ke dalam kategori tanpa urutan tertentu, seperti jenis kelamin atau warna rambut. Skala ordinal mengklasifikasikan data ke dalam kategori dengan urutan, seperti peringkat lomba atau tingkat pendidikan. Skala interval mengukur data dengan jarak antar nilai yang sama tetapi tidak memiliki nol mutlak, seperti suhu dalam Celsius atau Fahrenheit. Skala rasio, yang memiliki nol mutlak, mencakup data seperti berat badan, tinggi badan, atau pendapatan.

Dalam konteks analisis data, tipe dan skala pengukuran yang berbeda memerlukan metode analisis yang berbeda pula. Misalnya, data nominal paling baik dianalisis menggunakan frekuensi atau mode, sementara data ordinal lebih sesuai dianalisis menggunakan median atau rangking. Data interval dan rasio memungkinkan analisis yang lebih kompleks seperti mean dan stkitar deviasi. Dengan memahami tipe data dan skala pengukuran, kita dapat memilih metode analisis yang tepat dan membuat interpretasi yang akurat dalam penelitian atau analisis data.



Gambar 2: Tipe Data

1.6.1 Skala Nominal

Data yang dihimpun dapat dibedakan menjadi beberapa kategori tanpa memperhatikan urutan tertentu. Skala nominal sering juga disebut skala kualitatif adalah skala data yang berfungsi hanya untuk membedakan dan tidak ada tingkatan diantaranya.

Skala nominal adalah skala yang digunakan bukan untuk mengukur tetapi untuk membedakan secara klasifikasi. Bilangan atau angka digunakan untuk mawakili klasifikasi atau kategori. Sehingga fungsi bilangan hanyalah sebagai lambang pembeda.

Data nominal asalnya adalah sebagai non-numerik Contoh: suku bangsa, label, merek, nama, agama, dan sebagainya. Dalam mengidentifikasi hal-hal di atas digunakan **angka-angka sebagai symbol**. Sebagai contoh kita mengklasifikasi variable jenis kelamin: laki-laki kita beri simbol angka 1 dan wanita angka 2. Kita tidak dapat melakukan operasi arimatika dengan angka-angka tersebut, karena angka-angka tersebut hanya menunjukkan keberadaan atau ketidak-adanya karaktersitik tertentu.

1.6.2 Skala Ordinal

Skala Ordinal adalah skala pengukuran kualitatif dimana data diklasifikasikan ke dalam kelompok tertentu kemudian diberi kode, dan kode tersebut memiliki hierarki. Contoh: status sosial ekonomi (tinggi-menengah-rendah), kepangkatan dalam militer (perwira-bintara-tamtama), dan sebagainya.

Data ordinal biasanya terdapat dalam dua bentuk, yang pertama dalam bentuk ranking dan yang kedua dalam bentuk scalar. Rangking biasanya berbentuk tingkatan tertentu seperti, tidak suka, sedikit suka, suka, sangat suka. Biasanya digunakan untuk pengujian tingkat kesuakaan pada produk pangan, sebagai contoh, penggunaan skala likert. Skala Ordinal ini lebih tinggi daripada skala nominal, dan sering juga disebut dengan skala peringkat.

1.6.3 Skala Interval

Data yang dihimpun dapat diletakkan dalam skala dengan jarak (interval) antara dua titik skala diketahui dan skala tersebut tidak memiliki titik nol mutlak (titik pusat, memiliki titik nol yang didefinisikan dengan bebas). Contoh: tanggal lahir, suhu tubuh dalam skala Celcius, dan sebagainya.

Sebagai contoh, 20 derajat Celcius (20^0C) tidak berarti dua kali lipat panasnya dibandingkan 10 derajat Celcius. Misalnya pada pengukuran suhu. Kalau ada tiga daerah dengan suhu daerah A $= 10^0C$, daerah B $= 15^0C$ dan daerah C$= 20^0C$.
Kita bisa mengatakan bahwa selisih suhu daerah B, 5^0C lebih panas dibandingkan daerah A, dan selisih suhu daerah C dengan daerah B adalah 5^0C. (Ini menunjukkan pengukuran interval sudah memiliki jarak yang tetap).

Tetapi, kita **tidak bisa** mengatakan bahwa suhu daerah C dua kali lebih panas dibandingkan daerah A (artinya tidak bisa jadi kelipatan). Kenapa ? Karena dengan pengukuran yang lain, misalnya dengan Fahrenheit, di daerah A suhunya adalah 50^0F, di daerah B $= 59^0F$ dan daerah C$= 68^0F$.

Artinya, dengan pengukuran Fahrenheit, daerah C tidak dua kali lebih panas dibandingkan daerah A, dan ini terjadi karena dalam derajat Fahrenheit titik nolnya pada 32, sedangkan dalam derajat Celcius titik nolnya pada 0.

Skala interval mempunyai karakteristik seperti yang dimiliki oleh skala nominal dan ordinal dengan ditambah karakteristik lain, yaitu berupa adanya interval yang tetap.

1.6.4 Skala Rasio

Data yang dihimpun dapat diletakkan dalam skala dengan jarak antara dua titik skala diketahui dan skala tersebut memiliki titik nol mutlak. Contoh: usia, suhu ruang dalam skala Kelvin, dan sebagainya.

Hal lainnya adalah terdapat perbandingan yang sama antara data. Berat badan kelompok balita antara 0 sampai dengan 15 kg. Bayi 10 kg memiliki berat dua kali lipat dibandingkan dengan bayi 5 kg. Skala-skala pengukuran dalam ilmu pengetahuan alam sebagian besar adalah menggunakan skala ratio. Keunggulan dari skala ratio dibandingkan interval adalah kita dapat membandingkan suatu data dengan mudah.

Tabel 1: Ringkasan Skala Pengukuran

Kemampuan	Nominal	Ordinal	Interval	Rasio
Frekuensi	⋆	⋆	⋆	⋆
Modus	⋆	⋆	⋆	⋆
Median, Range		⋆	⋆	⋆
Mean,Stkitart Dev, Variansi			⋆	⋆
Mengurutkan		⋆	⋆	⋆
Membedakan Kuantitas			⋆	⋆
Penambahan, Pengurangan			⋆	⋆
Perkalian,Pembagian				⋆
Nol Mutlak				⋆

1.7 Analisis Statistik

Analisis data adalah proses memeriksa, membersihkan, mengubah, dan memodelkan data dengan tujuan menemukan informasi berguna, menarik kesimpulan, dan mendukung pengambilan keputusan. Proses ini melibatkan berbagai teknik statistik dan alat analisis untuk mengidentifikasi pola, tren, dan hubungan dalam data.

Dalam berbagai bidang, seperti bisnis, kesehatan, rekayasa, dan ilmu sosial, analisis data membantu dalam memahami fenomena kompleks dan membuat keputusan yang lebih baik berdasarkan bukti empiris. Langkah-langkah dalam analisis data biasanya mencakup pengumpulan data, pembersihan data untuk menghilangkan kesalahan atau data yang tidak relevan, analisis eksplorasi untuk memahami karakteristik dasar data, dan penggunaan model statistik atau algoritma pembelajaran mesin untuk membuat prediksi atau mengidentifikasi hubungan.

Hasil dari analisis data sering kali disajikan dalam bentuk visualisasi, seperti grafik atau tabel, untuk mempermudah pemahaman dan komunikasi temuan. Dengan semakin banyaknya data yang tersedia, analisis data menjadi keterampilan yang sangat penting dalam dunia modern, membantu organisasi dan individu dalam mengambil keputusan yang lebih tepat dan efektif.

Metode Analisis Data sangat ditentukan oleh: (1) tujuan penelitian, (2) banyaknya variabel, dan (3) sifat atau bentuk data.

1. Tujuan Penelitian
   Tujuan penelitian adalah faktor utama yang menentukan metode analisis data yang akan digunakan. Jika tujuan penelitian adalah untuk mendeskripsikan karakteristik dasar suatu populasi atau fenomena, maka analisis deskriptif seperti rata-rata, median, dan distribusi frekuensi akan digunakan. Jika tujuan adalah untuk memahami hubungan antara variabel, metode korelasi atau regresi mungkin lebih tepat. Untuk menguji hipotesis atau membuat prediksi, analisis inferensial atau teknik pembelajaran mesin dapat diterapkan. Misalnya, jika penelitian bertujuan untuk menentukan efektivitas sebuah pengobatan, metode seperti uji t atau ANOVA bisa digunakan untuk membandingkan kelompok perlakuan dan kontrol.

2. Banyaknya Variabel
   Jumlah variabel yang terlibat dalam analisis juga mempengaruhi metode yang dipilih. Analisis data univariat digunakan ketika hanya ada satu variabel, dengan fokus pada deskripsi statistik dasar. Analisis bivariat digunakan untuk dua variabel dan sering melibatkan metode seperti uji korelasi atau uji chi-square. Analisis multivariat, yang melibatkan lebih dari dua variabel, memerlukan teknik yang lebih kompleks seperti regresi bergkita, analisis faktor, atau analisis cluster. Metode yang dipilih harus mampu menangani kompleksitas data dan memberikan wawasan yang bermakna tentang interaksi antar variabel.
3. Sifat atau Bentuk Data
   Sifat atau bentuk data, termasuk jenis data (nominal, ordinal, interval, atau rasio) dan distribusi data (normal atau tidak normal), sangat mempengaruhi metode analisis yang sesuai. Data kategorikal, yang terdiri dari data nominal dan ordinal, sering dianalisis menggunakan tabel kontingensi atau uji chi-square. Data numerik, yang mencakup data interval dan rasio, dapat dianalisis menggunakan teknik seperti regresi atau ANOVA. Selain itu, jika data tidak mengikuti distribusi normal, metode non-parametrik seperti uji Mann-Whitney atau uji Kruskal-Wallis mungkin diperlukan. Bentuk data juga mencakup apakah data bersifat time-series atau cross-sectional, yang akan menentukan penggunaan metode seperti analisis deret waktu atau analisis regresi.

1.8 Istilah Penting

1. Data dan Variabel
   Dalam statistika dikenal beberapa jenis data. Data dapat berupa angka (metrik) dapat pula bukan berupa angka (non metrik). Data berupa angka disebut data kuantitatif dan data yang bukan angka disebut data kualitatif. Data merupakan kumpulan fakta atau angka atau segala sesuatu yang dapat dipercaya kebenarannya, sehingga dapat digunakan sebagai dasar menarik suatu kesimpulan.
2. Populasi dan Sampel
   - Populasi adalah keseluruhan dari data secara lengkap dari suatu masalah yang sedang dianalisa. Populasi adalah sekelompok manusia, komunitas, atau subjek lainnya yang diteliti dan memiliki karakteristik tertentu. Atau kumpulan yang lengkap dari suatu elemen atau unsur yang sejenis, akan tetapi dapat dibedakan satu sama lain karena nilai karakteristiknya berlainan.
   - Sampel adalah sebagian dari data yang dianggap mewakili data populasi yang digunakan untuk dianalisa karena sulitnya mendapatkan data jika populasi berjumlah sangat besar. Sampel adalah sebagian subjek yang diteliti yang diambil dari populasi (disebut juga contoh) yang dapat mewakili obyek yang akan diselidiki.
3. Statistik dan Parameter
   - Parameter adalah ukuran populasi seperti rata-rata populasi, stkitar deviasi populasi.
   - Statistik (statistic) adalah nilai (ukuran) yang diperoleh dari sampel seperti rata-rata sampel, stkitar deviasi sampel.
   - Nilai statistik merupakan penaksir (estimator) bagi nilai parameter, yang nilai sesungguhnya tidak pernah diketahui besarnya.

1.9 Latihan Soal

1. Apa yang dimaksud dengan skala nominal dalam pengukuran data? Berikan contoh.

   Jawaban:

Skala nominal adalah jenis skala pengukuran yang mengkategorikan data tanpa adanya urutan atau peringkat. Setiap kategori bersifat unik dan tidak ada nilai yang lebih tinggi atau lebih rendah dari yang lain. Contoh: Jenis kelamin (laki-laki, perempuan), warna mata (biru, hijau, coklat), atau jenis kendaraan (mobil, motor, sepeda).

2. Jelaskan perbedaan antara skala ordinal dan skala interval.

   Jawaban:

   - Skala ordinal mengukur data yang dapat diurutkan atau diberi peringkat, namun perbedaan antara nilai tidak dapat diukur secara tepat. Contoh: Peringkat lomba (juara 1, juara 2, juara 3).
   - Skala interval juga mengukur data yang dapat diurutkan, namun perbedaannya memiliki arti dan ukuran yang konsisten. Skala interval tidak memiliki titik nol absolut. Contoh: Suhu dalam Celsius atau Fahrenheit.

3. Sebutkan dua contoh variabel yang diukur dengan skala rasio dan jelaskan mengapa mereka termasuk dalam skala tersebut.

   Jawaban:

   - Contoh variabel yang diukur dengan skala rasio adalah berat badan dan tinggi badan.
   - Berat badan diukur dalam kilogram atau pon, memiliki jarak yang konsisten antara nilai, dan memiliki titik nol absolut (0 kg berarti tidak ada berat).
   - Tinggi badan diukur dalam meter atau sentimeter, memiliki jarak yang konsisten antara nilai, dan memiliki titik nol absolut (0 cm berarti tidak ada tinggi

4. Mengapa skala interval tidak memiliki titik nol absolut? Berikan contoh untuk memperjelas jawaban Anda.

   Jawaban:

   Skala interval tidak memiliki titik nol absolut karena nol pada skala interval tidak menunjukkan ketiadaan nilai, melainkan titik referensi yang arbitrer. Contoh: Suhu dalam Celsius. Nol derajat Celsius tidak berarti ketiadaan suhu, melainkan titik beku air. Oleh karena itu, perbandingan rasio tidak dapat dilakukan (misalnya, 20řC tidak dua kali lebih hangat dari 10řC)

5. Mengapa skala interval tidak memiliki titik nol absolut? Berikan contoh untuk memperjelas jawaban Anda.

   Jawaban:

   Skala interval tidak memiliki titik nol absolut karena nol pada skala interval tidak menunjukkan ketiadaan nilai, melainkan titik referensi yang arbitrer. Contoh: Suhu dalam Celsius. Nol derajat Celsius tidak berarti ketiadaan suhu, melainkan titik beku air. Oleh karena itu, perbandingan rasio tidak dapat dilakukan (misalnya, 20řC tidak dua kali lebih hangat dari 10řC).

6. Dalam sebuah survei kepuasan pelanggan, responden diminta untuk memilih jenis layanan yang mereka gunakan: Internet, TV Kabel, atau Telepon. Skala pengukuran apa yang digunakan dalam survei ini?

   Jawaban:

   Skala Nominal.

7. Sebuah perusahaan meminta karyawannya untuk menilai kinerja tahunan mereka dalam kategori: Sangat Buruk, Buruk, Cukup, Baik, dan Sangat Baik. Skala pengukuran apa yang digunakan dalam penilaian ini?

   Jawaban:

Skala Ordinal.

8. Dalam sebuah studi tentang preferensi warna mobil, responden diminta untuk memilih antara warna merah, biru, hijau, dan hitam. Jenis data dan skala pengukuran apa yang digunakan dalam studi ini?

   Jawaban:

   Jenis data: Kualitatif; Skala pengukuran: Skala Nominal.

9. Seorang peneliti sedang mempelajari hubungan antara tingkat pendidikan (SD, SMP, SMA, Sarjana) dan pendapatan bulanan (dalam rupiah) di sebuah kota. Peneliti menggunakan ANOVA untuk menganalisis data tersebut. Jelaskan tipe data dan skala pengukuran untuk masing-masing variabel, serta mengapa ANOVA merupakan metode analisis yang tepat.

   Jawaban:

   - Tingkat pendidikan adalah variabel kualitatif dengan skala Ordinal karena ada urutan dari SD ke Sarjana.
   - Pendapatan bulanan adalah variabel kuantitatif dengan skala Rasio karena memiliki nol absolut dan perbedaan antara nilai memiliki makna.
   - ANOVA merupakan metode analisis yang tepat karena digunakan untuk membandingkan rata-rata dari lebih dari dua kelompok (tingkat pendidikan) terhadap variabel kuantitatif (pendapatan bulanan).

1.10 Soal dan Tugas

1. Dalam sebuah penelitian tentang kepuasan pelanggan terhadap produk elektronik, peneliti mengumpulkan data berikut:
   (a) Nomor identifikasi pelanggan.
   (b) Tingkat kepuasan (1 = sangat tidak puas, 2 = tidak puas, 3 = netral, 4 = puas, 5 = sangat puas).
   (c) Usia pelanggan.
   (d) Pendapatan bulanan pelanggan.
   (e) Kategori produk yang dibeli (TV, Smartphone, Laptop, Tablet).

   Tentukan skala pengukuran untuk masing-masing jenis data yang dikumpulkan.

   Jawaban:

   - Nomor identifikasi pelanggan: Skala Nominal (hanya untuk identifikasi tanpa urutan atau makna numerik)
   - Tingkat kepuasan: Skala Ordinal (urutan kepuasan tanpa jarak yang sama antar nilai).
   - Usia pelanggan: Skala Rasio (memiliki nol absolut dan perbedaan antara nilai memiliki makna).
   - Pendapatan bulanan pelanggan: Skala Rasio (memiliki nol absolut dan perbedaan antara nilai memiliki makna).
   - Kategori produk yang dibeli: Skala Nominal (kategori tanpa urutan atau makna numerik).

2. Data merupakan elemen penting untuk membuktikan hipotesis ataupun menjawab pertanyaan penelitian. Dalam penelitian deduktif, data yang didapat dari kegiatan observasi digunakan untuk menguji hipotesis. Sementara itu, dalam penelitian induktif, generalisasi dari observasi dapat melahirkan teori. Dengan demikian, data menduduki posisi sentral dalam proses penelitian. Hal ini terlihat dari ungkapan "garbage in, garbage out (GIGO)", yang artinya jika data yang digunakan sebagai input adalah salah, maka output yang dihasilkan juga akan salah. Diskusikan!

   Jawaban:

Data adalah elemen krusial dalam penelitian, baik untuk membuktikan hipotesis (penelitian deduktif) maupun menghasilkan teori baru (penelitian induktif). Data menduduki posisi sentral karena kualitas hasil penelitian sangat bergantung pada kualitas data yang digunakan. Ungkapan "garbage in, garbage out (GIGO)" menekankan pentingnya akurasi data: jika data input salah, maka hasil penelitian juga akan salah. Oleh karena itu, validitas dan reliabilitas data harus dijaga.

Validitas dan reliabilitas data adalah dua konsep penting dalam penelitian yang berkaitan dengan kualitas data. Validitas mengukur sejauh mana data atau instrumen penelitian benar-benar mengukur apa yang seharusnya diukur. Data yang valid memastikan bahwa hasil penelitian benar-benar mencerminkan fenomena yang diteliti. Misalnya, jika tujuan penelitian adalah untuk mengukur kecerdasan, tes yang digunakan harus tepat mengukur aspek kecerdasan, bukan hal lain seperti memori atau keterampilan bahasa. Di sisi lain, reliabilitas mengukur konsistensi atau keandalan data atau instrumen penelitian dari waktu ke waktu. Data yang reliabel memberikan hasil yang konsisten ketika diukur ulang dalam kondisi yang sama. Sebagai contoh, jika sebuah tes kecerdasan diberikan kepada kelompok yang sama pada waktu yang berbeda, hasilnya harus serupa jika tes tersebut reliabel. Kedua konsep ini sangat penting untuk memastikan bahwa penelitian menghasilkan data yang akurat dan dapat dipercaya.

3. Statistik deskriptif atau deduktif adalah bagian dari statistik yang membahas suatu metode mengumpulkan, menyusun, menyajikan, mengolah, analisis, dan interpretasi hasil analisis dari data dengan cara yang informatif menggunakan gambar, grafik, atau diagram daripada menggunakan tulisan atau tabel. Berikan 5 contoh statistik deskriptif.

   Jawaban:

   1. Histogram: Grafik batang yang menunjukkan distribusi frekuensi data dalam interval tertentu. Berguna untuk menggambarkan bagaimana data terdistribusi dalam berbagai rentang nilai.
   2. Diagram Kotak (Box Plot): Menampilkan distribusi data dengan menunjukkan kuartil, median, dan nilai ekstrem (outliers). Ini membantu dalam memahami variabilitas dan skewness data.
   3. Grafik Lingkaran (Pie Chart): Menunjukkan proporsi atau persentase dari total data dalam bentuk irisan lingkaran. Ini cocok untuk menggambarkan komposisi kategori dari data.
   4. Diagram Batang (Bar Chart): Menampilkan data kategorikal dengan batang yang panjangnya mewakili nilai atau frekuensi dari kategori tersebut. Cocok untuk perbandingan antara kategori.
   5. Grafik Sebar (Scatter Plot): Menampilkan hubungan antara dua variabel dengan titik-titik di sumbu x dan y. Berguna untuk menganalisis korelasi atau pola antara dua variabel.

4. Statistik inferensi atau induktif adalah bagian dari statistik yang membahas suatu metode analisis, menaksir, meramalkan, menarik kesimpulan (konklusi), dan estimasi (perkiraan) terhadap data dari populasi berdasarkan sampel yang diambil secara acak dari populasi. Hasil dari analisis tersebut kemudian berlaku untuk populasi. Berikan 2 contoh statistik inferensi.

   Jawaban:

   1. Uji hipotesis digunakan untuk menentukan apakah perbedaan skor ujian antara mahasiswa Universitas A dan B signifikan secara statistik, sehingga bisa digeneralisasikan ke populasi mahasiswa kedua universitas.
   2. Interval kepercayaan memperkirakan rentang di mana rata-rata pengeluaran bulanan seluruh pelanggan kemungkinan berada, misalnya antara 4.8 juta hingga 5.2 juta, berdasarkan sampel 100 pelanggan.

5. Diberikan data gaji bulanan (dalam US$) pekerja perusahaan pertambangan tahun 2015. Apa

yang dapat kita baca dari data tersebut ?

1550	1310	1575	1675	1585	1590	1580	1475	1300	1650
1380	1730	1640	2000	1400	1323	1900	1600	1600	1555
1565	1320	1750	1750	1650	1740	1650	1875	1620	1550
1590	1570	2015	1620	1860	1625	2000	1850	1640	1900
1700	1380	1620	1650	2000	1455	1625	1340	1530	1410
1450	1815	1440	1420	1550	1550	1660	1760	1550	1650
1500	1620	1600	1580	1705	1780	1400	1550	1390	1600
1775	2025	1450	1425	1820	1990	1700	1900	1475	1850

Gambar 3: Gaji Bulanan Pekerja Tambang dalam US$ Tahun 2015

Pustaka

Alboukadel Kassambara. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, volume 1. STHDA, 2017.

D.T. Larose and C.D. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley Series on Methods and Applications in Data Mining. Wiley, 2014. ISBN 9780470908747. URL https://books.google.co.id/books?id=9hOpAwAAQBAJ.

Martin Maechler, Peter Rousseeuw, Anja Struyf, and Mia Hubert. Cluster analysis basics and extensions. Rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source), 2005.

Luis Alfonso Pérez Martos and Ángel Miguel García Vico. Clustering: an r library to facilitate the analysis and comparison of cluster algorithms, 12 2022. URL https://hdl.handle.net/10481/79296.