TUGAS INDIVIDU :

STATISTICS INTRODUCTION

Disusun untuk memenuhi tugas Mata Kuliah Pengantar Sains Data

Dosen Pengampu : **Eko Prasetio Widhi, S.Kom., M.Kom.**

Disusun oleh :

Ghulam Mushthofa 442023611060

PROGAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS DARUSSALAM GONTOR

2025

BAB I:

STATISTICS

1.1 Definisi Statistik

Perbedaan utama antara Statistik Deskriptif dan Statistik Inferensial terletak pada tujuan dan cakupannya.

1. **Statistik Deskriptif** Fokusnya adalah untuk meringkas, menggambarkan, dan menyajikan data dalam bentuk yang mudah dipahami. Tujuannya adalah untuk mendeskripsikan karakteristik dari data yang sedang diamati tanpa membuat kesimpulan apa pun tentang populasi yang lebih besar.
   - **Contoh:** Seorang dosen menghitung nilai rata-rata (mean), nilai tengah (median), dan nilai yang paling sering muncul (modus) dari hasil ujian di satu kelas. Hasilnya hanya digunakan untuk menggambarkan performa kelas tersebut pada ujian itu.
2. **Statistik Inferensial** Fokusnya adalah untuk membuat kesimpulan, prediksi, atau generalisasi tentang sebuah populasi besar berdasarkan data dari sampel yang lebih kecil yang diambil dari populasi tersebut. Statistik ini menggunakan probabilitas untuk mengukur tingkat ketidakpastian dalam kesimpulannya.
   - **Contoh:** Sebuah lembaga survei ingin mengetahui persentase penduduk Indonesia yang akan memilih kandidat A dalam pemilu. Alih-alih bertanya ke seluruh 270 juta penduduk (populasi), mereka mengambil sampel acak sebanyak 2.000 orang. Hasil dari sampel ini kemudian digunakan untuk mengestimasi (memperkirakan) persentase dukungan untuk kandidat A di seluruh Indonesia.

1.2 Jenis Data

Dalam statistik, data dapat diukur menggunakan empat skala pengukuran yang berbeda, yang menentukan jenis analisis apa yang bisa dilakukan terhadap data tersebut.

1. **Skala Nominal:** Skala ini digunakan untuk melabeli variabel tanpa memberikan nilai kuantitatif atau urutan apa pun. Ini adalah skala yang paling sederhana.
   - **Contoh:** Jenis Kelamin (Pria, Wanita), Warna Favorit (Merah, Biru, Hijau). Kita tidak bisa mengatakan bahwa Pria lebih tinggi dari Wanita atau sebaliknya.

2. **Skala Ordinal:** Skala ini mengkategorikan data dan juga memberikan urutan atau peringkat antar kategori tersebut. Namun, kita tidak bisa mengukur selisih atau jarak antar peringkat.
   - **Contoh:** Tingkat Kepuasan Pelanggan (Sangat Tidak Puas, Tidak Puas, Netral, Puas, Sangat Puas). Kita tahu "Puas" lebih tinggi dari "Netral", tetapi kita tidak tahu seberapa besar selisih kepuasannya.
3. **Skala Interval:** Skala ini memiliki semua sifat skala ordinal, ditambah dengan properti bahwa selisih atau jarak antar nilai dapat diukur dan memiliki arti. Namun, skala ini tidak memiliki titik nol mutlak (nilai nol tidak berarti ketiadaan).
   - **Contoh:** Suhu dalam Celcius. Selisih antara 30°C dan 20°C adalah 10°C, sama dengan selisih antara 20°C dan 10°C. Akan tetapi, 0°C bukan berarti tidak ada suhu sama sekali.
4. **Skala Rasio:** Ini adalah skala pengukuran tingkat tertinggi. Memiliki semua sifat skala interval dan juga memiliki titik nol mutlak yang bermakna.
   - **Contoh:** Tinggi Badan (cm). Seseorang dengan tinggi 180 cm adalah dua kali lebih tinggi dari seseorang dengan tinggi 90 cm. Nilai 0 cm berarti benar-benar tidak ada tinggi.

1.3 Ukuran Pemusatan Data

Diberikan data nilai ujian: 60, 75, 80, 85, 90.

- **Mean (Rata-rata)**

  Mean adalah jumlah semua nilai data dibagi dengan banyaknya data.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 75 + 80 + 85 + 90}{5} = \frac{390}{5} = 78$$

  Untuk *mean* dari data adalah **78.**

- **Median (Nilai Tengah)**

  Median adalah nilai yang berada tepat di tengah setelah data diurutkan dari yang terkecil hingga terbesar.

  Langkahnya, mengurutkan data: 60, 75, **80**, 85, 90.

  Dan nilai yang berada di posisi tengah ialah *median* dari data tersebut.

- **Modus (Nilai yang sering muncul)**

Modus adalah nilai yang memiliki frekuensi kemunculan tertinggi dalam set data. Analisis dalam data, yaitu : 60, 75, 80, 85, 90, setiap nilai hanya muncul satu kali Dan data tersebut tidak memiliki *modus*

1.4 Ukuran Penyebaran Data

Diberikan data tinggi badan (cm): 160, 165, 170, 175, 180.

- **Range (Jangkauan)**

  Range adalah selisih antara nilai maksimum dan nilai minimum dalam data.

  $$\text{Range} = \text{Nilai Maksimum} - \text{Nilai Minimum} = 180 - 160 = 20$$

  Jadi *range* dari data ialah **20 cm**

- **Variance (Varians)**

  Variance mengukur seberapa jauh setiap titik data tersebar dari nilai rata-rata (mean). Kita akan menggunakan rumus varians sampel (s^2).

  Langkah pertama, kita hitung mean $\bar{x} = \frac{160+165+170+175+180}{5} = \frac{850}{5} = 170$

  Lalu, hitung kuadrat selisih setiap data dengan mean:

  $$(160 - 170)2 = (-10)2 = 100$$

  $$(165 - 170)2 = (-5)2 = 25$$

  $$(170 - 170)2 = (0)2 = 0$$

  $$(175 - 170)2 = (5)2 = 25$$

  $$(180 - 170)2 = (10)2 = 100$$

  Lanjut, menjumlahkan semua hasil kuadrat selisih:

  $\sum(x_i - \bar{x})^2 = 100 + 25 + 0 + 25 + 100 = 250$

  bagi dengan $n - 1$ (dimana $n = 5$):

  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{250}{5-1} = \frac{250}{4} = 62.5$

  Jadi, *variance* sampelnya ialah **62.5**

- **Standart Deviation (Simpangan Baku)**

  Standard deviation adalah akar kuadrat dari variance. Ini memberikan ukuran penyebaran dalam satuan yang sama dengan data aslinya.

  $s = \sqrt{s^2} = \sqrt{62.5} \approx 7.91$

  Jadi, *standard deviation* sampelnya adalah sekitar **7.91 cm**.

BAB II:

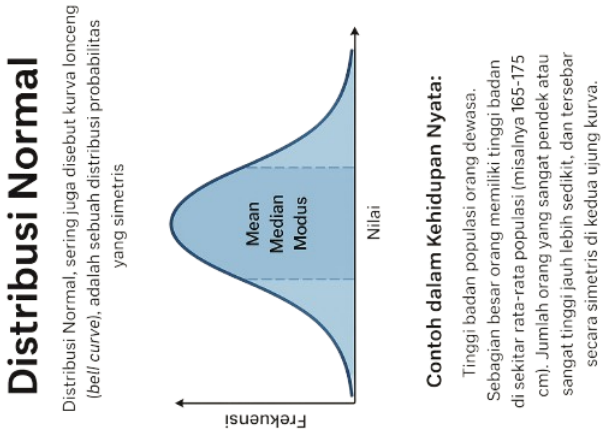
DISTRIBUSI DATA

2.1 Distribusi Normal

Distribusi Normal, sering juga disebut kurva lonceng (*bell curve*), adalah sebuah distribusi probabilitas yang simetris. Ciri utamanya adalah mayoritas data terkonsentrasi di sekitar nilai pusat (rata-rata), dan frekuensi data akan semakin menurun secara simetris seiring menjauhnya dari pusat. Pada distribusi normal yang sempurna, nilai mean, median, dan modus adalah sama.

- **Contoh dalam Kehidupan Nyata:**

  Tinggi badan populasi orang dewasa. Sebagian besar orang memiliki tinggi badan di sekitar rata-rata populasi (misalnya 165-175 cm). Jumlah orang yang sangat pendek atau sangat tinggi jauh lebih sedikit, dan tersebar secara simetris di kedua ujung kurva. Contoh lain yang sering mengikuti distribusi normal adalah skor IQ, tekanan darah, dan kesalahan pengukuran dalam eksperimen.



Gambar 2.1

BAB III:

PROBABILITY (PELUANG)

3.1 Dadu yang dilempar

Soal: Sebuah dadu dilempar sekali. Berapa peluang keluar angka genap?

Langkah nya, menentukan ruang sampel: Ruang sampel adalah himpunan semua kemungkinan hasil yang bisa terjadi. Untuk sebuah dadu, hasilnya adalah:

$S = 1,2,3,4,5,6$. Jumlah total kemungkinan hasil, $n(S) = 6$.

Tentukan kejadian yang diharapkan, kejadian yang diharapkan Adalah munculnya angka genap

$A = 2,4,6$. Jumlah hasil yang diharapkan, $n(A) = 3$.

Menghitung Peluang, peluang suatu kejadian dihitung dengan membagi jumlah hasil yang diharapkan dengan jumlah total kemungkinan hasil.

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

Peluang keluar angka genap adalah **1/2** atau **0.5**.

BAB IV:

SAMPLING

4.1 Pengambilan Sampling

Berikut adalah perbedaan antara tiga teknik sampling umum:

- **Random Sampling (Pengambilan Sampel Acak Sederhana):** Setiap anggota dalam populasi memiliki kesempatan yang sama persis untuk terpilih sebagai sampel. Prosesnya seperti undian.
    - **Contoh:** Untuk memilih 50 mahasiswa dari total 1.000 mahasiswa untuk sebuah survei, nama atau NIM semua mahasiswa dimasukkan ke dalam sistem komputer, lalu komputer akan memilih 50 nama secara acak tanpa preferensi apa pun.
- **Stratified Sampling (Pengambilan Sampel Berstrata):** Populasi dibagi terlebih dahulu menjadi beberapa sub-kelompok (strata) yang homogen berdasarkan karakteristik tertentu (misal: usia, jenis kelamin, jurusan). Kemudian, sampel acak sederhana diambil dari setiap strata tersebut, seringkali secara proporsional.
    - **Contoh:** Untuk mensurvei kepuasan mahasiswa di sebuah fakultas, populasi dibagi per angkatan (strata: 2022, 2023, 2024). Jika angkatan 2024 jumlahnya 50% dari total, maka 50% dari total sampel akan diambil secara acak dari mahasiswa angkatan 2024.
- **Systematic Sampling (Pengambilan Sampel Sistematis):** Sampel dipilih dari populasi dengan interval yang tetap (disebut interval *k*). Anggota pertama dipilih secara acak, lalu anggota-anggota berikutnya dipilih dengan melompati sebanyak k anggota dari daftar populasi.
    - **Contoh:** Sebuah perusahaan ingin menyurvei 100 karyawan dari total 1.000 karyawan. Intervalnya adalah

    $k = 1000/100 = 10$. Mereka memilih satu karyawan pertama secara acak (misalnya, karyawan nomor 7), maka sampel berikutnya adalah karyawan nomor 17, 27, 37, dan seterusnya hingga terpilih 100 karyawan.

BAB V:

OUTLIER

5.1 Definisi Outlier

Outlier (pencilan) adalah sebuah titik data yang nilainya secara signifikan berbeda atau ekstrem dibandingkan dengan sebagian besar data lainnya dalam satu kumpulan data.

Cara Mendeteksi Outlier:

1. **Visualisasi:** Menggunakan **Box Plot** adalah cara yang sangat umum. Titik data yang berada di luar "kumis" (*whiskers*) dari box plot biasanya dianggap sebagai outlier.
2. **Metode Statistik:** Menggunakan **Interquartile Range (IQR)**. Sebuah nilai dianggap outlier jika berada di bawah

   $Q1 - 1.5$
   $timesIQR\ atau\ di\ atas\ Q3 + 1.5$

   $timesIQR.$

   Pengaruh Outlier:

   Outlier dapat sangat **mendistorsi hasil analisis statistik**. Ukuran pemusatan data seperti **mean (rata-rata)** sangat sensitif terhadap outlier dan bisa bergeser secara drastis, membuatnya tidak lagi mewakili pusat data yang sebenarnya. Ukuran penyebaran seperti **standard deviation** juga akan meningkat, memberikan kesan bahwa data lebih tersebar dari yang sebenarnya.

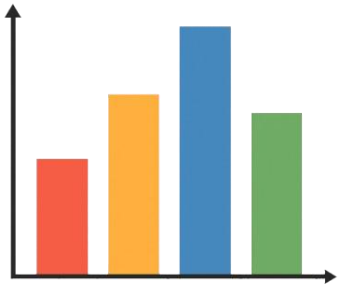
BAB VI:

DATA VISUALIZATION

6.1 Data Visualization

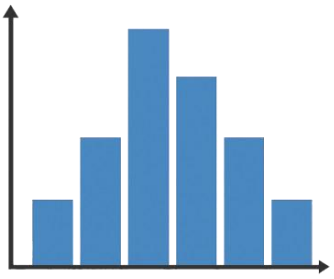
3 Jenis diagram yang umum digunakan dalam statistik untuk menampilkan data:

1. **Bar Chart (Diagram Batang):**



- **Kapan digunakan:** Digunakan untuk membandingkan nilai atau frekuensi antar beberapa kategori yang berbeda (data kualitatif/diskrit). Setiap batang mewakili satu kategori.
- **Contoh:** Membandingkan jumlah penjualan antara produk A, B, dan C pada bulan Juli.

2. **Histogram:**

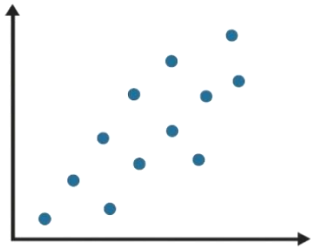


- **Kapan digunakan:** Digunakan untuk menampilkan distribusi frekuensi dari data numerik yang kontinu. Sumbu horizontalnya adalah rentang nilai data yang

dibagi menjadi interval (*bins*), dan sumbu vertikal menunjukkan frekuensi data yang jatuh dalam setiap interval.

- **Contoh:** Menampilkan distribusi tinggi badan mahasiswa dalam satu universitas.

3. **Scatter Plot (Diagram Pencar):**



- **Kapan digunakan:** Digunakan untuk menampilkan hubungan atau korelasi antara dua variabel numerik. Setiap titik pada plot mewakili satu pasangan nilai dari kedua variabel tersebut.
- **Contoh:** Memvisualisasikan hubungan antara lama waktu belajar (jam) dengan nilai ujian yang diperoleh.

BAB VII:

KORELASI

7.1 Definisi Koefisien Korelasi

Koefisien korelasi adalah sebuah nilai statistik yang mengukur kekuatan dan arah hubungan linear antara dua variabel kuantitatif. Nilainya selalu berada di antara -1 dan +1.

Cara Menginterpretasikan Nilainya:

- **Nilai mendekati +1:** Menunjukkan **korelasi positif yang kuat**. Artinya, ketika satu variabel meningkat, variabel lainnya cenderung meningkat juga. Contoh: Hubungan antara tinggi badan dan ukuran sepatu.
- **Nilai mendekati -1:** Menunjukkan **korelasi negatif yang kuat**. Artinya, ketika satu variabel meningkat, variabel lainnya cenderung menurun. Contoh: Hubungan antara jumlah jam bermain game dengan IPK.
- **Nilai mendekati 0:** Menunjukkan **tidak ada hubungan linear** atau hubungan yang sangat lemah antara kedua variabel. Bukan berarti tidak ada hubungan sama sekali, tetapi hubungannya tidak linear.