Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

Cara Berpikir STATISTIK dalam mengolah Data dengan Excel



YAYASAN PRIMA AGUS TEKNIK
UNIVERSITAS SAINS & TEKNOLOGI KOMPUTER
UNIVERSITAS STEKOM

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

Cara Berpikir STATISTIK

dalam mengolah Data dengan Excel



YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Cara Berpikir STATISTIK dalam mengolah Data dengan Excel

Penulis :
Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

ISBN :

Editor :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom.

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. 08122925000
Fax. 024-6710144
Email : info@stekom.ac.id

Hak cipta dilindungi undang-undang
Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara apapun tanpa ijin dari penulis

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat, karunia, kesehatan, serta kemudahan yang telah diberikan sehingga buku berjudul "*Cara Berpikir Statistik dalam Mengolah Data dengan Excel*" ini dapat diselesaikan dengan baik. Penyusunan buku ini merupakan upaya untuk menyediakan bahan ajar yang mampu menjembatani pemahaman konsep statistika dengan keterampilan pengolahan data secara praktis menggunakan Microsoft Excel.

Kehadiran buku ini dilatarbelakangi oleh kebutuhan untuk memahami data secara sistematis, kritis, dan bermakna di tengah melimpahnya informasi pada era digital. Statistika bukan sekadar kumpulan rumus atau perhitungan angka, melainkan sebuah cara berpikir untuk menggambarkan realitas, mengenali pola, menilai ketidakpastian, serta mendukung pengambilan keputusan berbasis data. Kemampuan ini menjadi semakin penting bagi mahasiswa, peneliti, praktisi bisnis, dan siapa pun yang berkecimpung dalam bidang sains data, teknologi informasi, maupun pengelolaan organisasi.

Buku ini disusun sebagai panduan belajar yang mengintegrasikan konsep dasar statistika dengan penerapan praktis menggunakan Microsoft Excel. Pembahasan diawali dengan hakikat dan esensi statistika, pemahaman data sebagai dasar analisis, peringkasan data dalam bentuk tabular, tendensi sentral dan variabilitas, bentuk distribusi, serta visualisasi data. Selanjutnya, buku ini mengantar pembaca memahami hubungan antarvariabel melalui korelasi dan regresi, konsep probabilitas dan distribusi probabilitas, teori pengambilan sampel, estimasi, pengujian hipotesis, hingga berbagai prosedur uji statistik yang umum digunakan.

Pembahasan dalam buku ini dirancang secara bertahap, mulai dari konsep yang bersifat deskriptif hingga inferensial. Setiap materi disertai contoh, studi kasus, dan panduan penerapan dalam Excel agar pembaca tidak hanya memahami teori statistika, tetapi juga mampu menerapkannya dalam pengolahan data nyata. Penggunaan Excel dipilih karena mudah diakses, familiar bagi banyak pengguna, serta cukup andal untuk melakukan perhitungan statistik, pembuatan tabel, PivotTable, grafik, dan analisis data dasar. Pendekatan ini sejalan dengan kebutuhan pembelajaran statistika terapan yang menekankan pemahaman konsep sekaligus keterampilan penggunaan perangkat lunak.

Melalui buku ini, pembaca diharapkan mampu mengembangkan pola pikir statistik yang tidak hanya berfokus pada hasil perhitungan, tetapi juga pada kualitas data, kecermatan interpretasi, etika pengelolaan data, dan relevansi temuan terhadap persoalan yang dihadapi. Seluruh materi dalam buku ini disusun untuk mendukung capaian pembelajaran mahasiswa, khususnya dalam membangun kompetensi dasar statistika dan pola pikir analitis sebagai bekal menuju sains data. Pada bagian penutup, pembaca diperkenalkan dengan arah pengembangan kompetensi lanjutan, termasuk penguasaan bahasa pemrograman Python, R, dan SQL, pemanfaatan business intelligence, serta strategi pembelajaran sepanjang hayat. Pemahaman statistika yang kokoh akan menjadi modal penting ketika pembaca melangkah ke analisis data tingkat lanjut, pemodelan prediktif, maupun penerapan kecerdasan buatan.

Penulis menyadari bahwa penyusunan buku ini masih memiliki keterbatasan, baik dari segi kedalaman materi, contoh kasus, penyajian, maupun cakupan pembahasan. Oleh karena itu, kritik dan saran yang membangun dari para pembaca, dosen, praktisi, maupun pengguna buku ini sangat diharapkan demi penyempurnaan pada edisi mendatang.

Akhir kata, penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, masukan, motivasi, dan bantuan selama proses penyusunan buku ini. Semoga buku ini dapat memberikan manfaat bagi mahasiswa, dosen, peneliti, praktisi data, pelaku usaha, serta siapa pun yang ingin mengembangkan kemampuan berpikir statistik dan mengolah data secara sistematis menggunakan Excel. Semoga Tuhan Yang Maha Esa senantiasa memberikan kemudahan dan keberkahan kepada seluruh pihak yang telah terlibat dalam proses penyusunan buku ini.

Semangat & Selamat Membaca….!!!!!

Semarang, Oktober 2026
Penulis

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM

DAFTAR ISI

KATA PENGANTAR ..... **i**
DAFTAR ISI ..... **iii**
BAB 1 HAKIKAT DAN ESENSI STATISTIK **1**
1.1 Statistika Sebagai Bahasa Untuk Memahami Dan Menginterpretasikan Data .. 1
1.2 Pengertian Dan Cabang Statistika 1
1.3 Peran Statistika Dalam Sains Data 3
1.4 Pola Pikir Statistik Dalam Analisis Data 5
1.5 Penerapan Statistika Dalam Praktik Sains Data 6
1.6 Statistik Dalam Kehidupan Sehari-Hari 8
1.7 Memahami Statistika Sebagai Keterampilan Praktis 8
1.8 Statistika Dalam Pengambilan Keputusan 9
1.9 Capaian Pembelajaran dan Penguasaan Kompetensi Statistik Sains Data 10
1.10 Ikhtisar Dasar Statistika Dalam Sains Data 11
BAB 2 DATA SEBAGAI DASAR ANALISIS STATISTIK **13**
2.1 Pengantar Data Dalam Statistika 13
2.2 Pengertian Data Dan Informasi 14
2.3 Struktur Dan Komponen Data 15
2.4 Jenis Data: Memahami Variasinya 1
2.5 Sumber Data Dan Pengaruhnya Terhadap Validitas Analisis Statistik 8
2.6 Kualitas Data: Faktor Penentu Keberhasilan Atau Kegagalan 12
2.7 Etika Dan Tanggung Jawab Data Dalam Pengelolaan Data 16
2.8 Siklus Hidup Data: Dari Perancangan Hingga Penghapusan 18
2.9 Pertemuan Pertama Anda Dengan Kumpulan Data (*Dataset*) 22
2.10 Ikhtisar Pengelolaan Dan Pemahaman Data 23
BAB 3 RINGKASAN DASAR & TAMPILAN TABULAR **25**
3.1 Agregasi Data Untuk Mengungkap Pola Dan Tren 25
3.2 Agregasi Data: Penghitungan, Persentase, Dan Ukuran Relatif 25
3.3 Distribusi Frekuensi Untuk Meringkas Pola Data 32
3.4 Tabel Dua Arah – Menjelajahi Hubungan 34
3.5 Analisis Data Dengan Pivottable Di Excel 37
3.6 Alur Kerja Peringkasan Data 43
3.7 Statistik SDM Untuk Mengidentifikasi Pola Dan Ketimpangan Organisasi 46
3.8 Ringkasan Data Sebagai Fondasi Penalaran Statistik 49
BAB 4 TENDENSI SENTRAL DAN VARIABILITAS **52**
4.1 Tendensi Sentral Dan Variabilitas Dalam Analisis Statistik Deskriptif 52
4.2 Ukuran Tendensi Sentral: Menemukan Inti Dari Data Anda 54
4.3 Ukuran Variabilitas: Memahami Sebaran Data 58
4.4 Studi Kasus Kepuasan Karyawan: Analisis Tendensi Sentral dan Sebaran Data . 65

4.5 Interpretasi Tendensi Sentral Dan Variabilitas Dalam Analisis Statistik 67
BAB 5 BENTUK DISTRIBUSI, KEMIRINGAN, & KURTOSIS 69
5.1 Bentuk Distribusi Dan Penerapannya Dalam Analisis Statistik 69
5.2 Bentuk Distribusi Dan Karakteristik Data 70
5.3 Kemiringan Distribusi (*Skewness*) Dan Interpretasinya 76
5.4 Kurtosis: Mengukur Bobot Ekor Dan Konsentrasi Distribusi 78
5.5 Memvisualisasikan Bentuk Distribusi 80
5.6 Studi Kasus: Analisis Distribusi Pendapatan 85
5.7 Statistik Deskriptif Dengan Excel Data Analysis Toolpak 87
5.8 Pemusatan Data Dan Menelaah Bentuk Distribusi Data 88
BAB 6 SENI VISUALISASI DATA 90
6.1 Visualisasi Data Dan Prinsip ART Dalam Analisis Data 90
6.2 Peran Visualisasi Dalam Sains Data 91
6.3 Kerangka Kerja ART Untuk Visualisasi 92
6.4 Memilih Jenis Grafik Yang Tepat 94
6.5 Alat Visualisasi Excel 105
6.6 Merancang Untuk Kejelasan Di Excel 111
6.7 Visualisasi Data Eksploratori 114
6.8 Visualisasi Dinamis Dan Interaktif 117
6.9 Bercerita Dengan Data Untuk Mendukung Pengambilan Keputusan 121
6.10 Prinsip Dan Teknik Visualisasi Data Untuk Analisis Dan Komunikasi Wawasan . 125
BAB 7 HUBUNGAN DALAM DATA ANDA – KORELASI DAN SELEBIHNYA 127
7.1 Memahami Hubungan Antarvariabel 127
7.2 Analisis Hubungan Antarvariabel: Kekuatan, Arah, Dan Metode Analisis 128
7.3 Hubungan Antara Variabel Kategorikal 129
7.4 Hubungan Antara Variabel Kategorikal Dan Numerik 133
7.5 Hubungan Antar Variabel Numerik – Analisis Korelasi 136
7.6 Menyatukan Semuanya – Kerangka Kerja Pengambilan Keputusan 146
7.7 Menganalisis Dan Menafsirkan Hubungan Antarvariabel 149
BAB 8 PENGANTAR ANALISIS REGRESI 151
8.1 Jembatan Antara Deskripsi Dan Prediksi 151
8.2 Dari Korelasi Menuju Prediksi: Menarik Garis 152
8.3 Memahami Analisis Regresi 152
8.4 Regresi Linear Sederhana: Kisah Tentang Garis Lurus 154
8.5 Regresi Linear Berganda 165
8.6 Ringkasan Regresi Linear Berganda 168
8.7 Asumsi Regresi Linear 169
8.8 Pedoman Praktis Analisis Regresi 171
8.9 Regresi Dalam Konteks: Penerapan Di Dunia Nyata 172
8.10 Keterbatasan Regresi Dan Faktor Yang Memengaruhi Keandalan Model 173
8.11 Ringkasan Analisis Regresi 174

8.12 Regresi Sebagai Fondasi Analitik Prediktif Dan Inferensi Statistik 175
BAB 9 PROBABILITAS — BAHASA KETIDAKPASTIAN 177
9.1 Peran Probabilitas Dalam Analisis Data 177
9.2 Memahami Probabilitas: Dari Intuisi Menuju Matematika 177
9.3 Jenis-Jenis Kejadian Dalam Probabilitas 181
9.4 Aturan Probabilitas 183
9.5 Probabilitas Bersyarat 187
9.6 Teorema Bayes: Pembaruan Keyakinan Dan Analisis Probabilitas Bersyarat 190
9.7 Ringkasan Probabilitas 193
BAB 10 DISTRIBUSI PROBABILITAS 195
10.1 Pengantar Distribusi Probabilitas 195
10.2 Dasar-Dasar Variabel Acak 196
10.3 Konsep Dasar Distribusi Probabilitas 198
10.4 Distribusi Diskrit 204
10.5 Distribusi Probabilitas Kontinu 212
10.6 Menghubungkan Probabilitas Dengan Data 226
10.7 Distribusi Probabilitas Dan Normal, Serta Penerapan Analisis Data 227
BAB 11 TEORI PENGAMBILAN SAMPEL DAN ESTIMASI 230
11.1 Teori Pengambilan Sampel Dan Estimasi Dalam Inferensi Statistik 230
11.2 Populasi, Sampel, Dan Bias Dalam Inferensi Statistik 231
11.3 Metode Pengambilan Sampel Dan Implikasinya 235
11.4 Pertimbangan Mengenai Ukuran Sampel 242
11.5 Distribusi Sampling 245
11.6 Distribusi Z Dan T Dalam Inferensi Statistik 248
11.7 Estimasi Titik Dan Estimasi Interval 257
11.8 Rangkuman Estimasi Dan Inferensi Statistik 268
BAB 12 DASAR-DASAR PENGUJIAN HIPOTESIS 270
12.1 Landasan Pengujian Hipotesis 270
12.2 Konsep Hipotesis Statistik 270
12.3 Logika Pengujian Hipotesis 271
12.4 Hipotesis: Null (H_0) Dan Alternatif (H_1) 272
12.5 Konsep Dasar Pengujian Hipotesis 274
12.6 Proses Pengujian Hipotesis: Prosedur Enam Langkah 279
12.7 Jenis Kesalahan Dan Kekuatan Statistik (*Statistical Power*) 285
12.8 Signifikansi Statistik Vs. Signifikansi Praktis 286
12.9 Selang Kepercayaan Dan Pengujian Hipotesis 287
12.10 Pemilihan Uji Statistik Berdasarkan Jenis Data Dan Desain Pembelajaran 289
12.11 Uji Hipotesis Statistik: Prosedur, Kesalahan, Dan Interpretasi Hasil 292
BAB 13 PROSEDUR STATISTIK UNTUK PENGUJIAN HIPOTESIS 294
13.1 Pemilihan Dan Penerapan Uji Statistik Parametrik Dan Nonparametrik 294
13.2 Gambaran Umum Uji Hipotesis Yang Umum Digunakan 294

13.3 Uji-Z (*Z-Test*): Penerapan Praktis 296
13.4 Uji-T (*T-Tests*): Penerapan Praktis 301
13.5 Uji Chi-Square (X^2): Penerapan Praktis 308
13.6 Uji F: Aplikasi Praktis (Contoh ANOVA) 313
13.7 Panduan Pengambilan Keputusan Dalam Pemilihan Uji Statistik 317
13.8 Penerapan Uji Statistik Dalam Analisis Data Dan Pengambilan Keputusan 318
BAB 14 MENGUJI HUBUNGAN — KORELASI DAN REGRESI 320
14.1 Korelasi Dan Regresi: Pengujian Signifikansi Hubungan Antarvariabel 320
14.2 Korelasi, Regresi, Dan Pengujian Hipotesis 321
14.3 Pengujian Untuk Korelasi Dan Regresi 324
14.4 Pengujian Signifikansi Korelasi Dan Regresi Linear Serta Validasi Model 329
BAB 15 MELANJUTKAN PERJALANAN ANDA DALAM SAINS DATA 330
15.1 Statistika Dan Analisis Data: Landasan Menuju Sains Data Dan AI 330
15.2 Kompetensi Dasar Statistika Dan Pola Pikir Analitis Dalam Sains Data 330
15.3 Perangkat Pemrograman Python, R, Dan SQL Untuk Analisis Data 331
15.4 Pembelajaran Python Untuk Pengembangan Kompetensi Sains Data 333
15.5 Peran BI Dalam Visualisasi Data Dan Pengambilan Keputusan Bisnis 333
15.6 Strategi Untuk Pembelajaran Sepanjang Hayat 334
15.7 Prinsip Dan Arah Pengembangan Diri Dalam Perjalanan Sains Data 335
DAFTAR PUSTAKA 337

BAB 1
HAKIKAT DAN ESENSI STATISTIK

1.1 STATISTIKA SEBAGAI BAHASA UNTUK MEMAHAMI DAN MENGINTERPRETASIKAN DATA

Maria menatap lembar kerja (*spreadsheet*) di layarnya. Ribuan baris data pembelian pelanggan terbentang tanpa henti berisi tanggal, jumlah transaksi, kategori produk, dan ID pengguna. Manajernya melontarkan pertanyaan yang terdengar sederhana:

> *"Apakah pelanggan kita lebih banyak berbelanja sejak kita merancang ulang situs web?"*

Datanya tersedia lengkap. Namun, lembar kerja itu tetap bungkam. Bagi Maria, rasanya seperti berdiri di tengah ruangan yang penuh sesak, di mana semua orang berbicara secara bersamaan tetapi dalam bahasa yang belum ia pahami. Lalu, tiba-tiba ia teringat sesuatu. Ia teringat pelajaran dari mata kuliah pengantar statistik: bandingkan nilai rata-rata sebelum dan sesudah perubahan, serta perhatikan seberapa besar variasi nilai-nilainya. Satu angka saja jarang bisa menceritakan keseluruhan kisah.

Dengan menggunakan Excel, ia menghitung nilai rata-rata pembelian untuk periode sebelum dan sesudah perancangan ulang situs web. Namun, ia tidak berhenti di situ; ia juga memeriksa variasinya sebaran dari nilai-nilai pembelian tersebut. Dalam hitungan menit, data itu mulai menyampaikan pesannya. Ternyata, rata-rata pembelian tidak hanya meningkat sebesar 15%, tetapi variasinya pun menjadi lebih sempit. Pelanggan tidak sekadar berbelanja lebih banyak secara rata-rata; mereka kini melakukan pembelian dengan jumlah yang lebih besar secara lebih konsisten. Data itu sebenarnya sudah berbicara sejak awal. Maria hanya perlu mempelajari bahasanya. Momen "aha!" saat angka-angka mentah berubah dari sekadar deretan digit yang membingungkan menjadi narasi, wawasan, dan bukti merupakan titik balik bagi setiap calon ilmuwan data (*data scientist*). Data, dengan sendirinya, tidak menyampaikan pesan apa pun. Data memerlukan interpretasi, struktur, dan cara pandang yang disiplin. Cara pandang yang disiplin itulah yang disebut sebagai statistika. Statistika bukanlah sekumpulan rumus yang menakutkan. Statistika juga bukan ujian untuk mengukur bakat matematika seseorang. Statistika adalah sebuah bahasa cara untuk mengungkap cerita yang ingin disampaikan oleh data. Statistika merupakan kerangka kerja yang mengubah "derau" (*noise*) menjadi pola, serta mengubah pertanyaan menjadi jawaban yang berlandaskan bukti. Dan inilah bahasa yang akan Anda pelajari.

1.2 PENGERTIAN DAN CABANG STATISTIKA

Mari kita mulai dengan menyingkirkan istilah-istilah teknis yang rumit dan menakutkan.

Statistika adalah ilmu untuk belajar dari data.

Statistika merupakan perpaduan antara seni dan disiplin ilmu yang mencakup proses pengumpulan, pengorganisasian, pemeriksaan, serta interpretasi informasi untuk pengambilan keputusan yang lebih baik di tengah ketidakpastian. Cara yang mudah untuk memahami statistika adalah dengan membayangkannya seperti pekerjaan seorang detektif. Anggaplah data sebagai petunjuk-petunjuk yang Anda miliki. Statistika menyediakan perangkat investigasi yang diperlukan untuk memeriksa petunjuk tersebut, membedakan pola yang bermakna dari "derau" acak, serta menyusun kesimpulan yang didasarkan pada bukti, bukan sekadar intuisi. Pada intinya, statistika membantu kita menjawab dua jenis pertanyaan mendasar:

1. Statistika Deskriptif: Apa yang Terjadi?

Statistika deskriptif berfokus pada meringkas dan menyajikan data yang sudah Anda miliki. Saat Anda menghitung nilai rata-rata ujian di sebuah kelas, menentukan menu yang paling sering dipesan, atau membuat grafik yang menunjukkan tren penjualan bulanan, Anda sedang menggunakan statistika deskriptif. Tujuannya adalah mengubah data mentah menjadi bentuk yang jelas dan mudah dipahami sehingga Anda dapat melihat pola, tren, dan karakteristik utama secara sekilas. Singkatnya, statistika deskriptif membantu Anda menceritakan kisah dari data yang telah Anda amati.

2. Statistika Inferensial: Apa yang Bisa Kita Simpulkan di Luar Data yang Ada?

Statistika inferensial melangkah lebih jauh. Metode ini memungkinkan Anda menggunakan sebagian data yang disebut sampel untuk menarik kesimpulan yang berdasar mengenai kelompok yang lebih besar, yaitu populasi. Sebagai contoh, ketika sebuah jajak pendapat politik menyurvei sekelompok kecil individu untuk memprediksi bagaimana jutaan orang mungkin memilih, atau ketika uji klinis menguji suatu pengobatan pada sejumlah pasien yang terbatas untuk mengevaluasi efektivitasnya, di situlah statistik inferensial berperan. Cabang statistik ini memungkinkan kita untuk memperluas wawasan melampaui apa yang kita amati secara langsung. Yang terpenting, hal ini dilakukan dengan memperhitungkan ketidakpastian secara eksplisit mengakui bahwa kesimpulan kita hanyalah estimasi, bukan kepastian mutlak.

Mengapa Statistik Ada?

Statistik muncul dari tantangan mendasar manusia: mengambil keputusan di dunia yang penuh ketidakpastian. Sepanjang sejarah:

- Para petani berupaya mengestimasi hasil panen di tengah cuaca yang tidak dapat diprediksi.
- Para pedagang mengevaluasi risiko saat mengangkut barang melalui rute yang panjang dan tidak menentu.
- Pemerintah membuat perencanaan bagi populasi yang tidak dapat diukur secara menyeluruh.

- Para ilmuwan menguji apakah hasil yang diamati mencerminkan pola yang nyata atau sekadar kebetulan. Dalam setiap kasus, masalahnya sama: informasi yang terbatas dan ketidakpastian yang melekat.

Kita tidak bisa mengamati segalanya, mengukur setiap kemungkinan, atau memprediksi masa depan dengan akurasi sempurna. Statistik menyediakan cara terstruktur untuk menghadapi keterbatasan ini membantu kita mengambil keputusan yang tepat sembari memahami tingkat keyakinan yang dapat kita berikan pada keputusan tersebut.

Sebuah Pemikiran Akhir

Sebelum Anda mempelajari perangkat canggih atau algoritma yang kompleks, ada satu keterampilan inti yang harus dikembangkan: kemampuan untuk berpikir secara statistik. Sebab dalam sains data, perangkat akan terus berkembang tetapi kemampuan untuk bernalar menggunakan data akan tetap menjadi aset Anda yang paling berharga.

1.3 PERAN STATISTIKA DALAM SAINS DATA

Jika sains data diibaratkan sebagai sebuah rumah, maka statistik adalah fondasinya. Setiap algoritma, setiap visualisasi, dan setiap model prediktif bersandar pada fondasi tersebut. Jika fondasi itu dihilangkan, bangunan mungkin masih bisa berdiri untuk sementara waktu tetapi akan menjadi tidak stabil, tidak dapat diandalkan, dan pada akhirnya tidak kokoh. Untuk memahami alasannya, mari kita telusuri siklus hidup sains data dan amati bagaimana cara berpikir statistik mendukung setiap tahapannya.

1. Pengumpulan dan Pembersihan Data

Sebelum pemeriksaan dimulai, kita harus bertanya: Data apa yang kita miliki? Data apa yang kurang? Apa saja yang mungkin terdistorsi? Penalaran statistik menjadi panduan dalam pengambilan keputusan tersebut. Hal ini membantu mengidentifikasi:

- Pencilan (*outlier*) yang mungkin menandakan kesalahan atau peristiwa langka namun bermakna
- Distribusi yang menceng (*skewed*) yang memengaruhi interpretasi
- Pola data yang hilang (*missingness*) yang dapat menimbulkan bias
- Masalah pengambilan sampel yang memengaruhi keterwakilan data

Tanpa pemahaman statistik, proses pembersihan data hanyalah sebuah tindakan coba-coba. Dengan pemahaman tersebut, penyiapan data menjadi sebuah evaluasi yang disiplin.

2. Pemeriksaan Data Eksploratif (*Exploratory Data Analysis*/EDA)

Di sinilah statistik benar-benar menunjukkan perannya. Anda mulai mengajukan pertanyaan seperti:

- Apa yang menjadi karakteristik umum dalam kumpulan data ini?
- Apa yang tidak biasa?
- Apakah variabel-variabel bergerak bersamaan?
- Apakah terdapat perbedaan yang berarti antar kelompok?

Ukuran-ukuran seperti nilai rata-rata (*mean*), median, simpangan baku (standar deviasi), dan korelasi memungkinkan kita untuk meringkas ribuan atau jutaan pengembangan teknologi masa depanke dalam bentuk yang mudah dipahami. Visualisasi seperti histogram dan diagram

kotak (*box plot*) bukan sekadar grafik; melainkan representasi visual dari prinsip-prinsip statistik. Melalui EDA (*Exploratory Data Analysis*), pola-pola mulai terlihat namun hanya dengan pola pikir statistik kita dapat menafsirkannya secara bertanggung jawab.

3. Pemodelan dan Pembelajaran Mesin (*Machine Learning*)

Pembelajaran mesin sering kali disajikan sebagai sesuatu yang terpisah dari statistik. Padahal, pada kenyataannya, bidang ini berakar kuat pada teori pembelajaran statistik (*statistical learning theory*). Saat membangun model prediktif, kita melakukan hal-hal berikut:

- Mengestimasi hubungan antarvariabel
- Mengukur ketidakpastian
- Mengoptimalkan parameter berdasarkan prinsip probabilitas
- Melakukan generalisasi dari data pelatihan ke data yang belum pernah dilihat sebelumnya

Baik menggunakan regresi linear maupun jaringan saraf tiruan (*neural network*) yang kompleks, tantangan utamanya tetap sama: mempelajari pola dari data masa lalu untuk memprediksi hasil di masa depan tanpa terkecoh oleh keacakan.
Pada dasarnya, ini adalah masalah inferensial.

4. Evaluasi Model

Sebuah model yang menghasilkan prediksi tidak serta-merta berguna. Evaluasi statistik menjawab pertanyaan-pertanyaan berikut:

- Seberapa sering model memberikan prediksi yang benar?
- Jenis kesalahan apa saja yang dihasilkan oleh model tersebut?
- Apakah kinerjanya benar-benar lebih baik daripada sekadar kebetulan?
- Seberapa stabil hasil yang diperoleh jika diterapkan pada sampel yang berbeda?

Metrik seperti akurasi, presisi, recall, selang kepercayaan (*confidence interval*), dan uji signifikansi menyediakan cara terstruktur untuk menilai keandalan model. Tanpa evaluasi statistik, sebuah model mungkin tampak mengesankan padahal sebenarnya memiliki cacat mendasar.

5. Komunikasi dan Pengambilan Keputusan

Yang mungkin paling krusial, statistik membentuk cara kita mengomunikasikan hasil. Terdapat perbedaan yang sangat besar antara pernyataan:

> *"Desain ulang kami meningkatkan penjualan."*

Dengan

> *"Pemeriksaan kami menunjukkan peningkatan nilai pembelian rata-rata sebesar 15%, dengan variabilitas yang terukur di berbagai segmen pelanggan serta bukti statistik yang mendukung perubahan tersebut."*

Statistik mendisiplinkan bahasa kita. Statistik mencegah klaim yang berlebihan, menyajikan kesimpulan dengan tingkat keyakinan yang tepat, serta menjamin kejujuran intelektual.

Kenyataannya sederhana: sains data tidak dapat dipisahkan dari statistik. Sains data adalah statistik yang diterapkan dalam skala besar. Anda mungkin bisa menulis kode, membuat grafik, bahkan membangun model yang tampak efektif; namun tanpa wawasan statistik, Anda ibarat berjalan dengan mata tertutup. Interpretasi menjadi rapuh, pengambilan keputusan menjadi berisiko, dan kesimpulan menjadi sulit untuk dipertanggungjawabkan. Statistika bukanlah pelengkap opsional dalam sains data (*data science*). Statistika adalah tulang punggung dari seluruh disiplin ilmu tersebut.

1.4 POLA PIKIR STATISTIK DALAM ANALISIS DATA

Apa yang membedakan seseorang yang sekadar menggunakan perangkat statistik dengan seseorang yang benar-benar berpikir secara statistik? Bukanlah jumlah rumus yang dihafal, bukan pula perangkat lunak yang dikuasai. Melainkan pola pikir.

1. Kenyamanan dalam Menghadapi Ketidakpastian

Pemikir statistik tidak mencari kepastian mutlak. Mereka menyadari bahwa sebagian besar keputusan di dunia nyata harus diambil dengan informasi yang tidak lengkap. Alih-alih mengabaikan ketidakpastian, mereka justru mengukurnya. Mereka menggunakan konsep seperti:

- Interval kepercayaan (*confidence intervals*)
- Probabilitas
- Margin kesalahan (*margins of error*)
- Kekuatan bukti

Mereka memahami bahwa estimasi yang diukur dengan cermat jauh lebih jujur dan bermanfaat dibandingkan klaim berani yang tidak didukung oleh bukti.

2. Kewaspadaan terhadap Asumsi

Sebelum pemeriksaan dimulai, muncul pertanyaan-pertanyaan mendalam:

- Apakah sampel yang digunakan sudah representatif?
- Mungkinkah ada bias yang mendistorsi hasil?
- Apakah yang kita ukur benar-benar sesuai dengan apa yang kita yakini sedang diukur?
- Apakah asumsi dasar model telah terpenuhi?

Tidak ada metode statistik yang berdiri sendiri tanpa konteks. Setiap teknik bersandar pada asumsi-asumsi tertentu. Mengenali dan mengevaluasi asumsi-asumsi tersebut merupakan bagian dari pemeriksaan yang bertanggung jawab. Pemeriksaan secanggih apa pun tidak akan mampu menyelamatkan data yang cacat atau metodologi yang keliru.

3. Sikap Skeptis yang Sehat terhadap Pola

Manusia secara alami cenderung mencari pola, bahkan sering kali melihat makna dalam sesuatu yang sebenarnya acak. Seorang pemikir statistik akan bertanya:

- Apakah tren ini nyata atau sekadar variasi acak?
- Apakah pola ini stabil di berbagai sampel yang berbeda?
- Apakah korelasi menunjukkan hubungan sebab-akibat?
- Atau, mungkinkah kedua variabel tersebut dipengaruhi oleh faktor lain?

Metode statistik membantu memisahkan sinyal yang bermakna dari sekadar derau (*noise*).

4. Berpikir dalam Distribusi, Bukan Sekadar Rata-rata

Rata-rata saja jarang menceritakan kisah lengkapnya. Jika rata-rata pelanggan menghabiskan Rp 500.000, seperti apa distribusinya?

- Apakah sebagian besar pelanggan berkumpul di dekat Rp 500.000?
- Atau apakah rata-rata menyembunyikan gabungan beberapa pembelanja besar dan banyak pembelanja kecil?
- Apakah variabilitas meningkat seiring waktu?

Memahami distribusi (penyebaran, kecondongan, konsentrasi) mengungkap wawasan yang disembunyikan oleh rata-rata.

5. Mengajukan Pertanyaan yang Lebih Baik

Alat statistik memperluas kualitas pertanyaan yang dapat diajukan oleh para ilmuwan data. Daripada berhenti pada "Berapa rata-ratanya?", mereka bertanya:

- Seberapa tepat perkiraan tersebut?
- Berapa besar variasinya?
- Apakah perbedaannya antar kelompok?
- Apakah hal tersebut berubah secara signifikan seiring berjalannya waktu?
- Mekanisme apa yang mungkin menjelaskan pola tersebut?

Pertanyaan yang lebih baik mengarah pada wawasan yang lebih dalam. Setiap alat statistik memperluas jangkauan pertanyaan yang dapat mereka ajukan dan jawab.

Pola pikir ini bukanlah bawaan; itu dikembangkan melalui latihan. Seiring kemajuan Anda dalam membaca buku ini, Anda akan mendapati diri Anda secara alami mengadopsi kebiasaan berpikir ini. Anda akan mempertanyakan klaim dengan lebih hati-hati. Anda akan menafsirkan grafik dengan lebih kritis. Anda akan merasa nyaman dengan bahasa probabilistik. Ketika pergeseran itu terjadi, Anda tidak lagi sekadar memeriksa data. Anda memikirkannya. Saat itulah Anda akan mengetahui bahwa Anda menjadi ilmuwan data sejati.

1.5 PENERAPAN STATISTIKA DALAM PRAKTIK SAINS DATA

Untuk memahami sepenuhnya peran statistik, mari kita lihat tanggung jawab sehari-hari seorang data scientist. Pada setiap tahap, penalaran statistik bukanlah suatu pilihan; itu operasional.

Menjelajahi Data

Ketika dihadapkan dengan jutaan pengembangan teknologi masa depan, inspeksi mentah tidak mungkin dilakukan. Statistik memberikan ringkasan terstruktur yang membuat kompleksitas dapat diinterpretasikan. Ukuran seperti mean, median, varians, dan persentil memadatkan kumpulan data yang besar menjadi indikator yang bermakna. Visualisasi histogram, plot kotak, plot sebar adalah ekspresi grafis dari ide statistik. Mereka mengungkapkan bentuk distribusi, kemiringan, cluster, dan anomali. Tanpa alat statistik, eksplorasi hanyalah dugaan. Dengan mereka, pola menjadi terlihat.

Membandingkan Grup

- Apakah Obat A lebih efektif dibandingkan Obat B?
- Apakah Kampanye Pemasaran 1 mengungguli Kampanye 2?

- Apakah desain ulang situs web meningkatkan konversi?

Ini adalah pertanyaan perbandingan dalam ketidakpastian.

Pengujian hipotesis statistik dan interval kepercayaan memungkinkan kita menentukan apakah perbedaan yang diamati kemungkinan besar asli atau hanya akibat variasi acak. Pengujian A/B, yang banyak digunakan di industri, pada dasarnya merupakan penerapan statistik inferensial dalam eksperimen terkontrol. Perbandingan tanpa ketelitian statistik dapat menyesatkan. Perbandingan dengan struktur statistik memberikan kesimpulan yang dapat dipertahankan.

Membuat Prediksi

Prediksi merupakan inti dari ilmu data modern. Apakah memperkirakan:

- Kemungkinan churn pelanggan
- Kemungkinan suatu transaksi menjadi penipuan
- Harga yang diharapkan dari sebuah properti

Kami mengandalkan model probabilistik. Bahkan sistem pembelajaran mesin paling canggih pun didasarkan pada prinsip statistic memperkirakan parameter, meminimalkan kesalahan, dan melakukan generalisasi di luar data yang diamati. Prediksi bukanlah sesuatu yang ajaib. Ini adalah inferensi terstruktur.

Mengevaluasi Model

Model prediktif hanya bernilai jika kinerjanya ditunjukkan. Seberapa akurat prediksi model Anda dibandingkan dengan kenyataan? Metrik statistik seperti:

- Kesalahan *Root Mean Squared* (RMSE)
- Rata-rata Kesalahan Absolut (MAE)
- Akurasi, presisi, dan recall

mengukur seberapa dekat prediksi dengan kenyataan. Interval kepercayaan dan teknik validasi menilai seberapa stabil hasil tersebut di berbagai sampel. Evaluasi mencegah kepercayaan palsu. Ini membedakan model yang kuat dari ilusi yang berlebihan.

Mengomunikasikan Wawasan

Mungkin kekuatan statistik yang paling sering diremehkan terletak pada aspek komunikasi. Anda adalah penerjemah resmi bagi data tersebut. Tugas Anda adalah berdiri dalam sebuah rapat dan menyampaikan cerita yang meyakinkan serta berbasis bukti. Statistik memberikan kredibilitas pada cerita Anda. Statistik mengubah pernyataan:

> *"Strategi ini akan meningkatkan pendapatan."*

menjadi:

> *"Pemeriksaan kami memperkirakan peningkatan pendapatan sebesar 12–18%, dengan bukti statistik yang menunjukkan bahwa dampak tersebut kemungkinan besar bukan sekadar variasi acak."*

Statistik menyeimbangkan antusiasme dengan bukti. Statistik mengubah opini menjadi wawasan yang dapat dipertanggungjawabkan. Dalam praktiknya, sains data (*data science*) adalah statistik yang diterapkan melalui komputasi berskala besar.

1.6 STATISTIK DALAM KEHIDUPAN SEHARI-HARI

Penalaran statistik tidak terbatas pada ruang kelas atau laboratorium studi. Anda sebenarnya sudah menggunakan penalaran statistik setiap hari, meskipun mungkin tidak melabelinya sebagai demikian.

- Saat Anda melihat sebuah produk memiliki "4,5 bintang dari 13.000 ulasan," Anda secara implisit memahami bahwa sampel yang lebih besar memberikan bukti yang lebih tepercaya dibandingkan "5 bintang dari 2 ulasan." Itu adalah penalaran mengenai ukuran sampel dan representativitas.
- Saat prakiraan cuaca memprediksi "peluang hujan 70%," Anda sedang menafsirkan probabilitas untuk menentukan tindakan apakah harus membawa payung atau menjadwalkan ulang suatu acara.
- Saat seorang pemain bola basket memiliki tingkat keberhasilan lemparan bebas (*free-throw*) sebesar 82%, persentase tersebut merangkum performa masa lalu dan menjadi dasar ekspektasi untuk lemparan berikutnya. Ini adalah data deskriptif yang memandu inferensi probabilistik.
- Platform streaming yang merekomendasikan film berdasarkan riwayat tontonan Anda mengandalkan model statistik berskala besar yang membandingkan pola di antara jutaan pengguna.

Anda tidak memulai dari nol.

Anda sudah terbiasa bernalar mengenai rata-rata, kemungkinan, dan bukti. Buku ini hadir untuk memberikan bahasa formal serta kemampuan luar biasa yang menyertainya; mengubah intuisi menjadi pemahaman yang terstruktur.

1.7 MEMAHAMI STATISTIKA SEBAGAI KETERAMPILAN PRAKTIS

Bagi banyak pembelajar, kata "statistik" sering kali memicu kecemasan; muncul ingatan tentang rumus-rumus yang rumit, perhitungan mekanis, dan simbol-simbol abstrak yang terasa tidak bermakna. Persepsi tersebut mencerminkan model pengajaran lama. Dalam sains data modern, perangkat komputasi menangani perhitungan yang berat. Perangkat lunak seperti Excel, Python, dan R dapat menjalankan prosedur statistik yang kompleks dalam hitungan milidetik. Melakukan perhitungan secara manual bukan lagi keterampilan utama yang diperlukan. Hal yang tetap menjadi ciri khas manusia dan oleh karena itu memiliki nilai yang unik adalah interpretasi. Nilai Anda sebagai seorang ilmuwan data (*data scientist*) terletak pada kemampuan Anda untuk:

- Merumuskan pertanyaan yang bermakna
- Memilih perangkat pemeriksaan yang tepat
- Menafsirkan hasil secara akurat
- Mengomunikasikan temuan secara bertanggung jawab

Dengan demikian, statistik bukan sekadar soal perhitungan. Statistik adalah soal perspektif.

- Statistik berarti mengenali ketidakpastian di saat orang lain melihat kepastian.
- Statistik berarti mempertanyakan pola sebelum menerimanya.
- Statistik berarti mengevaluasi bukti sebelum menarik kesimpulan.

Buku ini dirancang untuk menumbuhkan perspektif tersebut. Kami akan mengutamakan intuisi di atas formalisme, interpretasi di atas komputasi, dan penalaran di atas otomatisasi. Tujuannya bukan untuk mengubah Anda menjadi kalkulator, melainkan melatih Anda agar dapat melihat data dengan jernih.

1.8 STATISTIKA DALAM PENGAMBILAN KEPUTUSAN

Kursus-kursus tradisional sering kali menyajikan statistik sebagai persyaratan teknis berupa rumus untuk dihafal, prosedur untuk diikuti, dan ujian untuk dilalui. Padahal, pada kenyataannya, statistik merupakan salah satu perangkat intelektual paling berdaya yang bisa Anda kembangkan.

- **Statistik mempertajam persepsi**: Setiap hari, Anda menjumpai berbagai klaim yang didukung oleh "data" baik itu dalam iklan, laporan kebijakan, tajuk berita media, maupun ringkasan studi. Tanpa literasi statistik, respons kita sering kali bersifat instingtif: langsung menerima atau menolaknya begitu saja. Dengan pemahaman statistik, Anda dapat mengevaluasi bukti secara cermat. Anda mampu mengenali ukuran sampel yang terlalu kecil, grafik yang dimanipulasi, pelaporan yang selektif, serta klaim yang dilebih-lebihkan. Anda tidak hanya melihat apa yang disajikan di permukaan, tetapi mampu memahami struktur di baliknya.
- **Statistik memperkuat pengambilan keputusan**: Baik saat mengevaluasi rekomendasi medis, membandingkan pilihan investasi, menilai kinerja bisnis, maupun merencanakan inisiatif baru, penalaran statistik menghadirkan kedisiplinan. Statistik membantu Anda menimbang berbagai konsekuensi, menilai ketidakpastian, dan bertindak berdasarkan bukti, bukan sekadar dorongan sesaat.
- **Statistik mengubah rasa ingin tahu menjadi penemuan**: Pola-pola yang luput dari perhatian orang lain menjadi terlihat. Variabilitas menjadi sumber informasi, bukan lagi sesuatu yang membingungkan. Tren mendorong penyelidikan, bukan sekadar spekulasi. Baik dalam dunia bisnis, kesehatan, studi, maupun proyek pribadi, statistik memungkinkan Anda beralih dari sekadar pengamatan menuju penjelasan.
- **Statistik menciptakan nilai yang terukur**: Berbagai organisasi mengumpulkan data dalam jumlah besar, namun sering kali kesulitan menafsirkannya secara bertanggung jawab. Individu yang mampu menerjemahkan pola numerik menjadi wawasan yang koheren dapat menghadirkan kejelasan di tengah kompleksitas yang dilihat orang lain. Kemampuan ini bukan sekadar teknis melainkan strategis.
- **Statistik menghubungkan Anda dengan penalaran ilmiah**: Statistik menjadi landasan bagi studi modern di berbagai bidang, mulai dari kedokteran, ekonomi, ilmu sosial, kesehatan masyarakat, hingga teknologi. Memahami statistik memungkinkan Anda

untuk mengevaluasi studi secara kritis, menghindari informasi yang keliru, serta berpartisipasi secara bermakna dalam menghasilkan pengetahuan baru.

- **Di luar manfaat profesional, statistik mengubah cara pandang**: Anda mulai berpikir dalam kerangka probabilitas, bukan kepastian mutlak. Anda menjadi lebih nyaman dengan ketidakpastian, alih-alih merasa terancam olehnya. Anda pun mampu bersikap skeptis secara proporsional tanpa harus serta-merta menolak gagasan orang lain. Bayangkan Anda sedang dalam sebuah rapat di mana sekumpulan data disajikan untuk menjustifikasi keputusan besar. Orang lain mungkin menerima kesimpulan tersebut begitu saja. Namun, Anda menyadari adanya sampel yang kecil, variabel yang tidak terkendali, atau rentang kepercayaan (*confidence interval*) yang lebar. Dengan tenang, Anda mengajukan pertanyaan klarifikasi yang menyempurnakan pemeriksaan tersebut. Momen penalaran yang disiplin seperti ini dapat mencegah kesalahan yang merugikan. Statistik bukanlah penghalang sebelum seseorang bisa terjun ke dalam sains data yang "sesungguhnya". Statistik juga bukan sekadar "prasyarat membosankan" sebelum sampai ke bagian yang "menyenangkan". Statistik adalah mesin penggerak di balik sains data. Begitu Anda memahami tujuannya, bidang ini tidak lagi terasa mengintimidasi, melainkan menjadi sesuatu yang sangat diperlukan.

1.9 CAPAIAN PEMBELAJARAN DAN PENGUASAAN KOMPETENSI STATISTIK SAINS DATA

Buku ini merupakan perjalanan yang disusun dengan cermat, dirancang untuk membawa Anda dari seorang pemula di bidang statistik menjadi praktisi yang percaya diri. Dengan mempelajari isinya secara saksama, Anda akan mengembangkan:

- **Landasan Konseptual yang Kuat**: Sebelum mendalami berbagai teknik, Anda akan membangun intuisi yang mendalam mengenai konsep-konsep statistik inti apa maknanya, mengapa hal tersebut penting, dan bagaimana keterkaitannya satu sama lain.
- **Penguasaan Statistik Deskriptif**: Anda akan belajar untuk meringkas dan memvisualisasikan data secara efektif, serta memilih ukuran dan grafik yang tepat untuk setiap situasi.
- **Kemampuan Mendeteksi dan Mengukur Hubungan**: Anda akan mengeksplorasi cara mengidentifikasi dan mengukur hubungan antarvariabel menggunakan korelasi, regresi, dan teknik lainnya, sembari tetap berhati-hati dalam menafsirkan hubungan sebab-akibat.
- **Pemahaman tentang Probabilitas**: Anda akan menguasai konsep-konsep probabilitas yang menjadi dasar bagi pemodelan prediktif dan pembelajaran mesin (*machine learning*), termasuk probabilitas bersyarat, distribusi probabilitas, dan penalaran Bayesian.
- **Keterampilan Inferensial**: Anda akan menguasai pengujian hipotesis dan selang kepercayaan (*confidence intervals*) perangkat penting untuk menarik kesimpulan di luar sampel yang diamati, sekaligus mengukur tingkat ketidakpastian.

- **Penerapan Praktis dalam Excel**: Sepanjang buku ini, Anda akan mempelajari cara menerapkan konsep-konsep tersebut menggunakan Excel perangkat yang mudah diakses, canggih, dan digunakan secara luas dalam pemeriksaan data di dunia nyata.
- **Penilaian Analitis**: Lebih dari sekadar rumus, Anda akan mengembangkan kemampuan untuk:
  - Menentukan metode yang tepat dalam konteks tertentu
  - Mengenali pelanggaran asumsi
  - Menangani data yang tidak sempurna atau berantakan secara bertanggung jawab
  - Mengomunikasikan temuan dengan jelas kepada audiens teknis maupun non-teknis
- **Pola Pikir Statistik**: Yang terpenting, Anda akan menumbuhkan cara berpikir yang disiplin. Anda akan terbiasa mempertanyakan kualitas data, menantang kesimpulan yang tidak didukung bukti, mengenali berbagai keterbatasan, serta menyajikan wawasan dengan nuansa yang tepat.

Di akhir buku ini, Anda tidak hanya sekadar "mengetahui statistik", tetapi Anda akan mampu berpikir secara statistik. Anda akan siap untuk melangkah dengan percaya diri ke lingkungan berbasis pemrograman seperti Python, R, atau SQL, dengan berbekal pemahaman mendasar yang diperlukan untuk menafsirkan hasil keluaran, bukan sekadar menghasilkannya. Apa yang dulunya tampak abstrak akan menjadi terstruktur. Apa yang dulunya terasa mengintimidasi akan menjadi lebih mudah untuk dipahami dan dikelola. Statistik akan menjadi perangkat yang Anda gunakan secara sadar dan percaya diri dalam perjalanan Anda di bidang sains data.

1.10 IKHTISAR DASAR STATISTIKA DALAM SAINS DATA

Pada bab pembuka ini, kita telah meninjau ulang statistik bukan sebagai sekumpulan rumus, melainkan sebagai cara berpikir yang disiplin. Kita mulai dengan menyadari bahwa data tidak secara otomatis menyampaikan makna. Seperti bahasa, data harus ditafsirkan. Statistik menyediakan struktur yang memungkinkan angka-angka mentah berubah menjadi bukti, pola, dan wawasan. Kita mendefinisikan statistik sebagai ilmu untuk belajar dari data di tengah ketidakpastian. Dua cabang utama membentuk disiplin ilmu ini:

- **Statistik deskriptif**, yang meringkas dan mengatur data agar pola-pola di dalamnya dapat terlihat.
- **Statistik inferensial,** yang memungkinkan kita menarik kesimpulan di luar sampel yang diamati, sembari mengukur tingkat ketidakpastian secara eksplisit.

Kita telah menelaah bagaimana statistik mendukung setiap tahapan dalam siklus hidup sains data; mulai dari pembersihan dan eksplorasi data hingga pemodelan, evaluasi, dan komunikasi. Sains data tidak terpisah dari statistik; sains data adalah statistik yang diterapkan melalui komputasi dalam skala besar. Kita juga mengeksplorasi pola pikir seorang pemikir statistik:

- Kenyamanan dalam menghadapi ketidakpastian
- Kewaspadaan terhadap asumsi-asumsi yang digunakan
- Sikap skeptis yang sehat terhadap pola-pola yang tampak

- Perhatian pada distribusi, bukan sekadar nilai rata-rata tunggal
- Komitmen untuk mengajukan pertanyaan yang lebih baik

Kita pun menyadari bahwa penalaran statistik sebenarnya sudah menjadi bagian dari kehidupan sehari-hari. Mulai dari menafsirkan probabilitas dalam prakiraan cuaca hingga mengevaluasi peringkat dan rekomendasi, kita secara rutin berinteraksi dengan gagasan-gagasan statistik meskipun sering kali secara informal. Terakhir, kita meluruskan kesalahpahaman umum: statistik bukan sekadar tentang melakukan perhitungan yang rumit. Perangkat modern kini menangani komputasi secara efisien. Hal yang membedakan seorang ilmuwan data yang terampil adalah kemampuannya dalam menafsirkan hasil, mengevaluasi bukti, serta mengomunikasikan kesimpulan secara bertanggung jawab. Statistik bukanlah hambatan dalam perjalanan menuju sains data, melainkan fondasinya. Saat melangkah lebih jauh, Anda akan mulai membangun fondasi tersebut dengan cermat, dimulai dari elemen-elemen dasar: data, variabel, dan pengukuran. Kini, perjalanan berlanjut dari sekadar memahami semangat statistik menuju penguasaan atas strukturnya.

BAB 2
DATA SEBAGAI DASAR ANALISIS STATISTIK

2.1 PENGANTAR DATA DALAM STATISTIKA

Dalam era digital dan bisnis modern yang bergerak serba cepat, pengambilan keputusan strategis tidak lagi dapat sekadar mengandalkan intuisi, asumsi, atau pengalaman masa lalu semata. Setiap pemeriksaan metrik kinerja, setiap grafik yang menggambarkan tren pasar, setiap model prediktif yang meramalkan perilaku konsumen, hingga setiap keputusan bisnis yang berdampak besar, semuanya bermula dari satu bahan dasar yang sama dan paling krusial: data. Data adalah denyut nadi dari analitik bisnis; ia adalah fondasi di mana seluruh bangunan ilmu statistika didirikan.

Namun, sebelum Anda dapat melangkah lebih jauh seperti meringkas jutaan angka menjadi satu laporan *dashboard*, memvisualisasikan pola yang tersembunyi, atau membangun model pembelajaran mesin (*machine learning*) ada prasyarat mendasar yang mutlak harus dipenuhi. Anda harus terlebih dahulu memahami anatomi dari data itu sendiri. Anda perlu mengetahui secara persis apa itu data, bagaimana strukturnya terbentuk, dari mana sumber asalnya, serta parameter apa yang membuat suatu kumpulan data dapat dipercaya (*reliable*) atau justru berpotensi menyesatkan.

Banyak analis atau profesional yang sering kali terburu-buru melompat ke proses analisis tanpa benar-benar "berkenalan" dengan data mereka. Hal ini sering bermuara pada prinsip fundamental dalam ilmu data: *Garbage In, Garbage Out* (GIGO). Sehebat apa pun algoritma atau perangkat lunak yang Anda gunakan, jika data yang dimasukkan memiliki kualitas yang buruk, cacat, atau bias, maka kesimpulan yang dihasilkan akan sama buruknya. Bab ini dirancang khusus untuk mencegah hal tersebut dan meletakkan dasar pemahaman yang kokoh.

Prinsip-prinsip yang akan diperkenalkan dalam bab ini bersifat universal dan sangat fleksibel (skalabel). Baik Anda sedang mengolah puluhan baris tanggapan survei kertas dari pelanggan sebuah kedai kopi lokal, maupun sedang menambang jutaan baris data (*Big Data*) dari basis data transaksional sebuah perusahaan *e-commerce* multinasional, konsep-konsep dasar tentang cara memperlakukan data ini tetap tidak berubah.

Oleh karena itu, anggaplah bab ini sebagai panduan praktis dan kompas utama Anda mengenai literasi data. Ini adalah referensi pijakan yang akan terus Anda ingat dan gunakan sebagai landasan, bahkan jauh setelah Anda beralih ke metode pengujian hipotesis, analisis regresi, atau topik-topik statistika yang jauh lebih tingkat lanjut. Memiliki insting data yang baik dimulai dari sini.

Sebagai peta jalan pembelajaran, di akhir bab ini, Anda akan dibekali dengan kemampuan untuk:

1. **Mengklasifikasikan Data:** Membedakan anatomi data berdasarkan jenis (seperti kualitatif vs. kuantitatif) dan tingkat pengukurannya (nominal, ordinal, interval, dan rasio).

2. **Mengevaluasi Kualitas Kumpulan Data:** Mengidentifikasi data yang kotor, mendeteksi anomali (*outlier*), serta menilai apakah suatu data cukup valid untuk mewakili kondisi di dunia nyata.
3. **Memahami Asal-Usul dan Konteks Data:** Menelusuri dari mana data berasal apakah itu data primer yang dikumpulkan secara mandiri atau data sekunder dari pihak ketiga dan memahami implikasi bias di dalamnya.
4. **Mengajukan Pertanyaan Kritis:** Memiliki daftar periksa (*checklist*) mental mengenai pertanyaan-pertanyaan investigatif apa saja yang wajib diajukan sesaat setelah Anda membuka kumpulan data baru di *spreadsheet* (seperti Excel), sebelum satu pun rumus atau fungsi analisis dijalankan.

2.2 PENGERTIAN DATA DAN INFORMASI

Dalam percakapan sehari-hari di era digital ini, kata "data" sering kali digunakan secara longgar dan tumpang tindih. Banyak orang menggunakannya secara bergantian untuk merujuk pada fakta, informasi, wawasan (*insight*), atau bahkan sekadar angka statistik yang muncul di berita. Namun, dalam konteks disiplin ilmu statistika, sains data, dan analitik bisnis, istilah ini memiliki definisi dan batasan yang jauh lebih presisi dan terstruktur.

Pada level paling dasar, data adalah sekumpulan nilai, angka, simbol, atau fakta mentah mengenai suatu keadaan di dunia nyata yang telah direkam dan dikumpulkan untuk tujuan tertentu. Anda bisa membayangkan data sebagai bahan baku atau "minyak mentah" dalam dunia analitik. Sama seperti minyak mentah yang baru dipompa dari perut bumi, secara terpisah dan tanpa pengolahan, nilai-nilai data individual hampir tidak memiliki makna atau kegunaan praktis.

Untuk membuat data menjadi berharga, ia harus melalui proses transformasi: pengorganisasian, pemberian struktur, penyaringan, dan penambahan konteks. Melalui proses inilah hirarki terbentuk, yang sering dikenal dengan sebutan Piramida DIKW (*Data, Information, Knowledge, Wisdom*). Data mentah perlahan berubah menjadi informasi yang bermakna, dan pada akhirnya, berevolusi menjadi pengetahuan yang dapat mendorong pengambilan keputusan.

Studi Kasus: Evolusi Data Menjadi Pengetahuan

Untuk memahami perbedaan konseptual ini, mari kita bedah transformasi tersebut melalui sebuah contoh sederhana namun kuat di dunia medis:

- **Data (Fakta Mentah tanpa Konteks):** Bayangkan Anda melihat angka "98,6" tertulis di selembar kertas. Tanpa satuan, tanpa keterangan waktu, dan tanpa subjek, angka ini hanyalah sebuah goresan tinta. Anda tidak tahu apakah ini merupakan nominal harga, persentase nilai ujian, atau jarak dalam kilometer. Angka ini buta huruf secara makna.
- **Informasi (Data + Konteks):** Kini, kita menambahkan atribut 5W1H (*Who, What, Where, When, Why*) ke dalam angka tersebut. Anda mengetahui bahwa 98,6 tersebut mewakili suhu tubuh seorang pasien dewasa dalam derajat Fahrenheit, yang diukur oleh seorang perawat di ruang gawat darurat pada pukul 09.00 pagi di hari Senin.

Dengan adanya struktur dan konteks ini, angka tersebut kini memiliki arti. Data telah berubah menjadi informasi.

- **Pengetahuan (Informasi + Kerangka Pemahaman):** Langkah selanjutnya adalah menerapkan pemahaman ilmiah, pengalaman, atau aturan bisnis terhadap informasi tersebut. Seorang dokter membaca informasi "suhu 98,6°F". Berdasarkan ilmu medis yang ia miliki, ia tahu bahwa suhu tubuh normal manusia berada di kisaran 97°F hingga 99°F. Mengetahui bahwa angka tersebut berada di tepat di tengah rentang normal sehingga pasien tidak mengalami demam dan tidak ada intervensi medis darurat yang diperlukan saat itu adalah sebuah **pengetahuan**.

Cakupan Data dan Memisahkan Sinyal dari Kebisingan (*Signal vs. Noise*)

Dalam lanskap bisnis modern, data dapat menggambarkan hampir semua hal yang terukur (kuantitatif) atau teramati (kualitatif). Cakupannya sangat luas dan bervariasi:

- **Karakteristik Demografis:** Usia, tingkat pendapatan, atau lokasi geografis seseorang.
- **Metrik Operasional & Finansial:** Fluktuasi harga suatu produk setiap jamnya, total biaya produksi, atau jumlah unit yang terjual di musim liburan.
- **Jejak Digital (Digital Exhaust):** Jumlah klik, durasi kunjungan situs web per hari, atau rasio *bounce rate* pada sebuah kampanye pemasaran *email*.
- **Persepsi Psikologis:** Tingkat kepuasan pelanggan, sentimen merek di media sosial, atau ulasan bintang lima di *platform e-commerce*.

Namun, di era di mana segala hal bisa direkam, kita dihadapkan pada ancaman baru: ledakan kebisingan (*noise*). Apa yang membedakan data yang berkualitas dari sekadar "kebisingan acak"? Jawabannya terletak pada niat (*intention*) dan konsistensi.

Data yang sesungguhnya adalah hasil dari perekaman yang dilakukan secara sengaja, menggunakan metode ukur yang terstandarisasi, dan dicatat secara konsisten dari waktu ke waktu. Jika Anda mencatat pendapatan harian toko Anda kadang di pagi hari, kadang di malam hari, dan kadang menggabungkannya dengan hari lain, Anda hanya sedang mengumpulkan angka, bukan data yang valid. Konsistensi inilah yang memastikan bahwa data dapat diandalkan untuk dianalisis, diuji, dan dibandingkan secara *apple-to-apple*, sehingga Anda bisa melihat tren dan pola yang sebenarnya.

2.3 STRUKTUR DAN KOMPONEN DATA

Sebelum menjelajahi berbagai jenis dan sumber data, kita harus memahami struktur internalnya. Terlepas dari apakah Anda bekerja di Excel, basis data relasional, atau paket perangkat lunak statistik, kumpulan data memiliki komponen dasar yang sama. Memahami komponen-komponen ini memungkinkan Anda untuk menafsirkan, memanipulasi, dan memeriksa data dengan benar.

a) Titik Data

Titik data adalah nilai tunggal yang tercatat untuk satu variabel dalam satu pengamatan. Misalnya, nilai 34 yang mewakili usia pelanggan adalah satu titik data. Ini adalah unit terkecil yang bermakna dalam kumpulan data; satu pengukuran, satu fakta yang tercatat, satu nilai yang tersimpan. Dengan sendirinya, titik data hanya memberikan wawasan terbatas.

Maknanya hanya muncul dalam kaitannya dengan nilai-nilai lain dalam kumpulan data. Statistik tidak memeriksa titik data yang terisolasi; statistik memeriksa pola di banyak titik data tersebut.

b) Variabel (Fitur)

Variabel (sering disebut fitur dalam pembelajaran mesin) adalah karakteristik atau atribut terukur yang dikumpulkan untuk setiap pengamatan. Dalam spreadsheet, variabel biasanya muncul sebagai judul kolom: ID Pelanggan, Nama, Usia, Lokasi, Pembelian Seumur Hidup, Kepuasan, dll. Variabel mewakili pertanyaan yang dirancang untuk dijawab oleh dataset Anda. Setiap variabel dapat memiliki nilai yang berbeda di berbagai pengamatan, oleh karena itu disebut variabel. Tanpa variasi, pemeriksaan statistik tidak akan mungkin dilakukan. Dalam konteks analitis, variabel sering diklasifikasikan sebagai:

- **Variabel independen (prediktor atau input)**: digunakan untuk menjelaskan atau memprediksi hasil.
- **Variabel dependen (respons atau output)**: hasil yang diperiksa atau diprediksi.

Perbedaan ini menjadi sangat penting dalam pemeriksaan regresi dan pemodelan prediktif.

c) Pengembangan Teknologi Masa Depan (Catatan)

Sebuah pengembangan teknologi masa depan (juga disebut sebagai catatan, baris, atau contoh) mewakili satu unit pemeriksaan tunggal di mana data telah dikumpulkan. Unit pemeriksaan bergantung pada konteks studi:

- Dalam dataset pelanggan, setiap pengembangan teknologi masa depanadalah satu pelanggan.
- Dalam uji klinis, setiap pengembangan teknologi masa depan adalah satu pasien.
- Dalam dataset deret waktu, setiap pengembangan teknologi masa depandapat mewakili satu hari, minggu, atau bulan.

pengembangan teknologi masa depanmenjawab pertanyaan: Apa yang diukur?
Pertimbangkan contoh di bawah ini:

Tabel 2.1: Sebuah pengembangan teknologi masa depan

ID Pelanggan	Nama	Usia	Lokasi	Total Pembelian	Menjadi Anggota Sejak	Kepuasan
10472	Sunarko	34	Smg	Rp 28.475.000	2019-03-15	Tinggi

Setiap sel dalam baris tersebut adalah titik data. Setiap kolom mewakili sebuah variabel. Sebuah dataset dapat berisi ratusan, ribuan, atau jutaan pengamatan tersebut.

d) Dataset: Kumpulan Pengamatan

Dataset adalah kumpulan pengamatan dan variabel yang lengkap dan terorganisir. Ini adalah tabel lengkap baris dan kolom bersama-sama yang disimpan dalam spreadsheet, basis data, atau file data.

Tabel 2.2 Dataset Pelanggan

ID Pelanggan	Nama	Usia	Lokasi	Total Pembelian	Menjadi Anggota Sejak	Kepuasan
10472	Sunarko	34	Smg	Rp 28.475.000	2019-03-15	Tinggi
10473	Rendi	28	Sby	Rp 12.342.000	2020-07-22	Sedang
10474	Heni	45	Jkt	Rp 59.120.000	2018-01-08	Tinggi

Kumpulan data yang terstruktur dengan baik memiliki tiga karakteristik penting:

- Setiap baris mewakili satu pengembangan teknologi masa depan.
- Setiap kolom mewakili satu variabel.
- Setiap kolom berisi nilai-nilai dengan tipe data yang konsisten.

Susunan persegi panjang ini sering disebut sebagai data rapi, istilah yang dipopulerkan dalam ilmu data modern. Sebagian besar alat statistik termasuk Excel, R, Python, dan SQL mengharapkan data dalam format ini. Ketika data terstruktur dengan benar, pemeriksaan menjadi mudah. Ketika struktur rusak misalnya, ketika beberapa variabel digabungkan menjadi satu kolom kebingungan dan kesalahan analitis akan terjadi.

2.4 JENIS DATA: MEMAHAMI VARIASINYA

Tidak semua data sama. Jenis data yang Anda gunakan menentukan operasi statistik mana yang valid, grafik mana yang sesuai, dan teknik analitis mana yang berlaku. Ada dua kategori fundamental, masing-masing dengan sub-tipe penting.

1. Data Kualitatif vs Kuantitatif

Data Kualitatif (Data Kategorikal)

Data kualitatif, juga dikenal sebagai data kategorikal, menggambarkan atribut atau kategori daripada besaran numerik. Data kualitatif menjawab pertanyaan: Apa yang Anda cari? Contohnya meliputi:

- Kategori produk: elektronik, pakaian, furnitur, bahan makanan
- *Tingkat kepuasan pelanggan*: Tidak Puas, Netral, Puas, Sangat Puas

Meskipun data kualitatif terkadang dikodekan secara numerik (misalnya, 1 = Laki-laki, 2 = Perempuan), angka-angka tersebut tidak mewakili kuantitas yang terukur. Operasi aritmatika seperti penjumlahan atau rata-rata tidak bermakna secara inheren. Data kualitatif dibagi menjadi dua subtipe:

I. Data Nominal

Data nominal terdiri dari kategori tanpa urutan intrinsik. Contoh:

- Golongan darah (A, B, AB, O)
- Nama negara (Albania, Aljazair, Andorra, Angola)
- Warna mata (coklat, biru, hijau)

Tidak ada peringkat di antara kategori. Satu kategori tidak "lebih besar" atau "lebih kecil" dari kategori lainnya. Dengan data nominal, kita dapat:

- Menghitung frekuensi
- Menghitung proporsi

- Mengidentifikasi modus

Namun kita tidak dapat:

- Menghitung rata-rata yang bermakna
- Melakukan perbandingan aritmatika

II. Data Ordinal

Data ordinal terdiri dari kategori-kategori yang mengikuti urutan yang bermakna, tetapi interval antar kategori tidak selalu sama. Contoh:

Tingkat pendidikan (SMA < Sarjana < Magister < PhD)

- Peringkat kepuasan (Rendah, Sedang, Tinggi)
- Respons skala Likert (Sangat Tidak Setuju → Sangat Setuju)

Data ordinal memungkinkan pemeringkatan. Namun, perbedaan antara kategori yang berdekatan tidak dijamin seragam. Misalnya, kesenjangan antara "Setuju" dan "Sangat Setuju" mungkin tidak mewakili perbedaan intensitas yang sama seperti antara "Netral" dan "Setuju." Dalam praktiknya, peneliti terkadang menghitung rata-rata untuk skala ordinal (terutama skala Likert), tetapi secara tegas, interpretasi harus dilakukan dengan hati-hati.

Data Kuantitatif (Data Numerik)

Data kuantitatif mewakili kuantitas yang dapat diukur dan mendukung operasi aritmatika. Data ini menjawab pertanyaan: Berapa usia? atau Berapa total pembelian? Contohnya meliputi:

- Jumlah pelanggan
- Pendapatan tahunan
- Tinggi badan dalam sentimeter
- Waktu dalam menit

Data kuantitatif dibagi menjadi tipe diskrit dan kontinu.

I. Data Diskrit

Data diskrit terdiri dari nilai-nilai yang dapat dihitung, biasanya bilangan bulat. Contoh:

- Jumlah produk dalam keranjang belanja
- Jumlah karyawan
- Jumlah cacat dalam satu batch
- Jumlah klik pada situs web

Variabel diskrit muncul dari proses penghitungan. Anda tidak mungkin memiliki 3,7 karyawan atau 2,5 keluhan.

II. Data Kontinu

Data kontinu dapat mengambil nilai apa pun dalam rentang tertentu, termasuk pecahan dan desimal. Contoh:

- Tinggi
- Berat
- Suhu
- Waktu
- Pendapatan

Secara teori, data kontinu dapat mengambil nilai yang tak terhingga banyaknya dalam suatu interval. Dalam praktiknya, hal itu hanya dibatasi oleh presisi pengukuran. Misalnya:

- 12 kg
- 12,1 kg
- 12,06 kg
- 12,055 kg

Semakin presisi instrumennya, semakin akurat pengukurannya.

Tabel 2.3 Hierarki Tipe Data

Tipe Data Utama	Sub-Tipe	Keterangan & Contoh
Kualitatif (Kategorik)	Nominal	Tanpa urutan alami m.s. Warna mata, Golongan darah, Nama kota
Kualitatif (Kategorik)	Ordinal	Urutan alami m.s. Penilaian (1-5), Tingkat pendidikan, Kepuasan
Kuantitatif (Numerik)	Diskret	Nilai yang dapat dihitung m.s. No. pesanan, No. mahasiswa, Hitungan klik
Kuantitatif (Numerik)	Kontinu	Rentang tak terbatas m.s. Tinggi, Berat, Pendapatan

Memahami perbedaan-perbedaan ini bukan sekadar masalah akademis. Klasifikasi yang tepat merupakan langkah awal menuju pemeriksaan yang valid. Setiap jenis data juga berkaitan dengan tingkat pengukuran tertentu sebuah konsep yang akan kita bahas selanjutnya.

2. Empat Tingkat Pengukuran

Empat tingkat pengukuran membentuk salah satu kerangka kerja paling mendasar dalam statistika. Tingkat-tingkat ini menentukan:

- Ringkasan statistik apa yang tepat untuk digunakan
- Operasi matematika apa yang bermakna
- Visualisasi apa yang sesuai
- Kesimpulan apa yang valid

Mengabaikan tingkat pengukuran akan menghasilkan interpretasi yang keliru. Memperhatikan tingkat pengukuran tersebut akan menjamin integritas pemeriksaan.

Tingkat 1: Data Nominal (Kategori)

Definisi: Data yang merepresentasikan kategori tanpa urutan atau peringkat yang melekat. Karakteristik utama: Kategori-kategori tersebut berbeda satu sama lain, namun tidak ada yang "lebih" atau "kurang" dibandingkan yang lain.

Contoh:

- Jenis kelamin: Laki-laki, Perempuan
- Golongan darah: A, B, AB, O
- Kategori produk: Elektronik, Pakaian, Makanan, Buku

Meskipun nilai-nilainya mungkin diberi kode angka (contoh: 1 = Laki-laki, 2 = Perempuan), angka-angka tersebut hanyalah label.

Pemeriksaan yang diperbolehkan:

- Penghitungan frekuensi (Berapa jumlah pelanggan dalam setiap kategori?)
- Persentase dan proporsi
- Modus (Kategori yang paling sering muncul)
- Uji asosiasi (contoh: uji chi-square)

Tidak bermakna:

- Rata-rata (bagaimana menghitung rata-rata golongan darah?)
- Peringkat (Apakah "Merah" lebih besar daripada "Biru"? Tidak bermakna)
- Operasi aritmetika (Biru + Hijau =? Tidak masuk akal)

Kesalahan umum: Kesalahan yang sering terjadi adalah memperlakukan nomor identifikasi sebagai variabel kuantitatif. Pelanggan nomor 10473 tidak berarti "lebih besar" daripada pelanggan nomor 10472. Angka-angka tersebut hanya berfungsi sebagai penanda identitas.

Tingkat 2: Data Ordinal (Kategori Berperingkat)

Definisi: Data yang merepresentasikan kategori-kategori yang memiliki urutan, di mana interval atau jarak antar kategori tidak dijamin sama. Karakteristik utama: Terdapat urutan, namun jarak antar nilai tidak terdefinisi.

Contoh:

- Tingkat kepuasan: Sangat Tidak Puas < Tidak Puas < Netral < Puas < Sangat Puas
- Skala nyeri: Tidak ada < Ringan < Sedang < Berat < Sangat berat
- Peringkat bintang: ★ < ★★ < ★★★ < ★★★★ < ★★★★★
- Ukuran kaus: XS < S < M < L < XL < XXL

Pemeriksaan yang Diperbolehkan:

- Semua pemeriksaan yang berlaku untuk data nominal
- Pemeringkatan (*ranking*)
- Median
- Persentil

Kurang Bermakna secara Ketat:

- Operasi aritmetika
- Asumsi adanya jarak yang sama antar kategori

Dalam praktiknya, respons ordinal (seperti skala Likert 1–5) sering kali dihitung nilai rata-ratanya. Meskipun lazim dilakukan, nilai rata-rata tersebut harus diinterpretasikan dengan hati-hati. Hasilnya lebih merupakan indeks ringkasan daripada pengukuran kuantitatif yang presisi. Sebagai contoh, skor kepuasan rata-rata sebesar 3,7 menunjukkan bahwa respons terkonsentrasi di antara kategori “Netral” dan “Puas”, namun hal ini tidak menyiratkan adanya

intensitas fraksional yang dapat diukur secara pasti. Memahami keterbatasan ini dapat mencegah interpretasi yang berlebihan.

Tingkat 3: Data Interval (Interval Konsisten, Tanpa Titik Nol Mutlak)

Definisi: Data dengan nilai-nilai yang terurut dan memiliki interval yang konsisten serta setara di antara nilai-nilai tersebut, namun tidak memiliki titik nol mutlak. Karakteristik Utama: Selisih antar nilai memiliki makna; rasio antar nilai tidak memiliki makna.

Contoh:

- Suhu dalam skala Celsius atau Fahrenheit
- Tanggal dalam kalender
- Skor IQ

Peningkatan sebesar 10 derajat menunjukkan perubahan yang sama di titik mana pun pada skala Celsius. Akan tetapi, suhu 40°C tidak berarti dua kali lebih panas daripada 20°C, karena angka nol di sini tidak menunjukkan ketiadaan suhu.

Pemeriksaan yang Dapat Dilakukan:

- Semua pemeriksaan yang berlaku untuk data ordinal
- Rata-rata (*mean*)
- Simpangan baku (standar deviasi)
- Penjumlahan dan pengurangan
- Korelasi dan regresi

Pemeriksaan yang Tidak Bermakna:

- Perbandingan rasio
- Menafsirkan angka nol sebagai ketiadaan (absennya) suatu besaran

Ketiadaan titik nol mutlak membuat perbandingan perkalian (rasio) menjadi tidak bermakna.

Tingkat 4: Data Rasio (Data yang Paling Lengkap)

Definisi: Data dengan nilai-nilai yang terurut, interval yang setara, serta memiliki titik nol mutlak yang menunjukkan ketiadaan total dari suatu kuantitas. Karakteristik Utama: Semua operasi aritmetika memiliki makna, termasuk perbandingan rasio.

Contoh:

- Tinggi badan: 0 cm berarti tidak ada tinggi, 180 cm secara harfiah adalah dua kali lipat dari 90 cm
- Berat badan: 0 kg berarti tidak ada berat, 80 kg secara harfiah adalah dua kali lipat dari 40 kg
- Pendapatan: 0 berarti tidak ada pendapatan, Rp 1 Miliar secara harfiah adalah dua kali lipat dari Rp 500 juta
- Durasi waktu: 0 detik berarti tidak ada waktu, 60 detik secara harfiah adalah dua kali lipat dari 30 detik
- Jumlah pembelian: 0 berarti tidak ada pembelian, 10 secara harfiah adalah dua kali lipat dari 5
- Jarak: 0 km berarti tidak ada jarak, 100 km secara harfiah adalah dua kali lipat dari 50 km

- Usia: 0 tahun berarti baru lahir, 40 tahun secara harfiah adalah dua kali lipat dari 20 tahun

Pemeriksaan yang Dapat Dilakukan:

- Semua statistik deskriptif
- Semua operasi aritmetika
- Rasio dan tingkat pertumbuhan
- Pemodelan statistik tingkat lanjut

Data rasio menawarkan fleksibilitas pemeriksaan dan kemampuan interpretasi yang paling maksimal. Sebagian besar pengukuran fisik termasuk dalam kategori ini, itulah sebabnya pengukuran tersebut mudah untuk diperiksa.

Mengapa Tingkat Pengukuran Itu Penting

Memilih teknik pemeriksaan yang salah untuk tingkat pengukuran tertentu akan menghasilkan kesimpulan yang menyesatkan. Contoh:

- Menghitung rata-rata kategori nominal akan menghasilkan angka yang tidak bermakna.
- Memperlakukan skala ordinal sebagai skala interval tanpa kehati-hatian dapat mendistorsi interpretasi.
- Menghitung rasio menggunakan suhu dalam derajat Celsius akan menghasilkan kesimpulan yang tidak masuk akal.

Sebelum memilih rumus, model, atau visualisasi di Excel maupun perangkat statistik lainnya, tentukan terlebih dahulu tingkat pengukuran variabel Anda. Klasifikasi yang tepat akan menjaga integritas kesimpulan Anda.

Tabel 2.4 Tingkat Pengukuran dan Implikasi Analitisnya

Tipe Data	**Statistik yang Sesuai**	**Opsi Visualisasi**	**Contoh** pemeriksaan
Nominal	Modus, frekuensi, persentase	Diagram batang, diagram lingkaran	pemeriksaan pangsa pasar
Ordinal	Median, modus, persentil	Diagram batang, diagram batang bertumpuk	Tren kepuasan pelanggan
Interval	Rerata (*mean*), simpangan baku, korelasi	Histogram, plot garis, plot sebar	Kinerja dari waktu ke waktu
Rasio	Seluruh statistik, rasio, tingkat pertumbuhan	Semua jenis visualisasi	Pertumbuhan pendapatan, perhitungan ROI

3. Data Deret Waktu (Time Series): Dimensi Temporal

Definisi: Data deret waktu terdiri dari pengembangan teknologi masa depanyang dicatat secara berurutan sepanjang waktu, baik dalam interval yang teratur maupun tidak teratur. Karakteristik Utama: Urutan pengembangan teknologi masa depansangatlah penting. Waktu bukan sekadar variabel tambahan, melainkan struktur pengatur dari kumpulan data tersebut.

Contohnya meliputi:

- Harga saham yang dicatat setiap menit

- Pengukuran suhu harian
- Angka penjualan bulanan
- Lalu lintas situs web per jam
- Statistik populasi tahunan
- Pemantauan detak jantung secara berkelanjutan

Dalam setiap kasus, urutan pengembangan teknologi masa depanmemiliki makna. Mengubah urutan data akan menghilangkan nilai analitisnya.

Mengapa Data Deret Waktu Bersifat Khusus

Data deret waktu berbeda dengan data penampang melintang (*cross-sectional*) yaitu data yang dikumpulkan pada satu titik waktu tertentu karena pengembangan teknologi masa depandalam data deret waktu sering kali memiliki ketergantungan secara temporal.

1. Ketergantungan Temporal (Autokorelasi)

Dalam banyak data deret waktu, nilai saat ini dipengaruhi oleh nilai di masa lalu. Contohnya:

- Harga saham hari ini dipengaruhi oleh harga kemarin.
- Penjualan bulan ini dipengaruhi oleh penjualan bulan lalu.

Ketergantungan ini dikenal sebagai autokorelasi. Hal ini melanggar asumsi independensi yang disyaratkan oleh banyak metode statistik tradisional.

2. Tren

Tren mencerminkan pergerakan jangka panjang ke arah atas atau bawah dalam data. Contohnya:

- Peningkatan suhu global secara bertahap
- Pertumbuhan pendapatan jangka panjang
- Penurunan biaya manufaktur selama beberapa dekade.

Tren mencerminkan perubahan struktural dari waktu ke waktu.

3. Musiman (*Seasonality*)

Musiman merujuk pada pola yang berulang dalam interval tetap dan teratur dalam periode tertentu. Contohnya:

- Lonjakan penjualan ritel setiap bulan Desember
- Peningkatan konsumsi listrik setiap musim dingin
- Peningkatan pariwisata selama musim liburan tertentu

Pola musiman bersifat dapat diprediksi dan terikat pada siklus berbasis kalender.

4. Siklus (*Cycles*)

Pola siklikal mencerminkan fluktuasi jangka panjang yang tidak terikat secara ketat pada interval kalender yang tetap. Contohnya:

- Ekspansi dan resesi ekonomi
- Lonjakan dan penurunan pasar perumahan

Berbeda dengan pola musiman, siklus tidak terjadi dalam interval yang konsisten dan dapat diprediksi.

Pertimbangan Khusus dalam Pemeriksaan Deret Waktu

Data deret waktu memerlukan penanganan yang cermat:

- pengembangan teknologi masa depantidak dapat diacak begitu saja tanpa kehilangan maknanya.
- Nilai masa lalu sering digunakan untuk memprediksi nilai di masa depan.
- Teknik pemeriksaan khusus mungkin diperlukan, termasuk penghalusan eksponensial (*exponential smoothing*) dan model autoregresif.
- Pola-pola seperti tren dan musiman mungkin perlu dipisahkan sebelum dilakukan pemodelan.

Karena adanya ketergantungan temporal, metode statistik standar yang dirancang untuk pengembangan teknologi masa depanindependen dapat menghasilkan kesimpulan yang keliru jika diterapkan tanpa penyesuaian.

Ilustrasi Dunia Nyata: Data Penjualan Bulanan

Bayangkan sebuah perusahaan yang memeriksa kinerja penjualan bulanan. Data tersebut mungkin mencakup:

- **Tren**: Rata-rata penjualan meningkat 5% setiap tahun.
- **Musiman**: Penjualan di bulan Desember secara konsisten meningkat tiga kali lipat karena liburan.
- **Variasi acak**: Peristiwa tak terduga yang menyebabkan gangguan sementara.

Jika hanya menghitung rata-rata keseluruhan, komponen-komponen struktural ini akan terabaikan. Menguraikan deret data menjadi komponen tren, musiman, siklis, dan tidak beraturan akan memberikan dasar yang jauh lebih akurat untuk melakukan peramalan (*forecasting*).

Mengapa Hal Ini Penting

Data deret waktu (*time series*) memperluas pemikiran statistik dari deskripsi statis menjadi pemeriksaan dinamis. Hal ini memungkinkan kita untuk menjawab pertanyaan seperti:

- Bagaimana kinerja berkembang dari waktu ke waktu?
- Apakah pertumbuhan tersebut berkelanjutan?
- Pola apa yang berulang secara dapat diprediksi?
- Apa yang mungkin terjadi selanjutnya?

Memahami struktur temporal data sangatlah penting untuk peramalan, pemeriksaan ekonomi, pemodelan keuangan, dan perencanaan operasional.

2.5 SUMBER DATA DAN PENGARUHNYA TERHADAP VALIDITAS ANALISIS STATISTIK

Memahami asal-usul data merupakan hal mendasar bagi integritas statistik. Sumber data secara langsung memengaruhi:

- Tingkat keandalannya
- Potensi bias yang mungkin ada
- Biaya dan aksesibilitasnya
- Kekuatan kesimpulan yang dapat ditarik

Sebelum melakukan pemeriksaan apa pun, seorang ilmuwan data harus bertanya: Bagaimana data ini dihasilkan? Jawaban atas pertanyaan tersebut menentukan apa yang dapat dan tidak dapat didukung secara sah oleh data tersebut.

Data Primer vs. Data Sekunder

Data Primer

Data primer dikumpulkan secara langsung oleh peneliti untuk tujuan tertentu. Dalam hal ini, Anda memiliki kendali atas:

- Rancangan studi
- Strategi pengambilan sampel
- Instrumen pengukuran
- Proses pengumpulan data

Contohnya meliputi:

- Survei yang Anda lakukan
- Eksperimen yang Anda jalankan
- Wawancara yang Anda rancang
- Pengukuran lapangan yang Anda catat

Karena dirancang untuk tujuan khusus, data primer sangat selaras dengan pertanyaan studi Anda. Namun, pengumpulan data ini sering kali membutuhkan banyak waktu, biaya, dan perencanaan logistik. Keunggulannya adalah kendali penuh. Konsekuensinya adalah besarnya upaya dan sumber daya yang diperlukan.

Data Sekunder

Data sekunder dikumpulkan oleh pihak lain untuk tujuan yang berbeda atau lebih luas, kemudian digunakan kembali untuk pemeriksaan. Contoh umum meliputi:

- Data sensus pemerintah
- Laporan industri
- Kumpulan data akademis
- Repositori data terbuka
- Basis data perusahaan

Data sekunder biasanya lebih cepat dan lebih murah untuk diperoleh. Namun, Anda harus menerima keputusan rancangan, metode pengambilan sampel, dan keterbatasan pengukuran yang ditetapkan oleh peneliti sebelumnya. Saat menggunakan data sekunder, evaluasi kritis sangatlah penting:

- Mengapa data tersebut dikumpulkan?
- Bagaimana data tersebut diukur?
- Siapa yang disertakan atau dikecualikan?
- Definisi apa yang digunakan?

Kegagalan untuk menelaah pertanyaan-pertanyaan ini dapat berujung pada kesimpulan yang keliru.

Jenis-Jenis Sumber Data

Selain klasifikasi primer dan sekunder, data juga dapat dikategorikan berdasarkan cara data tersebut dihasilkan.

Data Observasional

Data observasional dikumpulkan dengan mencatat peristiwa atau perilaku tanpa memanipulasi variabel apa pun. Contoh:

- Memantau lalu lintas pengunjung di gerai ritel
- Meninjau rekam medis
- Melacak populasi satwa liar

Dalam studi observasional, peneliti berperan sebagai pengamat, bukan sebagai pihak yang melakukan intervensi. Jenis data ini dapat mengidentifikasi hubungan dan pola, namun menetapkan hubungan sebab-akibat (kausalitas) merupakan hal yang sulit karena mungkin terdapat variabel pengganggu (*confounding variables*).

Data Eksperimental

Data eksperimental dihasilkan melalui manipulasi terkontrol terhadap satu atau lebih variabel untuk menguji pengaruhnya terhadap suatu hasil. Contoh:

- Uji klinis
- Pengujian A/B (A/B testing) pada platform digital
- Eksperimen laboratorium terkontrol

Rancangan eksperimental memungkinkan peneliti untuk mengisolasi efek sebab-akibat dengan lebih andal dibandingkan studi observasional. Oleh karena itu, eksperimen yang dirancang dengan baik sering dianggap sebagai standar emas (*gold standard*) untuk inferensi kausal.

Data Transaksional

Data transaksional dihasilkan secara otomatis sebagai bagian dari operasional rutin. Contohnya:

- Catatan penjualan
- Transaksi perbankan
- Log pembelian dalam perdagangan elektronik (*e-commerce*)
- Catatan masuk (*login*) sistem
- Catatan pertemuan klinis pasien

Data ini sering kali berjumlah besar dan sangat terstruktur. Data ini memberikan wawasan mendetail mengenai perilaku operasional, namun mungkin kekurangan variabel kontekstual kecuali jika variabel tersebut sengaja dicatat.

Data yang Dihasilkan Pengguna (*User-Generated Data*)

Data yang dihasilkan pengguna muncul dari aktivitas sukarela atau perilaku individu saat berinteraksi dengan platform digital. Contohnya:

- Unggahan di media sosial
- Ulasan produk
- Kueri pencarian
- Log penggunaan aplikasi

Data jenis ini sangat melimpah, namun sering kali tidak terstruktur (berupa teks, gambar, atau interaksi). Data ini biasanya memerlukan pra-pemrosesan yang ekstensif sebelum dapat dipemeriksaan.

Data Sensor dan IoT

Data sensor dan *Internet of Things* (IoT) dihasilkan secara terus-menerus oleh perangkat fisik. Contohnya:

- Sensor suhu
- Pelacak GPS
- Meteran pintar (*smart meters*)
- Pemantau peralatan manufaktur
- Perangkat kesehatan yang dapat dikenakan (*wearable devices*)

Data ini sering kali memiliki frekuensi dan volume yang tinggi, sehingga menimbulkan tantangan dalam hal penyimpanan, pemrosesan, dan pemeriksaansecara waktu nyata (*real-time*).

Data Administratif dan Pemerintah

Dikumpulkan oleh lembaga publik untuk tujuan tata kelola dan regulasi, data ini mencakup:

- Catatan sensus
- Laporan pajak
- Data penerimaan rumah sakit
- Data pendaftaran sekolah

Kumpulan data ini sering kali berskala besar dan bersifat longitudinal (mencakup rentang waktu yang panjang), sehingga sangat berharga untuk pemeriksaan kebijakan. Namun, data ini mungkin tidak dikumpulkan dengan tujuan utama untuk studi dan bisa memiliki batasan dalam klasifikasinya.

Data Analitik Web dan Digital

Platform digital menghasilkan data perilaku yang mendetail, termasuk:

Jumlah tampilan halaman (*page views*)

- Tingkat pentalan (*bounce rates*)
- Metrik konversi
- Durasi sesi

Kumpulan data ini banyak digunakan dalam pemeriksaan bisnis dan evaluasi kinerja pemasaran. Data ini sering diekspor ke dalam perangkat pemeriksaan, seperti Excel, untuk keperluan pelaporan dan pemodelan.

Mengapa Sumber Data Itu Penting

Kredibilitas suatu kesimpulan statistik bergantung pada pemahaman mengenai asal-usul data (*data provenance*). Sebagai contoh:

- Data observasional mungkin menunjukkan korelasi, tetapi bukan kausalitas (sebab-akibat).
- Data eksperimental memperkuat klaim mengenai hubungan sebab-akibat.
- Data sekunder mungkin mengandung bias pengukuran yang tersembunyi.
- Data transaksional mungkin tidak mencakup faktor kontekstual yang tidak tercatat.

Data tidak berdiri sendiri. Data mencerminkan kondisi saat data tersebut dihasilkan. Seorang ilmuwan data yang piawai tidak sekadar mengpemeriksaan angka. Mereka menelaah asal-usul angka tersebut.

2.6 KUALITAS DATA: FAKTOR PENENTU KEBERHASILAN ATAU KEGAGALAN

Prinsip dasar dalam komputasi dan analitik adalah *Garbage In, Garbage Out* (GIGO) jika data yang dimasukkan buruk, maka hasil yang dikeluarkan pun akan buruk. Kecanggihan statistik secanggih apa pun tidak akan mampu menutupi kekurangan pada data masukan yang cacat. Model yang dibangun berdasarkan data berkualitas rendah akan menghasilkan kesimpulan yang tidak dapat diandalkan, bahkan sering kali memberikan tingkat keyakinan yang menyesatkan. Sebelum memercayai kumpulan data apa pun baik yang dikumpulkan sendiri maupun diperoleh dari sumber eksternal kualitasnya harus dievaluasi secara kritis. Kualitas data bukanlah sekadar masalah tampilan luar. Kualitas data berdampak langsung pada:

- Validitas ringkasan statistik
- Keandalan model prediktif
- Kredibilitas kesimpulan
- Kualitas keputusan yang diambil berdasarkan pemeriksaan

Kumpulan data yang cacat tidak hanya mengurangi presisi, tetapi juga dapat mendistorsi realitas.

Enam Dimensi Utama Kualitas Data

Dimensi kualitas data merupakan kriteria inti yang digunakan untuk mengevaluasi apakah data layak digunakan untuk tujuan tertentu. Dengan kata lain, dimensi-dimensi ini membantu menentukan apakah suatu kumpulan data cukup andal, akurat, dan relevan untuk mendukung pemeriksaan yang bermakna serta pengambilan keputusan yang tepat. Setiap dimensi menyoroti aspek spesifik yang menentukan apakah data tersebut dapat mendukung atau justru melemahkan kesimpulan yang valid.

1. Akurasi

Definisi: Apakah data merepresentasikan fenomena dunia nyata yang hendak diukur secara tepat?

Pertanyaan Analitis

- Apakah pengukuran dicatat dengan benar?
- Apakah nilai yang tercatat mencerminkan nilai yang sebenarnya?
- Apakah instrumen telah dikalibrasi dengan baik?
- Apakah terdapat kesalahan dalam penyalinan atau input data?

Masalah Umum

- Kesalahan pengetikan (misalnya 150, padahal seharusnya 15,0)
- Kerusakan sensor
- Kesalahan manusia (*human error*) dalam pencatatan manual
- Informasi yang sudah kedaluwarsa (misalnya alamat lama yang masih tersimpan)

- Kesalahan saat menyalin dan menempel (*copy-paste*) data dalam lembar kerja (*spreadsheet*)

Contoh: Basis data rumah sakit mencatat berat badan seorang pasien sebesar 500 kg. Kemungkinan besar terjadi kesalahan penempatan desimal (seharusnya 50,0 kg). Ketidakakuratan seperti ini dapat mendistorsi pemeriksaan dan, dalam konteks klinis, menimbulkan risiko serius.

Metode Penilaian

- Validasi silang dengan sumber tepercaya
- Pemeriksaan acak (*spot check*)
- Identifikasi nilai yang secara fisik tidak mungkin terjadi
- Perbandingan dengan tolok ukur historis atau eksternal

Akurasi berfungsi sebagai perlindungan terhadap kesalahan pengukuran; salah satu ancaman paling mendasar terhadap validitas statistik.

2. Kelengkapan (*Completeness*)

Definisi: Apakah semua data yang diperlukan tersedia, atau apakah ada nilai-nilai penting yang hilang?

Pertanyaan Analitis

- Apakah kolom-kolom yang wajib diisi telah terisi?
- Apakah variabel tertentu hilang secara sistematis?
- Apakah periode historis tercakup sepenuhnya?
- Apakah kelompok tertentu kurang terwakili (*underrepresented*)?

Masalah Umum

- Pertanyaan survei yang tidak dijawab
- Kegagalan sensor yang menyebabkan kekosongan data pada rentang waktu tertentu
- Migrasi sistem yang tidak tuntas
- Pencatatan selektif terhadap nilai-nilai ekstrem

Contoh: Jika 60% responden survei tidak mencantumkan informasi pendapatan, maka pemeriksaan apa pun yang berbasis pendapatan berisiko mengalami bias, terutama jika ketidakmauan untuk menjawab tersebut terjadi secara sistematis.

Dampak Data yang Hilang (*Missing Data*)

- Berkurangnya ukuran sampel, yang berdampak pada penurunan kekuatan statistik (*statistical power*)
- Estimasi parameter yang bias (jika data hilang secara tidak acak)
- Menghalangi pelaksanaan pemeriksaan tertentu
- Dapat mengarah pada kesimpulan yang keliru

Kondisi data yang hilang (*missingness*) diklasifikasikan sebagai berikut:

- *Missing Completely at Random* (MCAR) – Hilang sepenuhnya secara acak
- *Missing at Random* (MAR) – Hilang secara acak (bergantung pada variabel lain yang teramati)
- *Missing Not at Random* (MNAR) – Hilang tidak secara acak (bergantung pada nilai yang hilang itu sendiri)

Setiap jenis kondisi data yang hilang memiliki implikasi analitis yang berbeda dan memerlukan strategi penanganan yang berbeda pula. Kelengkapan data memengaruhi representativitas serta kekuatan inferensi (penarikan kesimpulan).

3. Konsistensi

Definisi: Apakah data seragam di seluruh catatan, sistem, dan rentang waktu?

Pertanyaan Analitis

- Apakah format sudah distandarisasi (tanggal, satuan, kategori)?
- Apakah kolom-kolom yang saling berkaitan sudah selaras secara logis?
- Apakah definisi telah berubah seiring berjalannya waktu?
- Apakah kumpulan data (*dataset*) yang digabungkan menggunakan standar yang kompatibel?

Masalah Umum

- Format tanggal yang beragam (contoh: 2024-10-15 vs. 10/15/2024 vs. 15-Okt-2024)
- Satuan pengukuran yang berbeda (contoh: cm vs. inci)
- Variasi kategori (contoh: “NY” vs. “New York” vs. “NYC”)
- Kontradiksi logis (contoh: tanggal lahir tercatat setelah tanggal kematian)
- Definisi variabel yang diubah di tengah tahun

Contoh: Jika pendapatan dicatat dalam Rp tanpa keterangan mata uang, maka total agregat menjadi tidak bermakna.

Dampak

- Munculnya duplikat semu atau kegagalan dalam mencocokkan data (*missed matches*)
- Kesalahan dalam agregasi data
- Distorsi pada pemeriksaan tren
- Ketidakakuratan dalam perbandingan antar-sistem

Konsistensi dapat dijaga dengan menstandarisasi proses, menetapkan aturan validasi data, mengikuti protokol pengumpulan data yang seragam, serta memastikan adanya dokumentasi yang jelas.

4. Ketepatan Waktu (*Timeliness*)

Definisi: Apakah data cukup mutakhir untuk tujuan penggunaannya? Ketepatan waktu bergantung pada konteks.

- Data sensus mungkin tetap berguna selama bertahun-tahun.
- Harga saham bisa menjadi usang dalam hitungan detik.
- Preferensi konsumen dapat berubah dalam hitungan bulan.

Pertanyaan Analitis

- Kapan data tersebut dikumpulkan?
- Seberapa cepat fenomena yang diamati mengalami perubahan?
- Apakah terdapat jeda waktu dalam pelaporan (*reporting lag*)?
- Apakah pembaruan data dilakukan dengan frekuensi yang memadai?

Contoh:

Menggunakan data demografis tahun 2020 untuk merencanakan ekspansi ritel tahun 2027 di suatu wilayah yang berkembang pesat dapat menghasilkan keputusan strategis yang keliru. Ketepatan waktu memengaruhi akurasi prediksi dan relevansi keputusan.

5. Validitas

Definisi: Apakah data sesuai dengan format, tipe, aturan logika, dan batasan bisnis yang telah ditetapkan sebelumnya?

Pertanyaan Analitis

- Apakah nilai-nilai yang ada berada dalam rentang yang dapat diterima?
- Apakah tipe data sesuai dengan definisinya?
- Apakah hubungan logis tetap terjaga?
- Apakah aturan bisnis terpenuhi?

Masalah Umum

- Usia bernilai negatif atau usia yang tidak mungkin (contoh: -5 atau 250 tahun)
- Tanggal di masa depan untuk peristiwa yang terjadi di masa lalu
- Alamat surel (email) tanpa simbol @
- Nomor telepon dengan jumlah digit yang tidak tepat
- Persentase di luar rentang 0–100
- Data kategorikal dengan nilai yang tidak terduga

Contoh: Jika total pesanan lebih kecil daripada jumlah dari setiap item di dalamnya, berarti terdapat kegagalan dalam logika perhitungan atau integritas data.

Validitas ditegakkan melalui:

- Validasi input
- Batasan basis data (*database constraints*)
- Pemeriksaan aturan otomatis
- Verifikasi logika bisnis

Validitas menjaga integritas struktural.

6. Keunikan (Uniqueness)

Definisi: Apakah setiap entitas di dunia nyata direpresentasikan tepat satu kali?

Pertanyaan Analitis

- Apakah terdapat catatan (*record*) ganda?
- Apakah tersedia pengidentifikasi unik yang dapat diandalkan?
- Apakah penggabungan kumpulan data menimbulkan redundansi?

Masalah Umum

- Beberapa ID pelanggan untuk satu individu yang sama
- Transaksi ganda
- Variasi kecil pada nama (contoh: “Bob Smith” vs. “Robert Smith”)

Contoh: Catatan surel ganda dapat menyebabkan pesan pemasaran dikirim berulang kali, meningkatkan angka berhenti berlangganan (*unsubscribe*), serta mendistorsi metrik pelanggan.

Dampak

- Jumlah yang membengkak (*inflated counts*)
- Statistik frekuensi yang menyesatkan
- Operasional yang tidak efisien
- Kerusakan hubungan dengan pelanggan

Keunikan dijaga melalui:

- Pengidentifikasi unik
- Prosedur penghapusan data ganda (*deduplikasi*)
- Algoritma pencocokan samar (*fuzzy matching*)
- Sistem pengelolaan data induk (*master data management*)

Kualitas Data sebagai Disiplin Profesional

Membangun kebiasaan untuk melakukan pemantauan kualitas data secara terstruktur merupakan salah satu kompetensi paling berharga yang dapat dikembangkan oleh seorang ilmuwan data (*data scientist*). Hal ini mencerminkan:

- Kematangan analitis
- Tanggung jawab profesional
- Penghargaan terhadap bukti

Pemeriksaan data tidak dimulai dengan pembuatan grafik atau model, melainkan dengan validasi. Pemantauan yang disiplin terhadap kualitas data akan melindungi kesimpulan Anda sekaligus menjaga kredibilitas Anda.

Memahami dan Mengelola Keterbatasan Data

Data yang sempurna hanyalah mitos. Setiap kumpulan data di dunia nyata mengandung ketidaksempurnaan:

- Nilai yang hilang (*missing values*)
- Kesalahan pengukuran
- Ketidakkonsistenan
- Penundaan (keterlambatan data)
- Bias

Tanggung jawab profesional seorang ilmuwan data (*data scientist*) bukanlah mencari kesempurnaan, melainkan untuk:

- Memahami keterbatasan kumpulan data
- Meningkatkan kualitas jika memungkinkan
- Memitigasi masalah secara metodis
- Mendokumentasikan asumsi secara transparan
- Mengomunikasikan keterbatasan secara jujur

Penilaian kualitas data bukanlah sekadar tugas awal, melainkan disiplin yang menjadi fondasi.

2.7 ETIKA DAN TANGGUNG JAWAB DATA DALAM PENGELOLAAN DATA

Data tidaklah netral. Setiap kumpulan data membawa jejak keputusan manusia apa yang diukur, bagaimana cara mengukurnya, siapa yang disertakan, siapa yang dikecualikan, dan bagaimana kategori didefinisikan. Sebagai seorang ilmuwan data, tanggung jawab Anda

melampaui akurasi teknis hingga mencakup pertimbangan etis. Kompetensi statistik tanpa kesadaran etis dapat menghasilkan dampak yang merugikan.

Privasi dan Perlindungan Data

Data pribadi termasuk nama, alamat, catatan kesehatan, informasi keuangan, dan pengenal digital adalah milik individu yang diwakilinya. Pengumpulan dan penggunaannya membebankan kewajiban bagi pihak yang menanganinya. Kerangka regulasi seperti *General Data Protection Regulation* (GDPR) menetapkan standar hukum yang mengatur:

- Pengumpulan data
- Penyimpanan dan keamanan
- Pemrosesan dan pemeriksaan
- Berbagi data dan akses pihak ketiga

Ketidakpatuhan dapat berujung pada sanksi hukum. Lebih penting lagi, hal ini merusak kepercayaan publik aset krusial dalam sistem yang berbasis data. Praktik etis menuntut tidak hanya kepatuhan terhadap hukum, tetapi juga penghormatan terhadap martabat manusia dan kerahasiaan.

Persetujuan dan Pembatasan Tujuan

Penggunaan data secara etis memerlukan persetujuan berdasarkan informasi yang jelas (*informed consent*). Individu harus memahami:

- Data apa yang dikumpulkan
- Mengapa data tersebut dikumpulkan
- Bagaimana data tersebut akan digunakan
- Siapa saja yang akan memiliki akses terhadap data tersebut

Menggunakan data yang dikumpulkan untuk satu tujuan guna menjawab pertanyaan yang sangat berbeda tanpa adanya persetujuan baru dapat melanggar norma etika, meskipun secara teknis mungkin diperbolehkan. Pembatasan tujuan melindungi individu dari penyalahgunaan dan mencegah pergeseran tujuan (*mission drift*) dalam pemeriksaan.

Bias dan Keadilan

Bias dapat muncul dalam data pada berbagai tahapan:

- **Bias pengambilan sampel (*sampling bias*)**: kelompok tertentu terwakili secara berlebihan atau justru kurang terwakili.
- **Bias pengukuran (*measurement bias*)**: instrumen secara sistematis mendistorsi hasil.
- **Bias pelaporan (*reporting bias*)**: pencatatan informasi yang bersifat selektif. Ketika data yang bias digunakan untuk melatih model prediktif, bias tersebut dapat diperkuat dan dilembagakan.

Sebagai contoh, model perekrutan yang dilatih menggunakan data rekrutmen yang secara historis mengandung bias dapat mereplikasi pola-pola diskriminatif, meskipun tidak ada aturan diskriminatif yang diprogram secara eksplisit. Sains data yang etis menuntut kesadaran bahwa kumpulan data mencerminkan konteks sosial. Ketelitian analitis harus dibarengi dengan kesadaran akan keadilan.

Transparansi dan Reproduksibilitas

Transparansi menuntut agar metode, asumsi, dan sumber data didokumentasikan dengan cukup jelas untuk memungkinkan verifikasi independen. Reproduksibilitas memastikan bahwa analis lain dapat mengikuti langkah-langkah yang sama dan memperoleh hasil yang sebanding. Dalam praktiknya termasuk dalam alur kerja Excel hal ini berarti:

- Menyimpan data mentah secara terpisah dari data yang telah diproses
- Mendokumentasikan langkah-langkah pembersihan data
- Memberi label yang jelas pada setiap transformasi data
- Mencatat rumus dan asumsi yang digunakan
- Menghindari penyesuaian manual yang tidak didokumentasikan

Pemeriksaan yang transparan membangun kredibilitas dan memungkinkan validasi oleh rekan sejawat.

Daftar Periksa Etika Praktis

Sebelum bekerja dengan kumpulan data apa pun, pertimbangkan hal-hal berikut:

1. Apakah saya memiliki hak hukum untuk mengakses dan menggunakan data ini?
2. Apakah individu yang terlibat telah diberi informasi dan memberikan persetujuan jika diperlukan?
3. Apakah kumpulan data tersebut merepresentasikan populasi yang diklaimnya secara adil?
4. Sudahkah saya memeriksa potensi sumber bias?
5. Apakah asumsi dan batasan saya didokumentasikan dengan jelas?
6. Apakah saya merasa nyaman untuk mempertahankan metodologi dan kesimpulan saya di depan publik?

Pertanyaan-pertanyaan ini melindungi individu yang tercakup dalam data sekaligus menjaga integritas pemeriksaan Anda.

Etika sebagai Disiplin Profesional

Etika dalam sains data bukanlah sekadar konsep filosofis yang abstrak, melainkan sebuah disiplin profesional. Pemeriksaan yang bertanggung jawab menuntut:

- Penghormatan terhadap privasi
- Kesadaran akan adanya bias
- Komitmen terhadap transparansi
- Akuntabilitas atas konsekuensi yang timbul

Keterampilan teknis memungkinkan Anda untuk menggali wawasan (*insight*). Pertimbangan etis memastikan bahwa wawasan tersebut digunakan secara bertanggung jawab. Di dunia yang digerakkan oleh data, kredibilitas sama berharganya dengan kompetensi. Jagalah keduanya.

2.8 SIKLUS HIDUP DATA: DARI PERANCANGAN HINGGA PENGHAPUSAN

Data tidak muncul dan menghilang begitu saja. Data bergerak melalui siklus hidup yang terstruktur, yang dibentuk oleh keputusan-keputusan yang diambil secara sadar pada setiap tahapannya. Memahami siklus hidup ini memungkinkan praktisi sains data untuk bekerja

secara sistematis, bertanggung jawab, dan efisien. Setiap tahapan memengaruhi integritas, kegunaan, dan implikasi etis dari data pada tahap-tahap selanjutnya.

Tahap 1: Perencanaan dan Perancangan

Sebelum data dikumpulkan, data tersebut dirancang terlebih dahulu. Pada tahap ini, keputusan dibuat mengenai:

- Masalah apa yang sedang ditangani
- Variabel apa yang harus diukur
- Bagaimana data akan dikumpulkan
- Siapa atau apa yang akan dimasukkan ke dalam populasi
- Batasan hukum dan etika

Perencanaan yang buruk akan menghasilkan kumpulan data yang tidak dapat digunakan atau mengandung bias. Perancangan yang matang akan meminimalkan kesalahan di tahap selanjutnya serta memastikan keselarasan antara tujuan bisnis dan hasil yang dapat diukur. Pertimbangan utama:

- Tujuan studi atau bisnis yang jelas
- Variabel dan standar pengukuran yang telah ditetapkan
- Persetujuan etis dan kepatuhan hukum
- Strategi pengambilan sampel
- Antisipasi terhadap kebutuhan penyimpanan dan pemrosesan

Tahap 2: Pengumpulan / Pembuatan

Data dibuat, diamati, dicatat, atau diperoleh. Keputusan dalam pengumpulan data menentukan kualitas, bias, dan keterwakilan data. Kesalahan yang terjadi pada tahap ini sering kali akan terus terbawa sepanjang siklus hidup data.

Pertimbangan utama:

- Sampel yang representatif (pengendalian bias)
- Persetujuan berdasarkan informasi (*informed consent*) harus diperoleh jika diperlukan
- Prosedur entri data yang terstandarisasi: memastikan konsistensi sejak awal
- Validasi saat pengambilan data: menolak nilai yang tidak valid secara sistematis pada titik entri
- Struktur dan format awal: Tentukan jenis data sejak dini

Tahap 3: Penyimpanan

Setelah dikumpulkan, data harus disimpan dengan aman dan efisien. Keputusan mengenai penyimpanan memengaruhi aksesibilitas, skalabilitas, kinerja, keamanan, dan biaya. Pertimbangan utama:

- Keamanan & Perlindungan Data: Terapkan kontrol yang ketat (enkripsi, autentikasi, *firewall*)
- Redundansi & Pemulihan Bencana: Lindungi data dengan strategi pencadangan otomatis dan redundansi (menyimpan salinan di lokasi yang berbeda).
- Format Penyimpanan yang Dioptimalkan: Pilih format yang tepat (misalnya, CSV, Parquet, tabel relasional)

- Tata Kelola & Kontrol Akses: Gunakan *Role-Based Access Control* (RBAC) untuk memastikan setiap pihak hanya dapat mengakses data yang mereka perlukan.

Data yang disimpan dengan buruk akan menjadi sulit diakses, tidak aman, atau memakan biaya besar untuk dikelola.

Tahap 4: Pemrosesan dan Persiapan

Data mentah jarang sekali langsung siap untuk diperiksa. Data tersebut harus dibersihkan, ditransformasikan, dan disusun. Tahap ini menentukan keandalan pemeriksaan. Pertimbangan utama:

- Pendokumentasian Transformasi: Jangan pernah mengubah data mentah tanpa adanya catatan.
- Menjaga Silsilah Data (*Data Lineage*): Anda harus mampu melacak perjalanan data, mulai dari asal-usulnya hingga kondisinya saat ini.
- Kontrol Versi (*Version Control*): Gunakan sistem penomoran versi yang terstruktur untuk kumpulan data dan skrip Anda, sehingga Anda dapat kembali ke kondisi sebelumnya jika terjadi kesalahan dalam proses transformasi.

Persiapan bukan sekadar masalah teknis, melainkan persoalan metodologis. Transparansi pada tahap ini mendukung prinsip reproduksibilitas dan akuntabilitas.

Tahap 5: Pemeriksaan dan Pemodelan

Data diperiksa untuk menggali wawasan, menguji hipotesis, atau membangun sistem prediktif. Tahap ini merupakan inti intelektual dari sains data, namun sangat bergantung pada tahapan-tahapan sebelumnya. Beberapa pertimbangan utama:

- Gunakan metode yang sesuai dengan jenis data dan tujuan yang ingin dicapai.
- Validasi asumsi: Sebagian besar uji statistik memiliki ‘syarat dan ketentuan’ tertentu.
- Kesadaran akan ukuran sampel dan bias
- Validasi silang atau tolok ukur (*benchmarking*): bandingkan hasil Anda dengan tolok ukur dasar (*baseline*) atau standar industri.
- Dokumentasi untuk reprodusibilitas: Dokumentasikan logika Anda agar analis lain atau auditor dapat mengikuti langkah-langkah yang Anda ambil dan mencapai kesimpulan yang sama.

Kecanggihan analitis tidak dapat menutupi kekurangan dalam proses-proses awal (*upstream processes*).

Tahap 6: Berbagi dan Publikasi

Wawasan, visualisasi, dasbor, laporan, atau model disampaikan kepada para pemangku kepentingan atau publik. Tahap ini mengubah pemeriksaan menjadi tindakan nyata. Pertimbangan utama:

- Melindungi privasi dalam hasil yang dibagikan
- Mengomunikasikan asumsi dan keterbatasan secara jelas
- Menghindari visualisasi yang menyesatkan
- Menyesuaikan komunikasi dengan audiens yang dituju
- Memastikan model dapat diinterpretasikan

Komunikasi yang bertanggung jawab sama pentingnya dengan pemeriksaan yang akurat.

Tahap 7: Pemantauan dan Pemeliharaan

Banyak sistem data bersifat dinamis. Dasbor diperbarui, model dilatih ulang, dan alur data (*pipeline*) berjalan secara berkelanjutan. Pemantauan berkelanjutan memastikan validitas tetap terjaga. Pertimbangan utama:

- Mendeteksi pergeseran data (*data drift*)
- Memantau kinerja model
- Memperbarui kumpulan data saat terjadi perubahan struktur
- Melakukan audit terhadap kerentanan keamanan
- Memvalidasi ulang asumsi secara berkala

Mengabaikan tahap ini dapat menyebabkan penurunan kualitas sistem pengambilan keputusan secara diam-diam.

Tahap 8: Pengarsipan (*Archival*)

Ketika penggunaan aktif menurun, data dapat dipindahkan ke penyimpanan jangka panjang. Pengarsipan menjaga memori institusional sekaligus mengurangi beban operasional. Pertimbangan utama meliputi:

- Menentukan nilai retensi
- Memilih format penyimpanan yang tahan lama
- Melestarikan metadata dan dokumentasi
- Memastikan data dapat dipahami di masa depan

Data yang diarsipkan tanpa konteks akan menjadi tidak dapat digunakan.

Tahap 9: Penghapusan (*Deletion*)

Data tidak seharusnya ada selamanya secara otomatis. Ketika data tidak lagi diperlukan atau ketika kewajiban hukum, kontraktual, maupun etis mengharuskan penghapusannya data tersebut harus dimusnahkan dengan aman. Pertimbangan utama meliputi:

- Persyaratan retensi secara hukum
- Protokol penghapusan yang aman
- Memenuhi permintaan penghapusan dari subjek data
- Dokumentasi mengenai apa yang dihapus dan alasannya

Penghapusan yang bertanggung jawab akan melindungi privasi, mengurangi risiko, serta menunjukkan tata kelola yang disiplin.

Prinsip Utama

Siklus hidup data bukan sekadar urutan teknis, melainkan kerangka kerja untuk akuntabilitas. Setiap tahapan membawa konsekuensi terhadap:

- Kualitas data
- Kepatuhan etis
- Validitas analitis
- Kepercayaan organisasi

Menguasai siklus hidup data berarti berpikir melampaui sekadar pemeriksaan. Hal ini berarti memahami bahwa praktik data yang bertanggung jawab dimulai sebelum pengumpulan data dan terus berlanjut bahkan setelah data dihapus. Manajemen data bukanlah sekadar tugas administratif, melainkan kompetensi inti dalam bidang sains data profesional.

2.9 PERTEMUAN PERTAMA ANDA DENGAN KUMPULAN DATA (*DATASET*)

Saat seorang ilmuwan data menerima kumpulan data baru, tugas pertamanya bukanlah menghitung rata-rata atau melatih model. Awal yang sesungguhnya dari sebuah pemeriksaan adalah orientasi. Pemeriksaan tidak dimulai dengan rumus, melainkan dengan pemahaman. Sebelum menarik kesimpulan, Anda harus memahami konteks, struktur, dan makna dari data tersebut. Langkah-langkah awal ini berfungsi sebagai kompas, memastikan bahwa pemeriksaan selanjutnya tetap akurat, relevan, dan selaras dengan tujuan di dunia nyata. Pemeriksaan awal yang disiplin dapat mencegah kesalahan yang merugikan di kemudian hari.

1. Tinjau Kembali Tujuan Pemeriksaan

Sebelum memeriksa satu nilai pun, pahami kembali tujuan dari pemeriksaan tersebut.

- Pertanyaan apa yang ingin Anda jawab?
- Keputusan apa yang akan didasarkan pada pemeriksaan ini?
- Siapa pemangku kepentingannya?

Memperjelas tujuan sejak awal akan menjaga fokus pekerjaan Anda dan mencegah eksplorasi yang tidak perlu. Setiap transformasi, ringkasan, dan model harus mendukung tujuan utama tersebut. Tanpa tujuan yang jelas, pemeriksaan yang secara teknis benar sekalipun bisa kehilangan arah.

2. Pratinjau Kumpulan Data (Dataset)

Mulailah dengan pemeriksaan visual secara menyeluruh.

- Lihat beberapa baris pertama.
- Telusuri isi kumpulan data tersebut.
- Amati pola dalam struktur dan formatnya.

Langkah sederhana ini sering kali mengungkap:

- Nilai yang hilang (*missing values*) yang tidak terduga
- Ketidakkonsistenan format
- Kesalahan entri data yang nyata
- Anomali struktural

Pratinjau singkat membantu membangun intuisi dan menyoroti bagian mana yang memerlukan perhatian lebih mendalam.

3. Memahami Struktur Data

Periksa bagaimana kumpulan data tersebut disusun.

- Berapa jumlah baris dan kolom yang ada?
- Apa yang diwakili oleh setiap kolom?
- Apa jenis datanya (numerik, kategorikal, tanggal, boolean)?
- Apakah pengidentifikasi (*identifier*) sudah dibedakan dengan benar dari variabel yang dapat diukur?

Tahap ini membentuk gambaran mental Anda mengenai kumpulan data tersebut. Anda tidak dapat memilih metode pemeriksaan yang tepat sebelum memahami jenis data yang sedang Anda tangani. Kesalahan dalam mengklasifikasikan pengidentifikasi sebagai variabel numerik,

atau memperlakukan data tanggal sebagai teks, dapat merusak hasil pemeriksaan sejak awal. Struktur menentukan strategi.

4. Membaca Metadata atau Kamus Data

Metadata adalah data mengenai data itu sendiri. Metadata menyediakan konteks yang penting dan menjelaskan:

- Definisi variabel
- Satuan pengukuran
- Kode kategori dan penyandian (*encoding*)
- Metode pengumpulan data
- Tingkat agregasi

Sebelum mengpemeriksaan kumpulan data yang dikumpulkan oleh pihak lain pastikan apa yang sebenarnya diwakili oleh setiap variabel. Langkah ini mencegah kesalahan mendasar seperti:

- Menafsirkan rasio atau tingkat (*rates*) sebagai jumlah mentah (*raw counts*)
- Keliru menganggap kategori yang disandikan sebagai variabel numerik
- Memperlakukan nilai agregat sebagai data tingkat individu

Asumsi yang dibuat pada tahap ini akan terus terbawa sepanjang proses pemeriksaan. Ketelitian di tahap ini akan menjaga validitas hasil di kemudian hari.

5. Periksa Integritas Data Dasar

Bahkan kumpulan data yang tampak bersih pun tetap memerlukan verifikasi. Lakukan pemeriksaan integritas awal:

- Apakah terdapat nilai yang tidak mungkin (contoh: usia bernilai negatif)?
- Apakah terdapat pengidentifikasi (*identifier*) ganda?
- Apakah terdapat format yang tidak konsisten?
- Apakah rentang nilai tampak masuk akal?

Pemeriksaan ini menentukan apakah kumpulan data tersebut cukup tepercaya untuk dieksplorasi lebih dalam. Kepercayaan harus dibangun, bukan sekadar diasumsikan.

Mengapa Tahap Ini Penting

Langkah-langkah awal ini membangun pemahaman dan kendali. Hal ini mengurangi risiko dalam pemeriksaan, mempertajam fokus, dan meningkatkan keyakinan sebelum melangkah ke tahap peringkasan atau pemodelan. Pemeriksaan data profesional tidak dilakukan dengan terburu-buru, melainkan dengan cermat dan terencana. Dengan memahami konteks, struktur, dan integritas data, Anda menciptakan landasan yang kokoh untuk menghasilkan wawasan yang bermakna. Pada bab-bab selanjutnya, kita akan mengembangkan landasan ini dengan memperkenalkan berbagai ukuran dan teknik utama yang memungkinkan Anda untuk meringkas, mengeksplorasi, serta menafsirkan data secara jelas dan akurat.

2.10 IKHTISAR PENGELOLAAN DAN PEMAHAMAN DATA

Dalam bab ini, kita telah membangun landasan yang komprehensif untuk memahami data sebelum melakukan pemeriksaan. Kita memulainya dengan menjawab pertanyaan

mendasar: Apa itu data dalam statistika? Data adalah kumpulan nilai atau fakta mentah mengenai dunia yang direkam untuk tujuan tertentu. Dalam pola pikir statistika, melalui pengorganisasian, pemeriksaan, dan konteks, data akan berubah menjadi informasi, dan pada akhirnya menjadi pengetahuan. Selanjutnya, kita mempelajari anatomi data bagaimana kumpulan data disusun dalam baris dan kolom, serta perbedaan antara variabel dan pengembangan teknologi masa depan. Memahami struktur sangatlah penting karena jika data disusun dengan benar, proses pemeriksaan akan menjadi jauh lebih mudah. Kemudian, kita mengeksplorasi berbagai jenis data kategorikal dan numerikal, nominal dan ordinal, diskret dan kontinu serta mengapa pembedaan ini penting.

Kita juga membahas tingkat pengukuran (*levels of measurement*) dan bagaimana jenis data serta tingkat pengukuran suatu variabel menentukan metode apa yang dapat diterapkan padanya. Ketepatan dalam pemeriksaan dimulai dengan mengidentifikasi secara benar jenis data yang sedang Anda tangani. Kita juga menelaah sumber data, dengan memahami bahwa data bisa bersifat primer atau sekunder, serta dapat berasal dari survei, eksperimen, sistem administratif, sensor, transaksi, atau platform digital. Setiap sumber memiliki implikasi terhadap keandalan, bias, dan cakupan data. Mengetahui asal-usul data sama pentingnya dengan mengetahui isi data tersebut. Bab ini kemudian membahas kualitas data, dengan menekankan bahwa wawasan yang dapat diandalkan bergantung pada input yang juga dapat diandalkan. Kami menelaah enam dimensi utama kualitas akurasi, kelengkapan, konsistensi, ketepatan waktu, validitas, dan keunikan serta melihat bagaimana pemeriksaan kualitas yang sistematis dapat menjaga integritas pemeriksaan.

Selanjutnya, kami mendiskusikan etika dan tanggung jawab dalam pengelolaan data. Data dibentuk oleh keputusan manusia dan dapat memengaruhi kehidupan manusia. Isu-isu seperti privasi, persetujuan, keadilan, bias, transparansi, dan reproduksibilitas bukanlah masalah sampingan, melainkan aspek sentral dalam praktik profesional di bidang data. Keterampilan teknis tanpa kesadaran etis akan merusak kepercayaan. Kami kemudian memperkenalkan siklus hidup data, dengan memahami bahwa data melalui berbagai tahapan perencanaan, pengumpulan, penyimpanan, persiapan, pemeriksaan, berbagi, pemantauan, pengarsipan, hingga penghapusan. Praktik sains data yang efektif menuntut pemahaman terhadap keseluruhan siklus hidup tersebut, bukan sekadar tahap pemeriksaannya saja.

Terakhir, kami membahas langkah awal saat berhadapan dengan sebuah dataset, dengan menekankan pentingnya orientasi sebelum melakukan komputasi. Meninjau kembali tujuan, melihat pratinjau data, memahami struktur, membaca metadata, serta memeriksa integritas data merupakan titik awal yang disiplin untuk setiap pemeriksaan yang bertanggung jawab. Pada bab berikutnya, kita akan beralih dari pemahaman mengenai sifat dan struktur data menuju pengukuran serta peringkasan data mengubah persiapan yang disiplin menjadi wawasan kuantitatif.

BAB 3
RINGKASAN DASAR & TAMPILAN TABULAR

3.1 AGREGASI DATA UNTUK MENGUNGKAP POLA DAN TREN

Maria mengelola sebuah kedai kopi kecil di kawasan bisnis pusat kota, tempat ia melayani puluhan hingga ratusan pelanggan setiap harinya. Ia mencatat setiap transaksi dengan cermat, meliputi tanggal, waktu, jenis minuman, metode pembayaran, dan harga. Setelah tiga bulan, Maria memiliki lembar kerja (*spreadsheet*) yang terus berkembang, berisi ribuan baris data. Ia tahu bisnisnya sedang tumbuh. Ia melihat pelanggan mengantre. Ia menghitung uang tunai saat kedai tutup. Namun, ketika seorang calon investor bertanya,

1. *"Apa produk Anda yang paling populer?"*
2. *"Bagaimana perbandingan penjualan di hari kerja dengan akhir pekan?"*

Maria membuka laptopnya. Ia menelusuri 8.000 baris data transaksi. Jawabannya ada di suatu tempat dalam lembar kerja tersebut ia yakin akan hal itu tetapi layar hanya menampilkan deretan tanggal, angka, dan nama minuman yang tak berujung. Setelah keheningan yang canggung, ia mengatakan akan memberikan jawaban melalui surel.Datanya lengkap tersedia, tetapi wawasannya tidak ada.

Situasi yang dihadapi Maria adalah salah satu hal paling umum dalam analisis data: data mentah, betapapun cermatnya dikumpulkan, tidak dapat berbicara dengan sendirinya. Lembar kerja yang penuh dengan transaksi individual tidak bisa menjawab pertanyaan mengenai pola. Untuk mengubah data mentah menjadi pemahaman, mulailah dengan agregasi seperti total, jumlah, frekuensi, dan proporsi. Pendekatan ini menyusun data Anda ke dalam format yang mengungkap tren di baliknya. Itulah topik pembahasan bab ini.

Keterampilan yang diperkenalkan dalam bab ini bersifat mendasar. Meski tampak sederhana, keterampilan ini merupakan langkah nyata pertama dalam penalaran statistik: beralih dari catatan individual menuju ringkasan yang terstruktur. Setiap teknik statistik tingkat lanjut mulai dari distribusi probabilitas hingga model regresi bergantung pada kemampuan untuk menyusun data secara bermakna. Kuasailah perangkat-perangkat ini, maka Anda akan mampu menghadirkan kejelasan di tengah kompleksitas. Tanpa perangkat tersebut, kumpulan data sebesar apa pun hanyalah deretan pertanyaan yang belum terjawab.

3.2 AGREGASI DATA: PENGHITUNGAN, PERSENTASE, DAN UKURAN RELATIF

Sebelum menghitung nilai rata-rata, varians, atau korelasi, kita harus menjawab pertanyaan yang lebih mendasar: Berapa jumlahnya? Menghitung mungkin tampak sederhana, namun ini merupakan langkah awal dalam penalaran statistik. Setiap analisis tingkat lanjut dimulai dengan penghitungan—menghitung kategori, hasil, atau kejadian. Tanpa penghitungan, tidak akan ada proporsi. Tanpa total, tidak akan ada nilai rata-rata. Penghitungan mengubah data mentah menjadi struktur yang dapat diukur.

Jumlah dan Total

Jumlah (atau frekuensi absolut) merujuk pada berapa kali suatu nilai atau kategori muncul dalam kumpulan data. Total merujuk pada hasil penjumlahan nilai numerik dari seluruh observasi, atau dari observasi yang memenuhi kriteria tertentu. Ini adalah bentuk ringkasan statistik yang paling sederhana dan sering kali justru yang paling informatif. Contoh: Metode Pembayaran, Maria memeriksa transaksi dalam satu hari dan mencatat 50 pembelian. Ia menghitung metode pembayaran yang digunakan:

- Tunai: 18
- Kartu Kredit/Debit: 27
- Pembayaran Seluler (*Mobile Payment*): 5

Ia juga menghitung total pendapatan harian: Rp 12.475.200. Dari ringkasan dasar ini, wawasan yang bermakna pun muncul:

- Pembayaran dengan kartu mendominasi.
- Pembayaran seluler relatif jarang digunakan.
- Rata-rata nilai transaksi adalah sekitar Rp 250.000 (total pendapatan ÷ jumlah transaksi).

Bahkan tanpa metode yang rumit, Maria sudah bisa mulai mengambil keputusan bisnis. Hal ini menunjukkan prinsip utama:nWawasan statistik bermula dari agregasi yang terstruktur.

Penghitungan di Excel

Excel menyediakan serangkaian fungsi yang dirancang untuk menghitung frekuensi secara efisien dan akurat. Alat-alat ini memungkinkan Anda untuk mengukur seberapa sering suatu nilai muncul baik secara umum maupun dalam kondisi tertentu. Fungsi-fungsi berikut menggambarkan pendekatan utama dalam melakukan penghitungan. Gunakan kumpulan data (*dataset*) `*Maria's_data*` yang disertakan dalam buku kerja Excel bab ini untuk mengikuti langkah-langkahnya dan melakukan analisis yang sama.

1. $COUNT()$ — Menghitung Entri Numerik

Tujuan: Menghitung sel yang berisi nilai numerik.

Sintaks:

$$= COUNT(range)$$

Contoh:

$$= COUNT(F2{:}F51)$$

Jika kolom F berisi jumlah transaksi, hasilnya adalah: 50. Interpretasi: Terdapat 50 catatan transaksi numerik yang valid. Fungsi ini digunakan untuk memverifikasi kelengkapan data numerik.

2. $COUNTA()$ — Menghitung Entri yang Tidak Kosong

Tujuan: Menghitung semua sel yang tidak kosong, termasuk teks maupun angka.

Sintaks:

$$= COUNTA(range)$$

Contoh:

$$= COUNTA(E2{:}E51)$$

Hasil: 50, ini mengonfirmasi bahwa semua transaksi memiliki catatan metode pembayaran.

3. $COUNTBLANK()$ — Mendeteksi Data yang Hilang

Tujuan: Menghitung sel yang kosong.
Sintaks:

$$= COUNTBLANK(range)$$

Contoh:

$$= COUNTBLANK(E2{:}E51)$$

Jika hasilnya 0, berarti tidak ada entri pembayaran yang hilang. Fungsi ini membantu dalam pemeriksaan kualitas data.

4. $COUNTIF()$ — Penghitungan Bersyarat

Tujuan: Menghitung sel yang memenuhi satu kondisi tertentu.
Sintaks:

$$= COUNTIF(rentang, kriteria)$$

Contoh:
Menghitung pembayaran tunai:

$$= COUNTIF(E2{:}E51, "Cash") \rightarrow 18$$

Menghitung pembayaran dengan kartu:

$$= COUNTIF(E2{:}E51, "Credit/Debit\ Card") \rightarrow 27$$

Menghitung pembayaran melalui perangkat seluler (mobile):

$$= COUNTIF(E2{:}E51, "Mobile\ Payment") \rightarrow 5$$

Ini merupakan frekuensi bersyarat—dasar dari analisis kategorikal.

5. $COUNTIFS()$ — Berdasarkan Beberapa Kondisi

Tujuan: Menghitung observasi yang memenuhi beberapa kriteria secara bersamaan. Sintaks:

$$= COUNTIFS(rentang1, kriteria1, rentang2, kriteria2, \dots)$$

Contoh: Pembayaran dengan kartu di atas Rp 300.000

$$= COUNTIFS(E2{:}E51, "Credit/Debit\ Card", F2{:}F51, " > 30")$$

Fungsi ini memungkinkan analisis berbasis filter, yang mendekati metode pemeriksaan data secara terstruktur.

Penjumlahan dalam Excel

Penghitungan (*counting*) memberi tahu kita seberapa sering sesuatu terjadi. Penjumlahan (*summing*) memberi tahu kita berapa jumlah totalnya.

1. $SUM()$ — Menjumlahkan Nilai

Tujuan: Menjumlahkan semua nilai numerik dalam suatu rentang.

Sintaks:

$$= SUM(angka1, [angka2], \dots)$$

- angka1 bisa berupa satu angka, satu sel, atau rentang seperti F2:F51.
- Anda dapat menyertakan beberapa rentang jika diperlukan.

Contoh:

$$= SUM(F2{:}F51)$$

Hasil: Rp 12.475.200. Ini menunjukkan total pendapatan Maria untuk hari tersebut. Nilai total menjadi dasar untuk menghitung rata-rata, persentase, dan analisis pendapatan.

2. $SUMIF()$ — Penjumlahan Bersyarat

Tujuan: Menjumlahkan nilai hanya jika suatu kondisi terpenuhi.

Sintaks:

$$= SUMIF(rentang, kriteria, rentang_jumlah)$$

- rentang = sel yang diperiksa untuk menentukan kondisi
- kriteria = kondisi yang ditetapkan (teks, angka, perbandingan, dll.)
- rentang_jumlah = nilai yang akan dijumlahkan jika kriteria terpenuh

Contoh: Total pendapatan dari pembayaran dengan kartu

$$= SUMIF(E2{:}E51, "Credit/Debit\ Card", F2{:}F51)$$

Ini menjawab pertanyaan:

- Berapa besar pendapatan yang berasal dari transaksi kartu?

- ❖ Metode pembayaran mana yang menghasilkan pendapatan tertinggi?

Ini merupakan langkah awal dalam melakukan analisis keuangan berdasarkan kategori.

3. $SUMIFS()$ — Berdasarkan Beberapa Kondisi

Tujuan: Menjumlahkan nilai ketika beberapa kriteria terpenuhi sekaligus.
Sintaks:

$$= SUMIFS(rentang_jumlah, rentang_kriteria1, kriteria1, rentang_kriteria2, kriteria2, \dots)$$

Setiap rentang_kriteria harus memiliki ukuran yang sama dengan rentang_jumlah.
Contoh: Pembayaran dengan kartu di atas Rp 200.000

$$= SUMIFS(F2:F51, E2:E51, "Credit/Debit\ Card", F2:F51, ">20")$$

Hal ini memungkinkan analisis pendapatan yang lebih mendalam berdasarkan kriteria tertentu.

- ❖ "Berapa total pengeluaran nasabah yang melakukan transaksi kartu bernilai tinggi hari ini?"
- ❖ "Berapa porsi pendapatan dari kartu yang berasal dari transaksi bernilai besar?"

Mengapa Penghitungan dan Total Itu Penting

Penghitungan menghasilkan distribusi frekuensi. Total menghasilkan ukuran agregat. Penghitungan dan penjumlahan bersyarat menghasilkan perbandingan yang terstruktur. Operasi-operasi ini mungkin tampak sederhana, namun merupakan transformasi awal dari data mentah menjadi wawasan yang terorganisasi.

- ❖ Tanpa penghitungan, tidak ada proporsi.
- ❖ Tanpa total, tidak ada nilai rata-rata.
- ❖ Tanpa agregasi bersyarat, tidak ada perbandingan.

Semua metode statistik tingkat lanjut dibangun berdasarkan ringkasan-ringkasan mendasar ini. Sebelum memodelkan hubungan, kita harus terlebih dahulu memahami strukturnya. Penghitungan adalah titik awal dari pemahaman tersebut.

Proporsi dan Persentase

Penghitungan memberikan gambaran mengenai besaran, namun tidak memiliki makna komparatif. Pernyataan bahwa "18 pelanggan membayar dengan uang tunai" hanya menunjukkan ukuran satu kelompok. Pernyataan ini tidak memberi tahu kita apakah jumlah tersebut besar atau kecil jika dibandingkan dengan total keseluruhan. Agar dapat diinterpretasikan dengan bermakna, hasil penghitungan harus dinyatakan dalam istilah relatif. Di sinilah proporsi dan persentase menjadi sangat penting.

Proporsi

Proporsi merepresentasikan bagian dari total keseluruhan yang termasuk dalam kategori tertentu. Proporsi dihitung sebagai berikut:

$$\text{Proportion} = \frac{\text{Count of interest}}{\text{Total count}}$$

Proporsi selalu berada di antara 0 dan 1. Menggunakan data Maria:

$$\frac{18}{50} = 0.36$$

Interpretasi: 0,36 dari seluruh pelanggan membayar dengan uang tunai.
Proporsi mengubah frekuensi absolut menjadi frekuensi relatif.

Persentase

Persentase pada dasarnya adalah proporsi yang dikalikan dengan 100.

$$Persentase = Proporsi \times 100.$$

Menggunakan contoh Maria:

$$0.36 \times 100 = 36\%$$

Sekarang kita dapat menyatakan:
36% pelanggan membayar dengan uang tunai.
Persentase meningkatkan kemudahan interpretasi dengan menempatkan nilai pada skala 0–100 yang sudah familier.

- **Aturan Penting: Tentukan Referensinya**
  Persentase hanya bermakna jika kelompok referensinya jelas.
  ❖ "36% pelanggan membayar dengan uang tunai" — jelas.
  ❖ "36% membayar dengan uang tunai" — ambigu, kecuali jika populasinya diketahui.
  Selalu jelaskan apa yang menjadi acuan dari persentase tersebut.
- **Pemeriksaan Konsistensi Internal**
  Ketika kategori-kategori membentuk satu kesatuan utuh, jumlah persentasenya harus mencapai 100%. Distribusi pembayaran Maria:
  ❖ Tunai: 36%
  ❖ Kartu: 54%
  ❖ Seluler: 10%
  Total: 100%

Jika jumlah persentase tidak mencapai 100%, mungkin terdapat kesalahan pembulatan atau kesalahan perhitungan. Pemeriksaan sederhana ini mencegah pelaporan yang keliru.

Rasio dan Laju (Rate)

Sementara proporsi menggambarkan hubungan antara bagian dan keseluruhan, rasio membandingkan dua kuantitas apa pun.

Rasio

Rasio menyatakan ukuran relatif dari dua kuantitas melalui pembagian. Rasio tidak mengharuskan salah satu kuantitas menjadi bagian dari kuantitas lainnya. Rasio dapat ditulis sebagai:

- $A: B$
- A/B

Contoh: Jika Maria menjual 60 minuman panas dan 40 minuman dingin, maka rasio minuman panas terhadap minuman dingin adalah:

$$60: 40 \ = \ 3: 2$$

Interpretasi: Untuk setiap 3 minuman panas yang terjual, terjual pula 2 minuman dingin.
Berbeda dengan persentase, rasio membandingkan kategori secara langsung, bukan membandingkannya terhadap total keseluruhan.

Laju (*Rate*)

Laju adalah jenis rasio khusus yang menyertakan unit standardisasi, yang sering kali melibatkan waktu, ukuran populasi, atau variabel skala lainnya. Contoh umum meliputi:

- Kata per menit
- Penjualan per hari
- Angka kejadian per 100.000 populasi
- Konversi situs web per 100 pengunjung

Rasio (tingkat) memungkinkan perbandingan di berbagai konteks dengan skala yang berbeda.

Contoh dalam E-Commerce: Tingkat Konversi (*Conversion Rate*)

Sebuah situs web menerima 1.500 pengunjung dalam satu hari dan mencatat 45 pembelian.

$$\frac{45}{1{,}500} = 0.03 = 3\%$$

Interpretasi:
3% pengunjung melakukan pembelian.
Tingkat ini menstandardisasi kinerja relatif terhadap volume lalu lintas dan memungkinkan perbandingan antarhari atau antarkampanye.

Mengapa Ukuran Relatif Itu Penting

- Jumlah mentah menggambarkan skala.
- Proporsi menggambarkan struktur.
- Rasio menggambarkan perbandingan.
- Tingkat (*rate*) menggambarkan intensitas yang telah distandardisasi.

Dalam analisis statistik, ukuran relatif memungkinkan perbandingan yang adil di antara kelompok-kelompok dengan ukuran yang berbeda. Tanpa ukuran ini, kelompok yang besar akan selalu tampak dominan, bahkan ketika dampak relatifnya sebenarnya kecil.
Beralih dari sekadar jumlah ke proporsi menandai pergeseran konseptual:

- Kita tidak lagi hanya bertanya, “Berapa banyak?”
- Kita mulai bertanya, “Seberapa besar dibandingkan dengan apa?”
- Pertanyaan tersebut merupakan inti dari penalaran statistik.

3.3 DISTRIBUSI FREKUENSI UNTUK MERINGKAS POLA DATA

Bayangkan seorang guru memegang 100 nilai ujian yang tertulis pada lembaran kertas terpisah. Melihat angka-angka tersebut satu per satu—87, 92, 76, 88, 95, 73...—tidak memberikan banyak informasi mengenai performa kelas secara keseluruhan. Namun, begitu nilai-nilai tersebut dikelompokkan ke dalam kategori tertentu dan dihitung jumlahnya, sebuah pola pun muncul. Distribusi frekuensi menyusun pengamatan dengan menunjukkan seberapa sering setiap nilai atau rentang nilai muncul. Alih-alih memeriksa titik data secara individual, kita meringkas pola kemunculannya. Peralihan dari pengamatan mentah ke penghitungan yang terstruktur ini menandai langkah substansial pertama dalam analisis deskriptif.

Tabel Frekuensi untuk Data Kategorikal

Ketika variabel mewakili kategori (contoh: jenis minuman, metode pembayaran, merek produk), tabel frekuensi menyajikan ringkasan yang jelas. Contoh: Pesanan Pagi di Tempat Maria. Maria mencatat kategori minuman yang dipilih oleh 60 pelanggan selama giliran kerja pagi. Data mentah (sebagian): Espresso, Latte, Latte, Cappuccino, Espresso, Latte, Teh, Espresso, Latte… Daftar mentah tidak memberikan banyak informasi, sedangkan tabel yang terstruktur mengungkapkan segalanya.

Langkah-langkah Menyusun Tabel Frekuensi

1. Daftarkan semua kategori yang unik (Latte, Espresso, dll.).
2. Hitung jumlah kemunculannya (Frekuensi).
3. Hitung Frekuensi Relatif: (Frekuensi / Total).
4. Hitung Persentase: (Frekuensi Relatif × 100).

Tabel 3.1 Preferensi Minuman

Kategori Minuman	Frekuensi	Frekuensi Relatif	Persentase
Latte	25	0.42	42%
Espresso	18	0.30	30%
Cappuccino	12	0.20	20%
Teh	5	0.08	8%
TOTAL	**60**	**1.00**	**100%**

Apa yang sebelumnya berupa daftar yang tersebar kini menjadi ringkasan yang dapat diinterpretasikan. Wawasan: Ceritanya kini menjadi jelas. Latte dan Espresso mencakup 72% dari seluruh pesanan. Teh berada di posisi paling bawah dengan hanya 8%—Maria mungkin bisa mempertimbangkan untuk mengalihkan ruang rak tersebut bagi produk yang lebih populer.

Distribusi Frekuensi untuk Data Numerik Diskrit

Jika data terdiri dari nilai numerik yang dapat dihitung (contoh: jumlah anak, kunjungan per bulan, jumlah barang yang dibeli), setiap nilai yang berbeda dapat dicantumkan beserta frekuensinya.

Contoh: Ukuran Rumah Tangga.

Sebuah survei terhadap 200 rumah tangga mengajukan pertanyaan: "Berapa jumlah anak di bawah usia 18 tahun yang tinggal di rumah Anda?"

Tabel 3.2 Tabel frekuensi untuk respons survei

Jumlah Anak	Frekuensi	Frekuensi Relatif	Persentase	% Kumulatif
0	46	0.23	23%	23%
1	62	0.31	31%	54%
2	58	0.29	29%	83%
3	24	0.12	12%	95%
4 atau lebih	10	0.05	5%	100%
Total	200	1.00	100%	

Memahami Persentase Kumulatif

Persentase kumulatif menunjukkan proporsi observasi yang nilainya sama dengan atau di bawah nilai tertentu. Contohnya:

- 54% rumah tangga memiliki satu anak atau kurang.
- 83% rumah tangga memiliki dua anak atau kurang.

Persentase kumulatif membantu menjawab pertanyaan mengenai ambang batas dan posisi dalam suatu distribusi. Persentase ini sangat berguna dalam analisis kinerja serta interpretasi kebijakan.

Distribusi Frekuensi Berkelompok untuk Data Kontinu

Data kontinu (seperti tinggi badan, pendapatan, suhu, atau nilai ujian) dapat mencakup banyak nilai yang unik. Mencantumkan setiap nilai secara terpisah menjadi tidak praktis. Solusinya adalah dengan mengelompokkan observasi ke dalam interval kelas (sering disebut juga sebagai *bin*). Pendekatan ini juga dapat diterapkan pada data diskret yang memiliki banyak nilai berbeda.

Contoh: Nilai Ujian.

Misalkan terdapat 200 nilai ujian yang berkisar antara 45 hingga 98. Jika setiap nilai (45, 46, 47, …, 98) dicantumkan satu per satu, maka akan dihasilkan 54 baris, hal ini tentu tidak efisien dan tidak informatif. Sebagai gantinya, nilai-nilai tersebut dikelompokkan ke dalam interval-interval tertentu.

Panduan untuk Menyusun Interval Kelas

1. Gunakan sekitar 5–15 kelas.
2. Usahakan agar lebar setiap kelas sama.
3. Pastikan antar-kelas tidak saling tumpang tindih.
4. Pastikan setiap nilai masuk ke dalam tepat satu kelas.
5. Gunakan batasan yang intuitif dan diberi label dengan jelas.

Tabel 3.3 Distribusi Frekuensi Berkelompok

Jumlah Anak	Frekuensi	Frekuensi Relatif	Persentase	% Kumulatif
90–100	28	0.14	14%	14%
80–89	52	0.26	26%	40%
70–79	67	0.34	34%	74%
60–69	38	0.19	19%	93%
50–59	12	0.06	6%	99%
Below 50	3	0.01	1%	100%
Total	200	1.00	100%	

Interpretasi

- Distribusi berkelompok menunjukkan:
- Rentang yang paling umum adalah 70–79 (34%).
- 74% mahasiswa memperoleh nilai 70 atau lebih.
- Hanya 7% yang memperoleh nilai di bawah 60 (jika 60 adalah ambang batas kelulusan).
- 40% memperoleh nilai 80 atau lebih tinggi.

Tabel berkelompok memungkinkan penilaian cepat terhadap pemusatan data dan ambang batas kinerja.

Pertimbangan (*Trade-Off*)

Pengelompokan data mengurangi presisi. Kita mengetahui bahwa 28 mahasiswa memperoleh nilai antara 90–100, tetapi kita tidak mengetahui distribusi persisnya di dalam rentang tersebut. Hilangnya detail ini merupakan kompromi yang disengaja. Dengan mengorbankan presisi pada tingkat mikro, kita memperoleh kejelasan pada tingkat makro. Statistik sering kali melibatkan pertimbangan seperti ini.

Mengapa Distribusi Frekuensi Itu Penting

- Jumlah (*counts*) mengungkapkan besaran.
- Proporsi mengungkapkan struktur.
- Distribusi frekuensi mengungkapkan pola.

Distribusi frekuensi menjadi dasar bagi:

- Histogram
- Distribusi probabilitas
- Ukuran pemusatan (central tendency)
- Ukuran penyebaran (*dispersi*)

Sebelum menghitung rata-rata atau memodelkan hubungan, kita harus memahami terlebih dahulu bagaimana data didistribusikan. Distribusi frekuensi merupakan gambaran nyata pertama dari suatu kumpulan data.

3.4 TABEL DUA ARAH – MENJELAJAHI HUBUNGAN

Melampaui Variabel Tunggal

Sejauh ini, kita telah memeriksa satu variabel dalam satu waktu: preferensi minuman, metode pembayaran, distribusi nilai ujian. Analisis-analisis ini mendeskripsikan struktur dalam

satu dimensi. Namun, sebagian besar pertanyaan bermakna dalam statistik melibatkan hubungan:

- Apakah pelanggan yang lebih muda lebih menyukai jenis minuman yang berbeda dibandingkan pelanggan yang lebih tua?
- Apakah tingkat kepuasan terhadap Produk B lebih tinggi daripada Produk A?
- Apakah perilaku pembelian berbeda berdasarkan metode pembayaran?

Menjawab pertanyaan-pertanyaan semacam itu memerlukan pemeriksaan terhadap dua variabel secara bersamaan. Di sinilah tabel dua arah menjadi sangat penting.

Tabel Kontingensi (Tabulasi Silang)

Tabel kontingensi, yang juga disebut tabulasi silang (*cross-tabulation*), menampilkan distribusi frekuensi gabungan dari dua variabel kategorikal. Tabel ini menunjukkan bagaimana observasi didistribusikan di seluruh kombinasi kategori. Dalam istilah statistik:

- Sel-sel di bagian dalam mewakili frekuensi gabungan.
- Total baris dan kolom disebut sebagai total marginal.

Contoh: Kepuasan Pelanggan Berdasarkan Produk

Sebuah perusahaan menyurvei 248 pelanggan mengenai tiga produk untuk menentukan apakah tingkat kepuasan berbeda di antara produk-produk tersebut.

Tabel 3.4 Kepuasan berdasarkan produk – Tabel Frekuensi Mentah

Tingkat Kepuasan	Produk A	Produk B	Produk C	Jumlah Baris
Puas	45	62	38	145
Netral	23	18	27	68
Tidak Puas	12	8	15	35
Jumlah Kolom	80	88	80	248

Tabel ini menyajikan frekuensi gabungan (*joint frequencies*). Sekilas, Produk B tampak sebagai yang terbaik karena memiliki jumlah pelanggan yang puas paling banyak (62). Namun, jumlah mentah saja tidak cukup untuk melakukan perbandingan karena setiap produk memiliki ukuran sampel yang berbeda (masing-masing 80, 88, dan 80 pelanggan). Oleh karena itu, kita belum dapat menjawab pertanyaan utama kita: "Produk manakah yang memiliki tingkat kepuasan tertinggi?" Untuk melakukan perbandingan secara adil, kita harus menggunakan persentase bersyarat (*conditional percentages*).

Persentase dalam Tabel Kontingensi

Persentase dalam tabel dua arah (*two-way table*) dapat dihitung dengan tiga cara berbeda. Masing-masing cara menjawab pertanyaan analisis yang berbeda pula.

1. Persentase Kolom (Bersyarat pada Variabel Kolom)

Persentase kolom menunjukkan bagaimana kategori baris terdistribusi di dalam setiap kategori kolom.

$$\text{Column Percentage} = \frac{\text{Cell Count}}{\text{Column Total}} \times 100$$

Tabel 3.5 Kepuasan berdasarkan kode produk—beserta persentasenya

Tingkat Kepuasan	Produk A	Produk B	Produk C
Puas	56,3%	70,5%	47,5%
Netral	28,8%	20,5%	33,8%
Tidak Puas	15,0%	9,1%	18,8%
Total	100%	100%	100%

Interpretasi

Kini perbandingan tersebut menjadi bermakna:

- Produk B memiliki tingkat kepuasan tertinggi (70,5%).
- Produk C memiliki tingkat kepuasan terendah (47,5%).
- Produk A berada di antara keduanya.

Persentase kolom menjawab pertanyaan:

"Dalam setiap produk, bagaimana pelanggan terdistribusi di berbagai tingkat kepuasan?"

Ini disebut sebagai distribusi bersyarat (*conditional distributions*) karena distribusinya bergantung pada variabel kolom.

2. Persentase Baris (Bersyarat pada Variabel Baris)

Persentase baris menunjukkan bagaimana kategori kolom terdistribusi di dalam setiap kategori baris.

$$\text{Row Percentage} = \frac{\text{Cell Count}}{\text{Row Total}} \times 100$$

Tabel 3.6 Kepuasan berdasarkan produk – Persentase baris

Tingkat Kepuasan	Produk A	Produk B	Produk C	Total
Puas	31,0%	42,8%	26,2%	100%
Netral	33,8%	26,5%	39,7%	100%
Tidak Puas	34,3%	22,9%	42,9%	100%

Interpretasi

Persentase baris menjawab pertanyaan:

"Di antara pelanggan yang puas, bagaimana distribusi mereka berdasarkan produk?"

Perspektif ini berguna saat menganalisis komposisi kelompok kepuasan, alih-alih membandingkan kinerja produk.

3. Persentase Total

Persentase total menyatakan setiap sel sebagai proporsi dari jumlah keseluruhan (248 pelanggan).

$$\text{Total Percentage} = \frac{\text{Cell Count}}{\text{Grand Total}} \times 100$$

Persentase ini menggambarkan pola distribusi secara keseluruhan, namun kurang efektif untuk perbandingan langsung antar kelompok jika ukuran sampelnya berbeda.

Memilih Persentase yang Tepat

Persentase yang tepat bergantung pada pertanyaan pembelajaran.

- Membandingkan kinerja antar kelompok? → Gunakan persentase kolom (jika kelompok ditempatkan dalam kolom).
- Menganalisis komposisi kategori? → Gunakan persentase baris.
- Menggambarkan struktur keseluruhan? → Gunakan persentase total.

Pemilihan penyebut (*denominator*) menentukan narasi yang disampaikan oleh tabel tersebut.

Mengapa Tabel Dua Arah (*Two-Way Tables*) Penting?

Tabel dua arah membawa kita dari sekadar deskripsi menuju perbandingan. Tabel ini memungkinkan kita untuk:

- Memeriksa potensi hubungan antar variabel.
- Mengidentifikasi pola bersyarat (*conditional patterns*).
- Mendeteksi kesenjangan antar kelompok.
- Merumuskan hipotesis untuk pengujian lebih lanjut.

Sementara distribusi frekuensi menggambarkan satu variabel, tabel kontingensi memperkenalkan pemikiran relasional, sebuah langkah dasar menuju korelasi, regresi, dan statistik inferensial. Sebelum menguji hubungan secara formal, kita harus melihatnya terlebih dahulu. Tabel dua arah membuat hubungan tersebut terlihat jelas.

Implementasi dalam Excel

Di Excel, salah satu cara paling efisien untuk menyusun tabel kontingensi adalah dengan menggunakan PivotTable dan menyesuaikannya untuk menampilkan jumlah (*counts*) atau persentase berdasarkan baris/kolom. PivotTable mengotomatiskan seluruh proses pembuatan tabulasi silang, sehingga memudahkan perhitungan persentase bersyarat serta eksplorasi hubungan antar variabel kategorikal. Pada bagian berikutnya, yaitu mengenai PivotTable, kita akan mempraktikkan konsep-konsep ini.

3.5 ANALISIS DATA DENGAN PIVOTTABLE DI EXCEL

Memahami PivotTable

Bayangkan Anda bekerja dengan lembar kerja (*spreadsheet*) yang berisi 5.000 transaksi penjualan. Setiap baris mencakup:

- Tanggal
- Produk
- Tenaga Penjual (*Salesperson*)
- Wilayah
- Kuantitas
- Jumlah (Nilai Penjualan)

Seorang manajer bertanya:

- “Berapa total penjualan berdasarkan wilayah?”
- “Tenaga penjual mana yang memiliki nilai transaksi rata-rata tertinggi?”

- "Bagaimana kinerja Produk A di wilayah Timur Laut dibandingkan dengan wilayah Barat Daya?"

Melakukan penyaringan (*filtering*), pengurutan (*sorting*), dan perhitungan secara manual akan memakan waktu dan rentan terhadap kesalahan. PivotTable memungkinkan Anda menjawab semua pertanyaan tersebut dalam hitungan menit.

PivotTable adalah alat ringkasan interaktif yang secara otomatis mengelompokkan dan mengagregasi data. Fitur ini memungkinkan Anda untuk mengatur ulang ("*pivot*") struktur analisis Anda menukar baris dengan kolom, mengubah ukuran ringkasan, serta mengeksplorasi berbagai dimensi dari kumpulan data yang sama tanpa mengubah data aslinya. Tabel pivot (*pivot table*) tidak menciptakan data baru. Tabel ini mengungkapkan struktur yang sudah ada di dalam data tersebut.

Empat Komponen Utama Tabel Pivot

Setiap tabel pivot terdiri dari empat area:

1. **Baris (*Rows*)**: Kategori yang digunakan untuk pengelompokan utama (ditampilkan secara vertikal).
2. **Kolom (*Columns*)**: Dimensi pengelompokan tambahan (ditampilkan secara horizontal).
3. **Nilai (*Values*)**: Variabel numerik yang diringkas (contoh: jumlah, rata-rata, hitungan).
4. **Filter (*Opsional*)**: Kriteria yang digunakan untuk membatasi analisis pada subset data tertentu.

Memahami keempat area ini jauh lebih penting daripada sekadar menghafal langkah-langkah teknis dalam perangkat lunak. Begitu Anda memahami strukturnya, penggunaan tabel pivot apa pun akan terasa intuitif.

Membuat Tabel Pivot di Excel

Gunakan kumpulan data `*sales_transactions*` yang tersedia dalam buku kerja Excel di bab ini untuk mempraktikkan pembuatan tabel pivot. Setiap baris dalam kumpulan data tersebut mewakili satu transaksi. Gambar di bawah ini menampilkan lima catatan pertama.

	A	B	C	D	E	F
1	**Tanggal**	**Produk**	**Tenaga Penjual**	**Wilayah**	**Kuantitas**	**Jumlah**
2	01/01/2024 0:00	Produk C	Aisha	Central	10	885.48
3	01/01/2024 1:00	Produk D	Denis	Southwest	18	218.38
4	01/01/2024 2:00	Produk A	Leroy	Northwest	12	490.56
5	01/01/2024 3:00	Produk C	Maya	Southwest	16	688.89
6	01/01/2024 4:00	Produk C	Moris	Central	1	42.8

Gambar 3.1 sates_trawsactiows - 5 data pertama

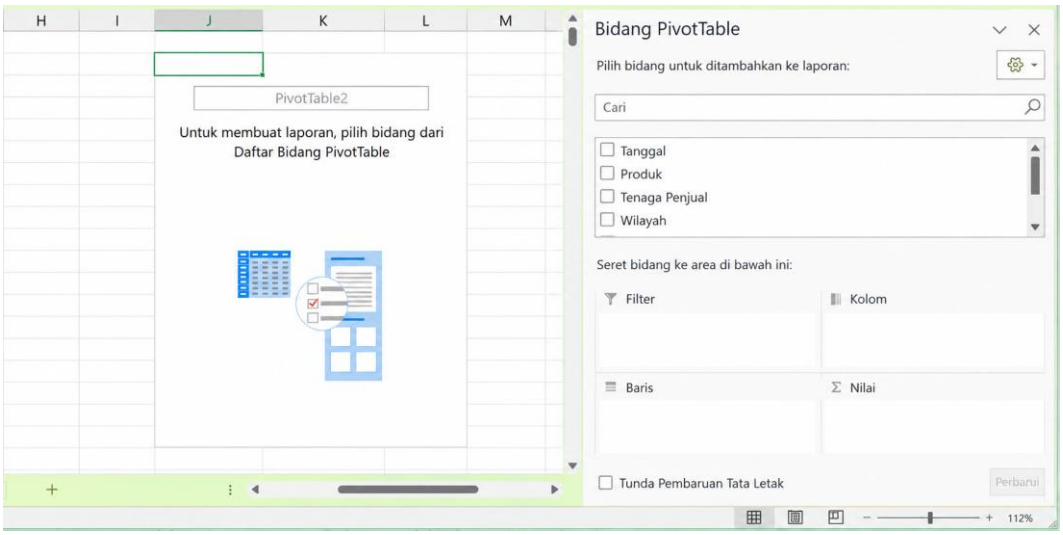
Petunjuk Pembuatan Pivot Table di Excel

Langkah demi Langkah:

- ❖ Pilih sel mana pun di dalam kumpulan data Anda.
- ❖ Buka menu Insert → PivotTable → From Table/Range.

- Konfirmasi rentang data yang terdeteksi.
- Tentukan apakah Pivot Table akan ditempatkan di lembar kerja (*worksheet*) baru atau di lembar kerja yang sudah ada. Jika di lembar kerja yang sama, pilih lokasi penempatannya.
- Klik OK.

Pivot Table baru akan terbuka dengan panel PivotTable Fields di sebelah kanan.



Gambar 3.2 Membuat tabel pivot (*pivot table*).

Seret (*drag*) kolom ke area yang sesuai:

- **Baris (*Rows*)**: Kategori untuk pengelompokan data
- **Kolom (*Columns*)**: Dimensi tambahan (*opsional*)
- **Nilai (*Values*)**: Data yang akan diringkas
- **Filter**: Kriteria untuk membatasi data (*opsional*)

Untuk mengubah cara suatu variabel diringkas:

- Klik kolom di bawah bagian Nilai (*Values*)
- Pilih Pengaturan Kolom Nilai (*Value Field Settings*)
- Pilih metode ringkasan seperti Jumlah (*Sum*), Cacah (*Count*), Rata-rata (*Average*), dll.
- Contoh 1: Total Penjualan berdasarkan Produk. Tujuan: Mengidentifikasi produk mana yang menghasilkan pendapatan terbesar. Pengaturan:
- Baris (*Rows*) → Produk
- Nilai (*Values*) → Jumlah dari Nominal (*Sum of Amount*)

Langkah ini akan menghasilkan total pendapatan untuk setiap produk.

Tabel 3.7 Analisis penjualan berdasarkan produk

Label Baris	**Jumlah Total**
Produk A	203,720.06
Produk B	319,386.23
Produk C	496,061.03

Produk D	733,452.25
Total Keseluruhan	1,752,619.57

Interpretasi

Misalkan Produk D menghasilkan Rp 7.334.520.000 dari total penjualan sebesar Rp 17.526.190.000. Kontribusinya adalah:

$$733.452 \div 1.752.619 = 41{,}8\%$$

Produk D menyumbang hampir 42% dari total pendapatan—jelas merupakan produk yang dominan. Pertanyaan untuk mendapatkan wawasan lebih lanjut:

- Apakah dominasi ini disebabkan oleh harga yang lebih tinggi?
- Volume yang lebih besar?
- Permintaan regional yang lebih kuat?

Tabel pivot (*pivot table*) menyediakan titik awal. Wawasan diperoleh dengan mengajukan pertanyaan lanjutan.

Contoh 2: Penjualan berdasarkan Wilayah dan Produk (Analisis Dua Dimensi)

Tujuan: Memeriksa apakah kinerja produk bervariasi di berbagai wilayah.

Pengaturan:

- Baris → Produk
- Kolom → Wilayah
- Nilai → Jumlah (*Sum of Amount*)

Ini menciptakan ringkasan dua arah yang serupa dengan tabel kontingensi, namun untuk data numerik.

Tabel 3.8 Penjualan berdasarkan wilayah dan produk

Jumlah Total	**Label Kolom**					
Label Baris	Tengah	Timur Laut	Barat laut	Tenggara	Barat daya	Total Keseluruhan
Produk A	41,421.01	46,273.19	37,467.60	40,526.50	38,031.76	203,720.06
Produk B	61,929.80	68,443.61	60,389.60	65,781.12	62,842.10	319,386.23
Produk C	117,025.18	86,681.45	96,450.83	99,746.23	96,157.34	496,061.03
Produk D	139,743.75	157,090.25	154,697.32	132,807.67	149,113.26	733,452.25
Total Keseluruhan	360,119.74	358,488.50	349,005.35	338,861.52	346,144.46	1,752,619.57

Interpretasi

Kini, pola-pola mulai terlihat:

- Produk D tampak mendominasi di semua wilayah.
- Produk A tampak secara konsisten menunjukkan kinerja yang kurang baik.
- Beberapa produk mungkin menunjukkan kinerja yang kuat hanya di wilayah tertentu saja.

Ringkasan multidimensi semacam ini mengubah data transaksi mentah menjadi wawasan strategis. Tanpa tabel pivot (*pivot table*), pola-pola seperti ini hampir mustahil untuk dideteksi secara efisien.

Mengubah Jenis Agregasi

Tabel pivot dapat meringkas nilai dengan berbagai cara:

- Sum (Jumlah) – Total pendapatan
- Count (Hitung) – Jumlah transaksi
- Average (Rata-rata) – Nilai transaksi rata-rata
- Minimum / Maximum – Nilai ekstrem (terendah/tertinggi)
- % of Grand Total (% dari Total Keseluruhan) – Kontribusi proporsional

Contoh: Rata-rata Nilai Transaksi berdasarkan Wilayah.

Pengaturan:

- Baris (*Rows*) → Wilayah
- Nilai (*Values*) → Rata-rata dari Jumlah (*Amount*)

Tabel 3.9 Rata-rata nilai transaksi berdasarkan wilayah

Label Baris	**Rata-rata Jumlah**
Tengah	350.31
Timur Laut	352.50
Barat Laut	349.01
Tenggara	346.13
Barat Daya	354.66
Total Keseluruhan	350.52

mengalihkan pertanyaan dari:

- *"Wilayah mana yang menghasilkan pendapatan terbesar?"*

menjadi:

- ✓ *"Wilayah mana yang memiliki transaksi dengan nilai tertinggi?"*

Ini adalah pertanyaan manajerial yang sangat berbeda.

- Total pendapatan mencerminkan volume.
- Nilai transaksi rata-rata mencerminkan perilaku pembelian.

Teknik Pivot Table yang Efektif

- Tambahkan beberapa kolom (*field*) ke bagian Baris (*Rows*) untuk membuat pengelompokan hierarkis.
- Jika Anda tidak lagi ingin menggunakan kolom tertentu, klik kolom tersebut → Hapus Kolom (*Remove Field*).

- ❖ Untuk melihat proporsi secara cepat, klik kanan pada nilai → Tampilkan Nilai Sebagai (*Show Values As*) → % dari Total Keseluruhan *(% of Grand Total*).
- ❖ Gunakan filter untuk berfokus pada periode waktu atau produk tertentu.
- ❖ Selalu segarkan (*Refresh*) pivot table setelah memperbarui data sumber.

Contoh: Proporsi Cepat.

Pada bagian sebelumnya, kita telah membahas persentase kolom, persentase baris, dan persentase total dalam tabel kontingensi. Sekarang, mari kita terapkan pengetahuan tersebut di Excel menggunakan pivot table. Pertanyaan bisnis: Produk manakah yang menghasilkan pendapatan terbesar di setiap wilayah, dan seberapa besar ketergantungan setiap wilayah terhadap masing-masing produk tersebut?

- **Pengaturan**
  - ❖ Baris (*Rows*) → Wilayah (*Region*)
  - ❖ Kolom (*Columns*) → Produk (*Product*)
  - ❖ Nilai (*Values*) → Jumlah dari Nilai (*Sum of Amount*)
- **Untuk menghitung persentase:**
  - ❖ Klik kanan pada nilai mana pun di dalam pivot table.
  - ❖ Pilih Tampilkan Nilai Sebagai (*Show Values As*).
  - ❖ Pilih % dari Total Baris *(% of Row Total*).

Perhatikan bahwa Anda juga dapat memilih % dari Total Kolom *(% of Column Total*) atau % dari Total Keseluruhan *(% of Grand Total*) untuk melihat distribusi proporsional yang berbeda.

Tabel 3.10 Penjualan berdasarkan wilayah dan produk

Jumlah Total	**Label Kolom**				
Label Baris	Produk A	Produk B	Produk C	Produk D	Total Keseluruhan
Tengah	11.50%	17.20%	32.50%	38.80%	100.00%
Timur Laut	12.91%	19.09%	24.18%	43.82%	100.00%
Barat Laut	10.74%	17.30%	27.64%	44.33%	100.00%
Tenggara	11.96%	19.41%	29.44%	39.19%	100.00%
Barat Daya	10.99%	18.15%	27.78%	43.08%	100.00%
Total Keseluruhan	11.62%	18.22%	28.30%	41.85%	100.00%

Interpretasi

Produk D merupakan penggerak pendapatan utama, dengan Produk C juga memberikan kontribusi yang signifikan. Di seluruh wilayah, Produk D menyumbang 39–44% dari total pendapatan. Baris "Total Keseluruhan" (*Grand Total*) menunjukkan kontribusi pendapatan berdasarkan produk, di mana Produk D menyumbang sekitar 42%.

Hal ini mengindikasikan bahwa:

- ❖ Bisnis sangat bergantung pada Produk D.
- ❖ Masalah apa pun yang terkait dengan Produk D seperti pasokan, penetapan harga, atau persaingan dapat berdampak serius terhadap pendapatan.

- ❖ Oleh karena itu, Produk D harus diprioritaskan dalam peramalan (*forecasting*), pengelolaan inventaris, dan pemasaran. Sebaliknya, Produk A merupakan penyumbang pendapatan yang lemah.

Mengapa Tabel Pivot (Pivot Table) Menjadi Fondasi Penting

Tabel pivot bukan sekadar fitur di Excel. Tabel ini mencerminkan cara berpikir tertentu:

- ❖ Mengelompokkan data
- ❖ Melakukan agregasi yang bermakna
- ❖ Membandingkan berbagai dimensi
- ❖ Merumuskan pertanyaan lanjutan yang lebih baik

Tabel pivot menjembatani kesenjangan antara data mentah dan analisis terstruktur. Sebelum statistik tingkat lanjut, sebelum regresi, dan sebelum pembelajaran mesin (*machine learning*) terdapat tahap peringkasan data. Tabel pivot adalah salah satu alat paling ampuh untuk melakukan hal tersebut dengan efektif.

3.6 ALUR KERJA PERINGKASAN DATA

Anda kini telah mengetahui cara menyusun:

- ❖ Tabel frekuensi
- ❖ Distribusi berkelompok
- ❖ Tabel kontingensi
- ❖ Tabel pivot

Namun, keterampilan analitis tidak ditentukan oleh seberapa banyak alat yang Anda kuasai, melainkan oleh pemahaman mengenai kapan dan mengapa setiap alat tersebut harus digunakan. Analisis data mengikuti alur kerja yang terstruktur. Tujuannya bukanlah sekadar menghasilkan tabel, melainkan menggali wawasan (*insight*).

1. Pahami Data Anda Sebelum Mengolahnya

Sebelum melakukan perhitungan apa pun, berhentilah sejenak dan periksa struktur kumpulan data Anda. Tanyakan:

- Jenis variabel apa saja yang ada?
  - ❖ Kategorikal (contoh: jenis produk, jenis kelamin, wilayah)
  - ❖ Numerik (contoh: pendapatan, usia, nilai penjualan)
- Berapa jumlah observasi yang tersedia?
- Pertanyaan apa yang ingin Anda jawab?

Langkah ini mencegah analisis yang dilakukan secara mekanis tanpa pemahaman. Tabel frekuensi yang dibuat untuk menjawab pertanyaan yang keliru tetaplah merupakan analisis yang salah. Kejelasan harus mendahului perhitungan.

2. Sesuaikan Metode dengan Pertanyaan

Pertanyaan yang berbeda memerlukan struktur peringkasan yang berbeda pula. Gunakan tabel di bawah ini sebagai panduan pengambilan keputusan:

Tabel 3.11 Berbagai struktur ringkasan – kerangka kerja pengambilan keputusan

Situasi	**Metode yang Direkomendasikan**	**Tujuan**
Satu variabel kategorikal	Tabel frekuensi dengan persentase	Memahami distribusi di berbagai kategori
Satu variabel kontinu	Distribusi frekuensi berkelompok	Mengidentifikasi pola dan sebaran secara keseluruhan
Dua variabel kategorikal	Tabel kontingensi (dengan persentase)	Mengeksplorasi hubungan
Satu variabel kategorikal + satu variabel numerik	Tabel pivot (rerata, total)	Membandingkan hasil numerik antar kelompok
Berbagai dimensi	Tabel pivot bertingkat	Mengeksplorasi pola yang kompleks

Perhatikan polanya:

- Jika pertanyaan berbunyi, “Berapa banyak?” (*How many?*), pikirkan tentang jumlah (hitungan).
- Jika pertanyaan berbunyi, “Seberapa banyak?” (*How much?*), pikirkan tentang total atau rata-rata.
- Jika pertanyaan berbunyi, “Apakah terdapat perbedaan antar kelompok?”, pikirkan tentang tabel perbandingan.

Metode mengikuti pertanyaan—bukan sebaliknya.

3. Lakukan Perhitungan dengan Konteks

Saat menyusun ringkasan:

- Sertakan total.
- Sertakan persentase saat membandingkan kelompok.
- Pastikan kategori bersifat saling lepas (*mutually exclusive*) dan mencakup keseluruhan (*collectively exhaustive*).
- Periksa apakah total sudah dihitung dengan benar (pengendalian kualitas itu penting).

Konteks mengubah angka menjadi makna.

Sebagai contoh, pernyataan:

> *“45 pelanggan merasa tidak puas.”*

tidak banyak berarti jika tidak diketahui apakah jumlah total pelanggannya 50 atau 5.000. Ukuran relatif (persentase, rata-rata) memungkinkan perbandingan. Ukuran absolut (hitungan, total) memberikan gambaran skala. Ringkasan yang baik sering kali mencakup keduanya.

4. Lakukan Interpretasi, Jangan Sekadar Melaporkan

Di sinilah analisis berubah menjadi wawasan (*insight*). Tanyakan:

- ❖ Pola apa yang paling menonjol?
- ❖ Apa yang tergolong sangat tinggi atau sangat rendah?
- ❖ Apakah terdapat perbedaan yang berarti antar kelompok?

- ❖ Apakah ada anomali yang memerlukan penyelidikan lebih lanjut?

Hindari sekadar mendeskripsikan isi tabel.
Interpretasi yang lemah:

> *"Produk B memiliki 62 pelanggan yang puas."*

Interpretasi yang lebih kuat:

> *"Produk B mencatat tingkat kepuasan tertinggi dibandingkan dengan basis pelanggannya, yang menunjukkan persepsi nilai yang lebih kuat dibandingkan Produk A dan C."*

Interpretasi menghubungkan angka dengan pengambilan keputusan.

5. Berkomunikasi secara Jelas dan Strategis

Tabel yang disusun dengan baik hanya akan berguna jika mudah dipahami. Saat menyajikan temuan:

- ❖ Gunakan judul yang jelas dan deskriptif.
- ❖ Beri label pada baris dan kolom secara tepat.
- ❖ Bulatkan persentase secara konsisten.
- ❖ Tulis kesimpulan dengan bahasa yang lugas dan langsung.

Nantinya, Anda dapat menyempurnakan ringkasan ini dengan visualisasi (lihat Bab 6).

Gambaran Besar

Proses dari data mentah menuju wawasan mengikuti alur yang terstruktur:

1. Pahami strukturnya.
2. Pilih metode peringkasan yang tepat.
3. Lakukan perhitungan dengan cermat.
4. Lakukan interpretasi secara kritis.
5. Berkomunikasilah dengan jelas.

Langkah-langkah ini mungkin tampak sederhana, namun merupakan fondasi dari semua analisis statistik.

- o Sebelum melakukan regresi.
- o Sebelum pengujian hipotesis.
- o Sebelum pemodelan prediktif.
- o Tahap awalnya adalah penyusunan ringkasan secara terstruktur.

Jika Anda menguasai alur kerja ini, Anda tidak akan pernah lagi merasa kewalahan menghadapi data mentah. Anda akan tahu persis dari mana harus memulai dan bagaimana melangkah secara sistematis untuk memperoleh wawasan.

3.7 STATISTIK SDM UNTUK MENGIDENTIFIKASI POLA DAN KETIMPANGAN ORGANISASI

Anda baru saja ditunjuk sebagai manajer SDM di sebuah perusahaan berskala menengah. Pada hari pertama, Anda menerima basis data karyawan yang berisi 200 catatan. Setiap catatan mencakup:

- Departemen
- Tingkat Jabatan
- Masa Kerja
- Gaji Tahunan
- Penilaian Kinerja

Tugas Anda sederhana namun penting:

Memahami tenaga kerja. Mengidentifikasi pola. Mendeteksi potensi masalah.

Anda tidak memulainya dengan model-model canggih, melainkan dengan ringkasan yang terstruktur.

1. Analisis 1: Distribusi Departemen

Pertanyaan: Bagaimana distribusi karyawan di berbagai departemen?

Metode: Tabel frekuensi beserta persentasenya.

Tabel frekuensi menunjukkan ukuran setiap departemen relatif terhadap keseluruhan organisasi.

Tabel 3.12 Distribusi karyawan berdasarkan departemen

Label Baris	Jumlah Departemen	Jumlah Departemen
Teknik	52	26.0%
Keuangan	17	8.5%
Pemasaran	28	14.0%
Operasional	45	22.5%
Penjualan	58	29.0%
Total Keseluruhan	200	100.0%

Wawasan

Departemen Penjualan dan Teknik muncul sebagai departemen terbesar, sementara departemen Keuangan relatif kecil. Hal ini segera memunculkan pertanyaan manajerial:

- Apakah departemen Keuangan kekurangan staf dibandingkan dengan operasional perusahaan?
- Apakah ukuran tersebut mencerminkan efisiensi akibat otomatisasi?
- Atau justru departemen tersebut kewalahan karena beban kerja yang terlalu berat bagi jumlah staf yang ada?

Bahkan tabel frekuensi sederhana pun dapat memicu refleksi strategis.

2. Analisis 2: Distribusi Gaji

Pertanyaan: Bagaimana kisaran gaji secara keseluruhan, dan di kelompok mana sebagian besar karyawan berada?

Metode: Distribusi frekuensi berkelompok untuk gaji tahunan.

Gaji dikelompokkan ke dalam interval-interval yang bermakna, sehingga pola-pola tertentu dapat terlihat.

Tabel 3.13 Distribusi gaji

Rentang Gaji	Frekuensi	Persentase	Persentase Kumulatif
Di bawah Rp 400.000.000	15	7,5%	7,5%
Rp 400.000.000–Rp 549.990.000	71	35,5%	43%
Rp 550.000.000–Rp 699.990.000	62	31%	74%
Rp 700.000.000–Rp 849.990.000	32	16%	90%
Rp 850.000.000–Rp 999.990.000	20	10%	100%
Rp 1.000.000.000 ke atas	0	0	100%
Total	200	100%	

Gaji minimum: 31.162,59 Gaji maksimum: 99.630,71

Wawasan

- Kisaran gaji yang paling umum adalah Rp 400.000.000–Rp 549.990.000, mencakup 35,5% dari seluruh karyawan.
- 74% dari tenaga kerja berpenghasilan kurang dari Rp 700.000.000.
- 7,5% karyawan berpenghasilan di bawah Rp 400.000.000.

Hal ini menunjukkan bahwa tenaga kerja terkonsentrasi pada rentang kompensasi tingkat menengah, dengan relatif sedikit yang berpenghasilan tinggi.

Implikasi manajerial:

- Apakah kompensasi selaras dengan tingkat pekerjaan?
- Apakah jalur promosi terlihat dalam struktur gaji?

Ringkasan yang dikelompokkan memberikan pemahaman struktural sebelum analisis statistik yang lebih mendalam.

3. Analisis 3: Kinerja per Departemen

Pertanyaan: Apakah peringkat kinerja berbeda antar departemen?

Metode: Tabel kontingensi dengan persentase kolom.

Karena ukuran departemen bervariasi, persentase kolom digunakan untuk membandingkan distribusi kinerja secara adil.

Tabel 3.14 kinerja berdasarkan departemen

Peringkat Kinerja	Teknik	Keuangan	Pemasaran	Operasional	Penjualan
Sangat Baik	26,92%	29,4%	17,86%	15,56%	12,07%
Baik	61,54%	29,4%	53,57%	57,78%	60,34%
Perlu Peningkatan	11.54%	41.2%	28.57%	26.67%	27.59%
Total Keseluruhan	100%	100%	100%	100%	100%

Wawasan

Departemen Keuangan menunjukkan proporsi tertinggi karyawan dengan kinerja "Sangat Baik" (29%). Namun, departemen ini juga memiliki proporsi karyawan yang dinilai "Perlu Perbaikan" yang luar biasa tinggi (41%). Polarisasi ini menunjukkan kemungkinan:

- ❖ Standar manajemen yang tidak konsisten
- ❖ Bias evaluasi kinerja
- ❖ Ketidakstabilan departemen
- ❖ Distribusi beban kerja yang tidak merata

Kuncinya bukan hanya mengidentifikasi persentase tetapi mengenali pola yang tidak biasa.
Membuat Analisis Ini Menggunakan Tabel Pivot (Excel)
Dataset → employee_database

- ❖ Baris: Peringkat Kinerja
- ❖ Kolom: Departemen
- ❖ Nilai: Jumlah kolom lain (misalnya, jumlah Peringkat Kinerja)

Klik kanan nilai apa pun di tabel pivot, buka 'Tampilkan Nilai Sebagai,' dan pilih '% dari Total Kolom.' Ini mengubah jumlah mentah menjadi proporsi yang dapat dibandingkan.

4. Analisis 4: Gaji Rata-Rata menurut Departemen dan Tingkat Jabatan

Pertanyaan: Apakah ada perbedaan gaji antar departemen pada tingkat jabatan yang serupa? Metode: Tabel pivot dengan gaji rata-rata. Pengaturan:

- ❖ Baris → Tingkat Jabatan
- ❖ Kolom → Departemen
- ❖ Nilai → Rata-Rata Gaji Tahunan

Tabel 3.15 Gaji Rata-Rata menurut Departemen

Tingkat Pekerjaan	Teknik	Keuangan	Pemasaran	Operasional	Penjualan
Pemula	49203,12	44985,17	45393.17	40721,83	42846,03
Tingkat Menengah	69351,54	71503,74	63698.96	59870	62210,52
Senior	94062,55	99630,71	88833.78	77963,53	85989,04
Total Keseluruhan	64753,15	60678,94	59688.2	52944,58	59193,42

Wawasan: Bidang Teknik dan Keuangan secara konsisten menawarkan gaji yang lebih tinggi untuk tingkat pekerjaan yang setara. Karyawan di bagian Operasional memperoleh penghasilan yang jauh lebih rendah dibandingkan rekan-rekan mereka di tingkat yang sama. Pola ini mengindikasikan potensi masalah kesetaraan:

- ❖ Apakah struktur kompensasi sudah selaras dengan standar pasar?
- ❖ Apakah bidang Operasional dinilai terlalu rendah dibandingkan dengan kontribusinya?
- ❖ Mungkinkah kesenjangan gaji ini menyebabkan tingkat pergantian karyawan (*turnover*) yang lebih tinggi?

Perbandingan rata-rata antar kategori mengungkapkan pola sistemik yang tidak dapat terlihat hanya dari daftar gaji mentah.

Hal-hal yang Belum Dapat Dijelaskan oleh Tabel-Tabel Ini

Meskipun ringkasan ini memberikan wawasan yang berarti, masih ada pertanyaan penting yang belum terjawab:

- Berapa nilai median gaji, dan bagaimana perbandingannya dengan nilai rata-rata (*mean*)?
- Seberapa besar variasi yang ada di dalam departemen?
- Apakah gaji terkonsentrasi pada satu titik atau tersebar luas?
- Apakah distribusi kinerja bersifat simetris atau condong (*skewed*)?
- Apakah terdapat nilai ekstrem (*outlier*) yang memengaruhi rata-rata?

Pertanyaan-pertanyaan ini memerlukan perangkat statistik yang lebih mendalam seperti ukuran tendensi sentral, variabilitas, dan visualisasi yang akan kita bahas pada Bab 4, 5, dan 6.

Pelajaran yang Lebih Luas

Perhatikan alur perkembangannya:

1. Mulai dari struktur (frekuensi).
2. Beralih ke distribusi (ringkasan yang dikelompokkan).
3. Mengeksplorasi hubungan (tabel kontingensi).
4. Membandingkan hasil antar kelompok (rata-rata melalui tabel pivot).

Tidak diperlukan matematika tingkat lanjut.

Namun, hanya dalam satu hari, manajer SDM yang baru telah berhasil:

- Mengidentifikasi potensi ketidakseimbangan jumlah staf.
- Mengamati pola konsentrasi gaji.
- Mendeteksi distribusi kinerja yang tidak wajar.
- Menandai kemungkinan masalah kesetaraan kompensasi. Inilah kekuatan dari peringkasan data yang terstruktur.

3.8 RINGKASAN DATA SEBAGAI FONDASI PENALARAN STATISTIK

Dalam bab ini, Anda beralih dari sekadar pengamat pasif data mentah menjadi pengelola informasi yang aktif. Melalui dilema kedai kopi Maria dan hari pertama seorang manajer SDM, Anda melihat kebenaran mendasar dalam analisis data: Data mentah tidak menyampaikan informasi; strukturlah yang menyampaikannya. Catatan transaksi, lembar kerja berisi data karyawan, atau daftar nilai ujian mungkin memuat jawaban tetapi tanpa pengorganisasian, jawaban tersebut akan tetap tersembunyi. Peran seorang analis data adalah mengubah pengamatan yang tersebar menjadi ringkasan yang bermakna. Kita memulainya dengan operasi paling mendasar dalam statistik: menghitung (*counting*).

Penghitungan bukanlah hal sepele. Penghitungan adalah fondasi dari semua penalaran statistik. Dengan menyusun data ke dalam tabel frekuensi, Anda telah mempelajari cara untuk:

- meringkas variabel kategorikal;
- mengubah jumlah mentah menjadi frekuensi relatif;
- menafsirkan persentase dalam konteks tertentu.

Anda melihat bagaimana sebuah tabel sederhana dapat mengungkap produk yang dominan, kategori yang populer, serta segmen yang terabaikan. Jika dilakukan dengan cermat, penghitungan dapat menghasilkan wawasan.

Selanjutnya, Anda mengeksplorasi distribusi frekuensi untuk variabel diskret maupun kontinu. Untuk data numerik diskret, mencantumkan nilai beserta frekuensinya akan memperlihatkan struktur data tersebut. Untuk data kontinu, pengelompokan nilai ke dalam interval kelas mengubah detail yang terlalu rumit menjadi pola yang dapat terlihat dengan jelas. Anda mempelajari bahwa:

- Interval kelas harus bersifat saling lepas (*mutually exclusive*) dan mencakup seluruh kemungkinan (*collectively exhaustive*).
- Persentase kumulatif memungkinkan Anda untuk membuat pernyataan mengenai distribusi.
- Pengelompokan memang mengurangi presisi, namun meningkatkan kejelasan.

Yang terpenting, Anda mulai mengenali bentuk data secara keseluruhan bukan sekadar nilai-nilai individualnya saja.

Statistik menjadi jauh lebih kuat ketika kita mengkaji hubungan antarvariabel. Melalui tabel dua arah (tabel kontingensi), Anda melangkah lebih jauh dari sekadar ringkasan satu variabel dan mulai membandingkan berbagai kategori secara bersamaan. Anda belajar untuk:

- membedakan antara persentase baris, kolom, dan total;
- memilih jenis persentase yang sesuai dengan pertanyaan pembelajaran;
- menafsirkan perbedaan proporsional alih-alih sekadar melihat jumlah mentah.

Hal ini menandai pergeseran pola pikir: dari sekadar mendeskripsikan distribusi menjadi menyelidiki keterkaitan antarvariabel.

Pivot table (tabel pivot) memperluas logika ini lebih jauh lagi. Anda melihat bagaimana satu kumpulan data dapat ditinjau dari berbagai sudut pandang:

- Total berdasarkan produk;
- Rata-rata berdasarkan wilayah;
- Jumlah berdasarkan tingkat kinerja;
- Ringkasan bertingkat yang mencakup beberapa dimensi sekaligus.

Pivot table menunjukkan bahwa analisis bukanlah proses yang linier, melainkan proses yang bersifat iteratif (berulang) dan eksploratif. Dengan menyusun ulang struktur data yang sama, pola-pola baru pun akan muncul.

Kekuatan analisis tidak terletak pada perangkat lunaknya, melainkan pada pertanyaan-pertanyaan yang Anda ajukan. Anda juga mempelajari kerangka kerja analitis yang disiplin:

1. Memahami struktur data Anda.
2. Menyesuaikan metode dengan pertanyaan yang ingin dijawab.
3. Menghitung ringkasan data dengan cermat.
4. Menafsirkan hasil secara kritis.
5. Mengomunikasikan temuan dengan jelas.

Alur kerja ini mengubah analisis dari sekadar perhitungan mekanis menjadi penalaran yang terstruktur. Teknik-teknik yang Anda pelajari dalam bab ini bukanlah metode statistik tingkat

lanjut, melainkan metode-metode yang menjadi fondasi dasar. Setiap model yang canggih, setiap pengujian hipotesis, dan setiap algoritma prediktif bertumpu pada kemampuan untuk meringkas serta memahami data secara jelas.

BAB 4
TENDENSI SENTRAL DAN VARIABILITAS

4.1 TENDENSI SENTRAL DAN VARIABILITAS DALAM ANALISIS STATISTIK DESKRIPTIF

Pada Bab 3, kita telah mempelajari cara mengorganisasi data. Melalui penghitungan, persentase, tabel frekuensi, tabel kontingensi, dan tabel pivot, kita mengubah observasi mentah menjadi ringkasan yang terstruktur. Kita menjawab pertanyaan mengenai komposisi, distribusi di berbagai kategori, serta hubungan antarvariabel. Namun, struktur saja tidaklah cukup. Setelah informasi terorganisasi, muncul dua pertanyaan analitis mendasar:

1. **Di manakah pusatnya?**
   Nilai apa yang paling mewakili kumpulan data tersebut hasil yang tipikal, diharapkan, atau paling sering muncul? Konsep ini dikenal sebagai tendensi sentral (*central tendency*).
2. **Seberapa besar variasi yang ada?**
   Apakah observasi mengelompok rapat di sekitar pusat tersebut, atau justru tersebar luas? Sifat ini dikenal sebagai variabilitas (atau *dispersi*).

Kedua konsep ini pusat dan sebaran membentuk fondasi analisis data numerik. Tendensi sentral mengidentifikasi apa yang dianggap tipikal. Variabilitas mengukur seberapa stabil atau tidak stabilnya nilai tipikal tersebut. Keduanya secara bersama-sama memberikan gambaran deskriptif yang lebih baik.

Mengapa Pusat Saja Tidak Cukup

Bayangkan Anda sedang menganalisis gaji tahunan 500 karyawan di sebuah perusahaan teknologi. Manajer Anda bertanya:

"Berapa gaji yang kita berikan kepada karyawan kita?"

Anda tidak bisa menjawab dengan menyebutkan 500 angka sekaligus. Anda memerlukan sebuah ringkasan misalnya: "Gaji rata-ratanya adalah Rp 800.000.000." Namun, bagaimana jika CEO perusahaan tersebut memperoleh gaji Rp 100.000.000, sementara sebagian besar karyawan hanya menerima Rp 300.000.000? Rata-rata yang dihitung mungkin benar secara matematis, tetapi sangat menyesatkan. Angka tersebut gagal merepresentasikan kondisi karyawan pada umumnya.

Contoh ini mengungkapkan prinsip penting: Ukuran pusat tanpa pemahaman mengenai sebaran dapat mendistorsi realitas. Bahkan jika kita telah menentukan nilai pusat, kita tetap harus bertanya:

- Apakah sebagian besar observasi mengelompok di dekat nilai tersebut?
- Atau, apakah variasi ekstrem membuat nilai pusat tersebut menjadi tidak dapat diandalkan?

Kredibilitas nilai pusat bergantung pada variabilitas di sekitarnya.

Dua Pilar Statistik Deskriptif

Tendensi sentral dan variabilitas bukanlah gagasan yang berdiri sendiri. Keduanya saling melengkapi.

- Tendensi sentral menjawab: Di mana data memusat?
- Variabilitas menjawab: Seberapa konsisten pemusatan tersebut?

Setiap bidang yang bermakna bergantung pada sepasang konsep ini:

- **Kesehatan**: Berapa lama waktu pemulihan yang lazim, dan seberapa besar variasinya di antara pasien?
- **Keuangan**: Berapa imbal hasil rata-rata, dan seberapa fluktuatif imbal hasil tersebut?
- **Pemasaran**: Berapa peringkat rata-rata dari pelanggan, dan seberapa konsisten pengalaman pelanggan tersebut?
- **Pendidikan**: Berapa skor tes yang lazim, dan seberapa lebar kesenjangan kinerjanya?

Konteksnya mungkin berubah, tetapi logikanya tetap sama. Tidak ada analisis kuantitatif yang serius dapat dilakukan tanpa mengestimasi nilai pusat maupun sebarannya.

Apa yang Akan Anda Pelajari dalam Bab Ini

Dalam bab ini, kita akan mengkaji ukuran-ukuran formal untuk:

- **Tendensi Sentral (Pusat Data)**
  - Rata-rata (*Mean*)
  - Median
  - Modus
- **Variabilitas (Sebaran Data)**
  - Rentang (*Range*)
  - Rentang Interkuartil (*Interquartile Range*/IQR)
  - Varians
  - Simpangan Baku (*Standard Deviation*)
  - Koefisien Variasi

Ukuran-ukuran ini memberikan koordinat bagi suatu kumpulan data yaitu pusat dan sebarannya. Ukuran-ukuran tersebut memungkinkan kita untuk membandingkan kelompok, mendeteksi ketidakseimbangan, mengevaluasi stabilitas, serta mempersiapkan diri untuk pemodelan statistik yang lebih lanjut.

Perspektif Kritis

"Rata-rata" mungkin merupakan statistik yang paling dikenal di dunia. Istilah ini muncul dalam laporan berita, debat kebijakan, ringkasan keuangan, dan dasbor bisnis. Namun, statistik ini juga merupakan salah satu yang paling sering disalahpahami dan disalahgunakan.Di akhir bab ini, Anda akan memahami:

- Kapan nilai rata-rata (*mean*) memberikan informasi yang berguna.
- Kapan median lebih dapat diandalkan.
- Mengapa variabilitas menentukan apakah suatu nilai rata-rata bermakna atau tidak.
- Dan mengapa menyajikan satu angka saja tanpa menyertakan sebarannya dianggap tidak lengkap secara statistik.

Singkatnya, Anda akan beralih dari sekadar menghitung nilai rata-rata menjadi menafsirkannya secara bertanggung jawab. Sebab dalam statistik, mengetahui nilai pusat hanyalah permulaan.

4.2 UKURAN TENDENSI SENTRAL: MENEMUKAN INTI DARI DATA ANDA

Ukuran tendensi sentral menjawab pertanyaan yang tampak sederhana: Apa yang dianggap tipikal dalam kumpulan data ini? Ukuran-ukuran ini meringkas keseluruhan distribusi ke dalam satu nilai representatif. Penyederhanaan ini sangat berguna, namun bisa menjadi berbahaya jika digunakan tanpa kecermatan. Pada bagian ini, kita akan membahas tiga ukuran tendensi sentral utama:

- *Mean* (Rata-rata)
- *Median* (Nilai Tengah)
- *Modus* (Nilai yang Paling Sering Muncul)

Masing-masing ukuran ini menangkap konsep "pusat" dengan cara yang berbeda. Seorang analis yang terampil tidak hanya memahami cara menghitungnya, tetapi juga mengetahui kapan masing-masing ukuran tersebut memberikan gambaran yang paling akurat.

Mean (Rata-rata): Titik Keseimbangan Matematis

Mean (sering disebut sebagai rata-rata) merupakan titik keseimbangan matematis dari suatu kumpulan data. Nilai ini dihitung dengan menjumlahkan seluruh nilai yang ada, kemudian membaginya dengan jumlah observasi.

Rumus (*Mean Sampel*)

$$\bar{x} = \frac{\sum x_i}{n}$$

Di mana:

- $\bar{x}$ = rata-rata sampel
- $\sum x_i$ = jumlah dari semua nilai yang diamati
- n = banyaknya pengamatan

Rumus (Rata-rata Populasi)

$$\mu = \frac{\sum x_i}{N}$$

Di mana:

- μ = rata-rata populasi
- N = jumlah total observasi dalam populasi

Kapan Menggunakan Rata-rata (*Mean*)

Rata-rata paling tepat digunakan:

- Untuk data numerik
- Ketika distribusi data cenderung simetris
- Ketika Anda ingin setiap nilai memengaruhi hasil akhir
- Dalam pemodelan dan inferensi statistik formal

Karena mencakup setiap observasi, nilai rata-rata menjadi sensitif terhadap nilai-nilai ekstrem.

Contoh: Pengeluaran Pemasaran Bulanan.

Misalkan Anda menganalisis anggaran pemasaran bulanan dari 12 usaha kecil:

Bisnis 1	2.400
Bisnis 2	3.100
Bisnis 3	2.800
Bisnis 4	3.500
Bisnis 5	2.900
Bisnis 6	3.200
Bisnis 7	2.700
Bisnis 8	3.300
Bisnis 9	2.600
Bisnis 10	3.400
Bisnis 11	2.500
Bisnis 12	15.000

Total pengeluaran pemasaran:

$$47.400$$

Rata-rata:

$$47.400 \div 12 = 3.950$$

Interpretasi

Nilai rata-rata (*mean*) menunjukkan bahwa perusahaan menghabiskan sekitar Rp 39.500.000 per bulan untuk pemasaran. Namun, ada satu perusahaan yang menghabiskan Rp 150.000.000, angka yang jauh lebih tinggi dibandingkan perusahaan lainnya. Nilai ekstrem tunggal ini menarik nilai rata-rata ke atas, sehingga anggaran pemasaran yang "tipikal" tampak lebih besar daripada kondisi yang sebenarnya bagi sebagian besar perusahaan. Nilai rata-rata tersebut secara matematis tepat, namun belum tentu representatif.

Implementasi di Excel

Fungsi:

$$= AVERAGE\ (number1, [number2], \dots)$$

Contoh (untuk data di sel B1:B12):

$$= AVERAGE(B1{:}\,B12)$$

Median: Posisi Tengah

Median merupakan nilai tengah dari suatu kumpulan data setelah data tersebut diurutkan dari yang terkecil hingga yang terbesar. Berbeda dengan nilai rata-rata (*mean*), median bergantung pada posisi, bukan pada besaran nilai.

Menentukan Median

- Urutkan data dari yang terkecil hingga terbesar (urutan menaik).
- Jika n ganjil → Median adalah nilai pada posisi $(n + 1)/2$.
- Jika n genap → Median adalah rata-rata dari nilai-nilai pada posisi:

$$\frac{n}{2} \quad \text{dan} \quad \binom{n}{2} + 1$$

Contoh (Data yang telah diurutkan)

2.400
2.500
2.600
2.700
2.800
2.900
3.100
3.200
3.300
3.400
3.500
15.000

Dengan 12 observasi (jumlah genap):

Posisi tengah:

Nilai ke-6 = 2.900
Nilai ke-7 = 3.100

Median:

$$(2.900 + 3.100) \div 2 = 3.000$$

Interpretasi

Setengah dari bisnis yang ada mengeluarkan biaya kurang dari Rp 30.000.000, dan setengahnya lagi mengeluarkan biaya lebih dari Rp 30.000.000. Berbeda dengan nilai rata-rata (*mean*), median tidak terdistorsi oleh nilai pencilan (*outlier*) sebesar Rp 150.000.000. Median memberikan ukuran yang lebih tangguh (*robust*) untuk menggambarkan pengeluaran tipikal dalam data yang menceng (*skewed*).

Kapan Menggunakan Median

- Saat data menceng (*skewed*)
- Saat terdapat nilai pencilan (*outlier*)

- Untuk data ordinal (contoh: peringkat, tingkat kepuasan)
- Saat Anda memerlukan ukuran pemusatan yang tahan terhadap nilai ekstrem

Implementasi di Excel

Fungsi

$$= MEDIAN(number1, [number2], \dots)$$

Contoh:

$$= MEDIAN(B1{:}B12)$$

Modus: Nilai yang Paling Sering Muncul

Modus adalah nilai yang paling sering muncul dalam suatu dataset. Suatu dataset dapat berupa:

- Unimodal (satu modus)
- Multimodal (beberapa modus)
- Tanpa modus (semua nilai sama seringnya muncul)

Tidak seperti rata-rata dan median, modus sangat berguna untuk data kategorikal.

Contoh: Hari Kerja dari Rumah yang Disukai

Jawaban:

$$0, 1, 2, 2, 3, 2, 4, 2, 1, 2, 3, 2, 5, 2, 1$$

Nilai 2 muncul 7 kali.

Modus = 2

Interpretasi

Sebagian besar karyawan lebih suka bekerja dari rumah dua hari per minggu. Ini adalah informasi yang dapat ditindaklanjuti secara langsung untuk desain kebijakan SDM.

Kapan Menggunakan Modus

- Untuk data kategorikal (merek, warna, tanggapan survei)
- Untuk mengidentifikasi kategori yang paling umum
- Untuk data numerik diskrit ketika dominasi frekuensi penting

Implementasi Excel

Fungsi:

$$= MODE.SNGL(number1, [number2], \dots) ,$$
$$= MODE.MULT(number1, [number2], \dots)]$$

Contoh:

Untuk modus tunggal:

$$= MODE.SNGL(B1:B12)$$

Untuk beberapa modus:

$$= MODE.MULT(B1:B12)$$

(Beberapa versi Excel mungkin memerlukan entri khusus untuk beberapa modus.)

Memilih Ukuran yang Tepat

Ukuran yang tepat bergantung pada struktur data dan tujuan analitis.

Tabel 4.1 Memilih ukuran yang tepat untuk data kategorikal

Situasi	**Ukuran Terbaik**
Data numerik simetris	Mean
Data yang menceng (*skewed*) dengan nilai pencilan (*outliers*)	Median
Data kategorikal	Mode
Menentukan pilihan yang paling populer	Mode
Pemodelan statistik formal	Mean

Aturan yang berguna:

- Jika nilai-nilai ekstrem mendistorsi realitas → Gunakan median.
- Jika setiap nilai harus memengaruhi hasil → Gunakan rata-rata (*mean*).
- Jika dominasi frekuensi itu penting → Gunakan modus.

Wawasan Penting

Rata-rata (*mean*) sangat berguna karena mencakup semua titik data. Median sangat berguna karena tahan terhadap distorsi. Modus sangat berguna karena menunjukkan dominasi. Namun, tidak ada satu ukuran pun yang secara universal lebih unggul. Analisis statistik yang bertanggung jawab sering kali melibatkan pelaporan lebih dari satu ukuran pemusatan terutama ketika distribusi data tidak diketahui atau tidak simetris. Ukuran pemusatan (*central tendency*) memberikan gambaran mengenai inti dari data tersebut. Pada bagian selanjutnya, kita akan membahas pasangannya: variabilitas ukuran seberapa rapat atau renggang inti data tersebut.

4.3 UKURAN VARIABILITAS: MEMAHAMI SEBARAN DATA

Jika ukuran pemusatan memberi tahu kita di mana data mengelompok, ukuran variabilitas mengungkapkan seberapa rapat kelompok data tersebut. Dua kumpulan data bisa memiliki rata-rata dan median yang sama, namun sangat berbeda dalam hal konsistensi, stabilitas, dan prediktabilitas. Satu kumpulan data mungkin terkonsentrasi rapat di sekitar pusatnya, sementara yang lain tersebar luas. Tanpa mengukur sebaran, kita tidak dapat menilai keandalan, risiko, atau ketidakpastian.

Variabilitas menjawab pertanyaan mendasar kedua dalam statistik deskriptif:

Seberapa stabil nilai-nilai yang ada di sekitar pusat data?

Pada bagian ini, kita akan mengkaji lima ukuran sebaran utama:

- Rentang (*Range*)
- Rentang Antarkuartil (*Interquartile Range*/IQR)
- Varians (*Variance*)
- Simpangan Baku (*Standard Deviation*)
- Koefisien Variasi (*Coefficient of Variation*)

Masing-masing memberikan sudut pandang berbeda mengenai sebaran data.

Rentang (Range): Ukuran Sebaran yang Paling Sederhana

Definisi: Rentang adalah selisih antara pengamatan terbesar dan terkecil:

$$Rentang = Maksimum - Minimum$$

Kapan Harus Digunakan

- Untuk mendapatkan gambaran awal yang cepat mengenai sebaran data
- Untuk kumpulan data yang kecil
- Ketika nilai-nilai ekstrem memiliki makna penting

Contoh (Data Pengeluaran Pemasaran)

- Maksimum = 15.000
- Minimum = 2.400
- Rentang = 15.000 − 2.400 = 12.600

Interpretasi

Anggaran pemasaran memiliki rentang sebesar Rp 126.000.000. Namun, ukuran ini hanya mencerminkan dua observasi dan mengabaikan semua data di antaranya.

Keterbatasan

Rentang (*range*) sangat sensitif terhadap nilai pencilan (*outlier*) dan tidak memberikan informasi apa pun mengenai distribusi nilai di bagian tengah.

Implementasi di Excel

Fungsi:

$$Minimum: = MIN(number1, [number2], \dots)$$
$$Maksimum: = MAX(number1, [number2], \dots)$$

Contoh: Menghitung rentang

$$= MAX(B1:B12) - MIN(B1:B12)$$

Rentang Antarkuartil (*Interquartile Range*/IQR): Bagian Tengah yang Tangguh

Untuk memahami sebaran nilai-nilai tipikal, kita membagi data menjadi empat bagian yang sama besar yang disebut kuartil:

- Q1 (persentil ke-25) – 25% nilai berada di bawahnya.

- Q2 (Median) – 50% nilai berada di bawahnya.
- Q3 (persentil ke-75) – 75% nilai berada di bawahnya.

Rentang Antarkuartil (IQR) mengukur sebaran dari 50% data di bagian tengah:

$$IQR = Q3 - Q1$$

Berbeda dengan rentang (*range*), IQR tidak mudah dipengaruhi oleh nilai-nilai ekstrem.

Contoh (Data Pemasaran)

$Q1 = 2.675$
$Q3 = 3.325$
$IQR = 3.325 - 2.675 = 650$

Interpretasi

Meskipun rentang (*range*) keseluruhan cukup besar, 50% data di bagian tengah mengelompok dalam rentang yang sempit, yaitu Rp 6.500.000. Hal ini menunjukkan bahwa sebagian besar perusahaan memiliki anggaran pemasaran yang sangat mirip, dan variabilitas besar yang terlihat terutama disebabkan oleh adanya pencilan (*outlier*) ekstrem.

IQR sangat berguna ketika data bersifat menceng (*skewed*).

Kapan Menggunakan IQR

- ❖ Saat terdapat pencilan (*outlier*)
- ❖ Saat digunakan bersama dengan median
- ❖ Saat mendeskripsikan sebaran yang "tipikal" (khas)

Implementasi di Excel

Fungsi:

$$= QUARTILE.INC(array, quart)$$
$$= QUARTILE.EXC(array, quart)$$

Contoh:

$$Q3 = QUARTILE.INC(B1{:}B12,3)$$
$$Q1 = QUARTILE.INC(B1{:}B12,1)$$
$$IQR = QUARTILE.INC(B1{:}B12,3) - QUARTILE.INC(B1{:}B12,1)$$

(Catatan: QUARTILE.INC bersifat inklusif dan umumnya lebih disukai. QUARTILE.EXC bersifat eksklusif.)

Varians dan Standar Deviasi

Varians: Rata-rata Kuadrat Deviasi

Sementara rentang (*range*) dan IQR mengukur sebaran menggunakan logika posisi, varians mengukur sebaran secara matematis. Varians adalah rata-rata dari kuadrat selisih antara setiap observasi dengan nilai rata-rata (*mean*). Penguadratan memiliki dua tujuan:

1. Mencegah deviasi negatif saling meniadakan dengan deviasi positif.

2. Memberikan bobot lebih besar (penalti lebih berat) terhadap deviasi yang besar.

Rumus

Varians Populasi

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Varian Sampel

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

Di mana:

- σ^2 adalah varians populasi
- $\sum(xi - \mu)^2$ adalah jumlah kuadrat selisih dari rata-rata populasi
- N adalah jumlah observasi populasi
- s^2 adalah varians sampel
- $\sum(xi - \bar{x})^2$ adalah jumlah kuadrat selisih dari rata-rata sampel
- $n - 1$ adalah derajat kebebasan (*degrees of freedom*)

Penggunaan $n - 1$ (derajat kebebasan) berfungsi untuk mengoreksi bias saat mengestimasi varians populasi berdasarkan sampel.

Interpretasi

Varians mencerminkan rata-rata kuadrat jarak dari nilai rata-rata (*mean*). Namun, karena dinyatakan dalam satuan kuadrat (contoh: rupiah²), varians menjadi kurang intuitif untuk diinterpretasikan secara langsung.

Simpangan Baku (Standar Deviasi): Variabilitas dalam Satuan Asli

Simpangan baku adalah akar kuadrat dari varians:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Hal ini mengembalikan ukuran tersebut ke satuan data yang asli, sehingga jauh lebih mudah untuk diinterpretasikan.

- ❖ Standar deviasi kecil → nilai-nilai mengelompok di dekat nilai rata-rata (*mean*).
- ❖ Standar deviasi besar → nilai-nilai tersebar luas.

Contoh (Data Pemasaran)

Perhitungan manual mengikuti rumus populasi

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

1) **Hitung nilai rata-rata (μ)**

$$Nilai\ rata - rata\ (\mu)\ =\ 3950$$

2) **Selisih dari nilai rata-rata dan kuadrat selisih tersebut**

Value x	$x - \mu$	$(x - \mu)$ ^2
2400	(2400-3950) = -1550	2402500
2500	(2500-3950) = -1450	2102500
2600	(2600-3950) = -1350	1822500
2700	(2700-3950) = -1250	1562500
2800	(2800-3950) = -1150	1322500
2900	(2900-3950) = -1050	1102500
3100	(3100-3950) = -850	722500
3200	(3200-3950) = -750	562500
3300	(3300-3950) = -650	422500
3400	(3400-3950) = -550	302500
3500	(3500-3950) = -450	202500
15000	(15000-3950) = 11050	122102500

Sekarang jumlahkan selisih-selisih yang telah dikuadratkan:

$$\sum (x - \bar{x})^2 = 134{,}630{,}000$$

3) **Varians populasi (bagi selisih kuadrat dengan N)**

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{134{,}630{,}000}{12} = 11{,}219{,}166.\overline{6}$$

$$\sigma^2 = 11{,}219{,}166.67$$

4) **Simpangan baku populasi (akar kuadrat dari varians)**

$$\sigma = \sqrt{\sigma^2} = \sqrt{11{,}219{,}166.6667} \approx 3{,}349.50$$

Simpangan baku populasi:

$$\sigma \approx 3{,}349.50$$

Interpretasi

Standar deviasi sebesar Rp 33.490.000 menunjukkan dispersi yang sangat tinggi dibandingkan dengan anggaran pemasaran pada umumnya. Namun, nilai yang besar ini terutama disebabkan oleh adanya pencilan (*outlier*) sebesar Rp 150.000.000. Jika pencilan tersebut dihapus, standar deviasi yang dihitung ulang akan turun secara drastis (menjadi sekitar Rp 3.600.000), yang menunjukkan bahwa sebagian besar bisnis sebenarnya memiliki pola pengeluaran yang sangat konsisten. Contoh ini memberikan wawasan penting: Standar deviasi sensitif terhadap nilai-nilai ekstrem.

Implementasi dalam Excel

Pilih Fungsi yang Tepat: Sampel vs. Populasi

- Varians populasi $= VAR.P(B1:B12)$
- Varians sampel $= VAR.S(B1:B12)$
- Standar deviasi populasi $= STDEV.P(B1:B12)$
- Standar deviasi sampel: $= STDEV.S(B1:B12)$

Rumus Populasi vs. Sampel: Memilih Perspektif yang Tepat

Keputusan statistik yang mendasar adalah menentukan apakah kumpulan data Anda mewakili keseluruhan populasi yang diteliti atau sekadar sampel yang diambil dari kelompok yang lebih besar. Rumus yang Anda pilih mencerminkan perbedaan tersebut.

Rumus populasi

Gunakan rumus populasi jika data Anda mencakup setiap anggota dari kelompok yang ingin Anda deskripsikan dan Anda tidak berupaya untuk melakukan generalisasi di luar kelompok tersebut. Dalam kasus ini, variabilitas dihitung dengan menggunakan N sebagai penyebut karena tidak diperlukan estimasi.

Contoh umum meliputi:

- Nilai dari seluruh 30 mahasiswa di kelas tertentu
- Suhu harian untuk setiap hari dalam satu tahun tertentu
- Gaji seluruh karyawan di sebuah perusahaan

Dalam hal ini, Anda mendeskripsikan kelompok yang lengkap dan tertutup.

Rumus sampel

Gunakan rumus sampel jika data Anda hanya mewakili sebagian kecil (*subset*) dari populasi yang lebih besar dan tujuan Anda adalah mengestimasi variabilitas dari kelompok

yang lebih luas tersebut. Dalam kasus ini, penyebutnya menjadi $n - 1$ sebuah koreksi yang dikenal sebagai penyesuaian Bessel (*Bessel's adjustment*) yang berfungsi mengompensasi kecenderungan alami sampel untuk meremehkan variabilitas populasi yang sebenarnya. Contoh umum meliputi:

- Survei terhadap 200 rumah tangga yang diambil dari total 20.000 rumah tangga.
- Sebagian mahasiswa yang dipilih dari beberapa sekolah.
- Sampel acak transaksi dari total penjualan selama satu tahun penuh.

Dalam praktiknya, jika Anda melakukan analisis statistik inferensial seperti penentuan selang kepercayaan (*confidence interval*), pengujian hipotesis, regresi, atau pemodelan prediktif rumus sampel hampir selalu menjadi pilihan yang tepat. Singkatnya, gunakan N saat Anda mendeskripsikan keseluruhan kelompok; gunakan $n - 1$ saat Anda mengestimasi sesuatu yang lebih besar daripada apa yang Anda amati secara langsung.

Koefisien Variasi (CV): Sebaran Relatif

Ukuran sebaran absolut (seperti simpangan baku) bergantung pada skala pengukuran. Simpangan baku sebesar 10 memiliki makna yang sangat berbeda jika nilai rata-ratanya 20 dibandingkan jika nilai rata-ratanya 2.000.

Koefisien Variasi (CV) menyatakan variabilitas secara relatif terhadap nilai rata-rata:

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\backslash\%$$

CV tidak memiliki satuan, sehingga memungkinkan perbandingan antar-kumpulan data yang memiliki skala atau satuan berbeda. Nilai CV yang lebih rendah menunjukkan variabilitas yang lebih kecil relatif terhadap nilai rata-rata, yang mengindikasikan konsistensi data yang lebih tinggi. Sebaliknya, nilai CV yang lebih tinggi menunjukkan variabilitas yang lebih besar.

Contoh (Data Pemasaran)

$$CV = \frac{3{,}349.5}{3{,}950} \times 100\backslash\% \approx 84.8\backslash\%$$

Interpretasi

Nilai CV sebesar 84,8% menunjukkan variabilitas relatif yang sangat tinggi yang sekali lagi disebabkan oleh adanya pencilan (*outlier*). Setelah nilai ekstrem tersebut dihapus:

$$CV \approx 12{,}2\%$$

Angka ini mencerminkan konsistensi yang kuat di antara sebagian besar bisnis. CV sangat berguna dalam bidang:

- Keuangan (perbandingan risiko)
- Manufaktur (pengendalian kualitas)
- Membandingkan variabilitas di berbagai unit yang berbeda

Implementasi di Excel

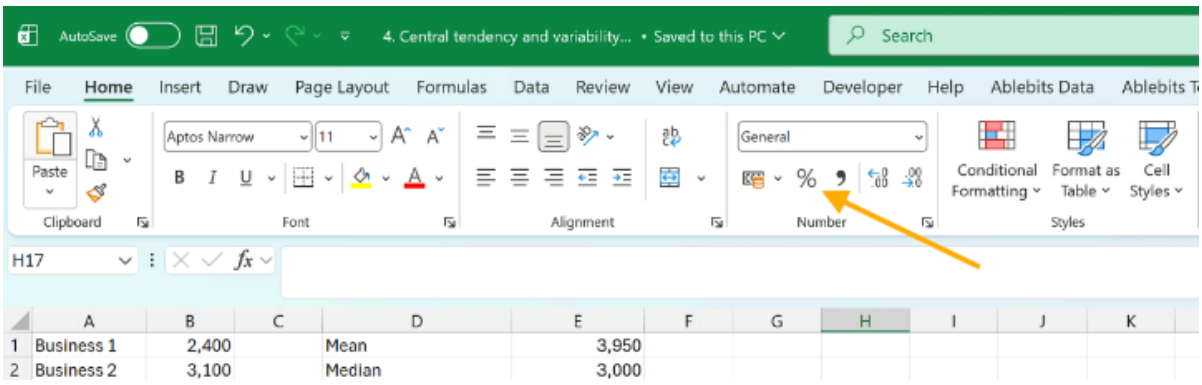
Excel tidak memiliki fungsi khusus bawaan untuk menghitung Koefisien Variasi (*Coefficient of Variation*), namun menyediakan semua perangkat yang diperlukan untuk menghitungnya dengan mudah.

Contoh:

$$= STDEV.S(B1{:}B12) \,/\, AVERAGE(B1{:}B12)$$

Format sebagai persentase

- Pilih sel atau rentang sel yang ingin Anda format.
- Pada tab Beranda (*Home*), klik tombol Persentase (%) di grup Angka (*Number*).
- Excel akan mengubah nilai tersebut menjadi persentase dan menambahkan simbol % secara otomatis.



Gambar 4.1 Nilai yang Ditetapkan sebagai Persentase

4.4 STUDI KASUS KEPUASAN KARYAWAN: ANALISIS TENDENSI SENTRAL DAN SEBARAN DATA

Studi Kasus: Survei Kepuasan Karyawan

Untuk memperkuat pemahaman kita tentang kecenderungan sentral dan variabilitas, mari kita analisis kumpulan data yang realistis. Misalkan sebuah perusahaan mensurvei sampel 20 karyawan dan meminta mereka untuk menilai kepuasan kerja mereka pada skala 1 (sangat tidak puas) hingga 10 (sangat puas).

Kumpulan Data:

7, 8, 6, 9, 7, 5, 8, 7, 9, 6, 8, 7, 4, 8, 7, 9, 6, 8, 7, 10

Masukkan nilai-nilai tersebut ke dalam satu kolom Excel (misalnya, F1:F20). Karena 20 karyawan ini hanya mewakili sebagian kecil dari semua karyawan di perusahaan, kita akan menggunakan rumus sampel jika sesuai.

Langkah 1: Mengukur Pusat

Rata-rata

$$= RATA - RATA(F1{:}F20) \;\rightarrow\; 7{,}3$$

Median

$$= MEDIAN(F1{:}F20) \rightarrow 7$$

Modus

$$= MODE.SNGL(F1{:}F20) \rightarrow 7$$

Interpretasi

- Skor kepuasan rata-rata adalah 7,3, menunjukkan sentimen yang umumnya positif.
- Median dan modus keduanya adalah 7, menunjukkan bahwa karyawan pada umumnya menilai kepuasan pada angka 7.
- Kedekatan rata-rata dan median menunjukkan distribusi simetris tanpa kemiringan ekstrem (akan dibahas lebih lanjut di Bab 5).

Pada tahap ini, pimpinan mungkin menyimpulkan: "Karyawan cukup puas." Tetapi kesimpulan ini tetap tidak lengkap tanpa memahami sebaran.

Langkah 2: Mengukur Sebaran

Rentang

$$= MAX(F1{:}F20) - MIN(F1{:}F20) \rightarrow 6$$
$$(Tertinggi = 10, Terendah = 4)$$

Rentang Antar Kuartil (IQR)

$$= QUARTILE.INC(F1{:}F20{,}3) - QUARTILE.INC(F1{:}F20{,}1) \rightarrow 1{,}25$$

Deviasi Standar Sampel

$$= STDEV.S(F1{:}F20) \rightarrow 1{,}45$$

Koefisien Variasi (CV)

$$= (STDEV.S(F1{:}F20)/AVERAGE(F1{:}F20)) * 100 \rightarrow \approx 20\%$$

Menginterpretasikan Variabilitas

- Rentang 6 menunjukkan kesenjangan yang nyata antara karyawan yang paling tidak puas dan yang paling puas.
- IQR sebesar 1,25 menceritakan kisah yang lebih terfokus: 50% karyawan di tengah berkumpul dalam rentang yang sempit. Ini menunjukkan bahwa sebagian besar karyawan memiliki perasaan yang serupa tentang pekerjaan mereka.

- Deviasi standar sebesar 1,45 menunjukkan dispersi yang relatif rendah pada skala 10 poin.
- Koefisien variasi (~20%) mengonfirmasi bahwa variabilitas tergolong moderat jika dibandingkan dengan nilai rata-rata.

Gambaran Statistik

Meskipun tingkat kepuasan bervariasi antara 4 hingga 10, sebagian besar karyawan berada di kisaran angka 7. Data ini menunjukkan konsistensi dengan variabilitas yang rendah, bukan polarisasi ataupun ketidakstabilan. Dari perspektif manajerial:

- Tidak ada bukti adanya ketidakpuasan yang meluas.
- Namun, skor di angka yang lebih rendah (4 dan 5) menunjukkan bahwa ada kelompok minoritas yang mungkin memerlukan perhatian.

Contoh ini mengilustrasikan prinsip krusial dalam statistik: Nilai rata-rata (*mean*) menunjukkan tingkatannya. Variabilitas menunjukkan stabilitasnya. Keduanya secara bersama-sama mengungkapkan fakta yang sebenarnya. Dengan mengintegrasikan ukuran tendensi sentral dan variabilitas, kita beralih dari sekadar menghitung angka menjadi memperoleh wawasan organisasional yang berdasar.

4.5 INTERPRETASI TENDENSI SENTRAL DAN VARIABILITAS DALAM ANALISIS STATISTIK

Dalam bab ini, kita melangkah lebih jauh dari sekadar mendeskripsikan posisi data, dan mengembangkan keterampilan yang lebih mendalam untuk memahami bagaimana data tersebut berperilaku. Kita memulainya dengan ukuran tendensi sentral mean, median, dan modus serta mempelajari bahwa meskipun ketiganya sama-sama mengidentifikasi "pusat", masing-masing melakukannya dari sudut pandang filosofis yang berbeda. Mean menyeimbangkan setiap nilai, median tahan terhadap distorsi akibat nilai ekstrem, sementara modus menangkap realitas yang paling sering muncul. Pemilihan di antara ketiganya bukanlah proses mekanis semata; hal ini bergantung pada sifat dan tujuan dari data tersebut.

Selanjutnya, kita memperluas perspektif ke ukuran variabilitas, dengan menyadari bahwa nilai pusat saja tidaklah cukup. Dua kumpulan data bisa memiliki nilai rata-rata yang sama, namun menceritakan kisah yang sangat berbeda bergantung pada sebaran datanya. Melalui rentang (*range*), rentang interkuartil (IQR), varians, dan standar deviasi, kita menelaah seberapa rapat nilai-nilai mengelompok atau seberapa lebar nilai-nilai tersebut tersebar. Kita melihat bagaimana IQR melindungi interpretasi dari pengaruh pencilan (*outlier*), sementara varians dan standar deviasi memberikan ukuran sebaran yang secara matematis ketat.

Sebuah wawasan penting muncul: variabilitas bukanlah sekadar pelengkap matematis, melainkan properti struktural dari data. Standar deviasi mengukur stabilitas. Koefisien variasi memperluas pemahaman ini dengan memungkinkan perbandingan lintas skala yang berbeda, melalui konsep variabilitas relatif. Kita juga memperjelas perbedaan statistik yang mendasar: rumus untuk populasi versus rumus untuk sampel. Penggunaan N mencerminkan pengetahuan yang lengkap, sedangkan penggunaan $n-1$ mencerminkan estimasi dan mempersiapkan kita untuk penalaran inferensial. Perbedaan ini menjadi jembatan konseptual antara statistik deskriptif dan inferensi statistik.

Di sepanjang bab ini, Excel tidak sekadar berfungsi sebagai kalkulator, melainkan sebagai laboratorium mengubah rumus-rumus abstrak menjadi wawasan yang dapat diukur. Setiap fungsi memperkuat satu pelajaran utama: statistik bukanlah soal menghafal rumus, melainkan soal menafsirkan struktur. Yang terpenting, Anda telah belajar untuk berpikir secara berpasangan: pusat dan sebaran, tingkat dan stabilitas, nilai dan variasi. Kesadaran ganda inilah yang mengubah angka mentah menjadi sebuah wawasan. Pada bab berikutnya, kita akan mulai mengembangkan landasan deskriptif ini menuju penalaran analitis yang lebih mendalam di mana data tidak hanya menggambarkan realitas, tetapi juga mulai menjelaskannya.

BAB 5
BENTUK DISTRIBUSI, KEMIRINGAN, & KURTOSIS

5.1 BENTUK DISTRIBUSI DAN PENERAPANNYA DALAM ANALISIS STATISTIK

"Bentuk data menceritakan sebuah kisah yang tidak dapat diungkapkan oleh angka-angka saja." Pada Bab 4, kita telah mempelajari cara menentukan pusat data dan mengukur penyebarannya. Kita mengidentifikasi di mana nilai-nilai mengelompok dan seberapa besar variasinya. Namun, meskipun telah menggunakan ringkasan-ringkasan yang informatif tersebut, ada satu dimensi penting yang belum dibahas: bentuk distribusi. Data bukan sekadar kumpulan nilai; data memiliki bentuk. Bentuk atau distribusi ini menunjukkan bagaimana pengembangan teknologi masa depan tersusun di sepanjang rentang nilainya. Bentuk tersebut memberi tahu kita apakah nilai-nilai terkumpul secara simetris di sekitar pusat, menjulur tidak merata ke satu sisi, mengelompok dengan rapat, atau menghasilkan ekor yang ekstrem. Dua kumpulan data bisa saja memiliki nilai rata-rata dan simpangan baku yang sama, namun memiliki karakteristik yang sangat berbeda karena bentuk distribusinya yang berlainan. Memahami bentuk distribusi mengubah analisis statistik dari sekadar perhitungan menjadi sebuah wawasan. Bentuk distribusi memengaruhi:

- Apakah nilai rata-rata dapat diandalkan atau justru menyesatkan
- Apakah uji parametrik tepat untuk digunakan
- Apakah nilai-nilai ekstrem merupakan hal yang wajar atau justru pengecualian
- Apakah model prediktif akan memberikan hasil yang tangguh (*robust*)

Singkatnya, bentuk distribusi menentukan seberapa yakin kita dalam menarik kesimpulan berdasarkan data. Bab ini memperkenalkan dua ukuran formal untuk bentuk distribusi: *skewness* (kemiringan), yang mengukur asimetri, dan *kurtosis*, yang menggambarkan bobot ekor serta konsentrasi data. Kita juga akan mengeksplorasi bagaimana teknik visualisasi seperti histogram dan kurva densitas dapat mengungkap pola yang tidak dapat dijelaskan sepenuhnya hanya melalui rumus. Di akhir bab ini, Anda akan mampu untuk:

- Mengidentifikasi dan mendeskripsikan bentuk-bentuk distribusi yang umum.
- Menghitung dan menafsirkan *skewness* serta *kurtosis*.
- Menilai simetri, ekor, dan konsentrasi data secara visual.
- Memahami bagaimana bentuk distribusi memengaruhi pengambilan keputusan statistik.
- Menerapkan analisis distribusi pada kumpulan data di dunia nyata.

Bab ini menandai langkah maju yang penting. Kita beralih dari pertanyaan "Di mana letak datanya?" dan "Seberapa besar variasinya?" menuju pertanyaan yang lebih mendalam: "Bagaimana struktur data tersebut?" Pertanyaan itulah yang menjadi inti dari penalaran statistik yang tepat.

5.2 BENTUK DISTRIBUSI DAN KARAKTERISTIK DATA

Distribusi data menggambarkan keseluruhan nilai yang mungkin ada dalam suatu kumpulan data beserta frekuensi kemunculan nilai-nilai tersebut. Distribusi bukan sekadar daftar angka, melainkan pola struktural yang dibentuk oleh angka-angka tersebut secara kolektif. Sementara ukuran seperti rata-rata (*mean*) dan simpangan baku (*standard deviation*) merangkum karakteristik spesifik dari data, distribusi mengungkapkan bagaimana keseluruhan kumpulan data diorganisasikan di sepanjang rentangnya. Distribusi menjawab pertanyaan yang lebih mendalam:

Bagaimana pengembangan teknologi masa depan tersebut tersusun?

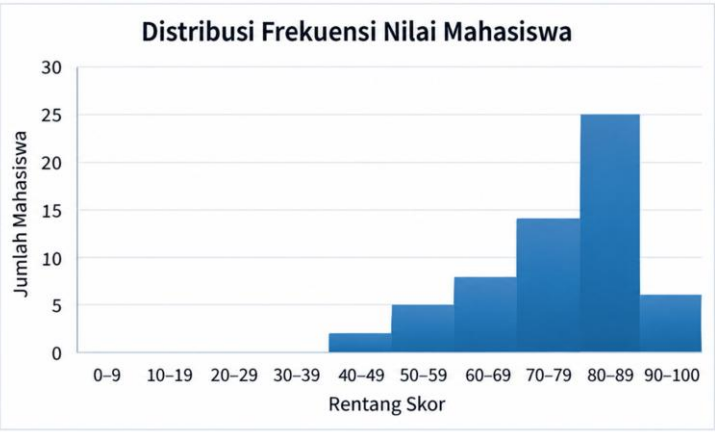
Analogi Sederhana: Nilai Ujian

Misalkan Anda mengumpulkan nilai ujian dari 60 mahasiswa: 85, 92, 77, 46, 85, 64, 78, 88, ... Dalam bentuk mentah, daftar tersebut tidak memberikan informasi yang berarti. Untuk memperoleh wawasan, kita perlu mengaturnya. Langkah pertama adalah mengurutkan data. Langkah kedua adalah menghitung seberapa sering setiap nilai muncul. Ketika kumpulan data menjadi besar mencakup ratusan atau ribuan pengembangan teknologi masa depan menghitung nilai satu per satu menjadi tidak praktis. Sebagai gantinya, nilai-nilai tersebut dikelompokkan ke dalam interval (kelompok/bin), seperti:

- 40–49
- 50–59
- 60–69
- 70–79
- 80–89
- 90–100

Selanjutnya, kita menghitung berapa banyak pengembangan teknologi masa depan yang masuk ke dalam setiap kelompok. Hasilnya dapat disajikan dalam bentuk tabel frekuensi atau divisualisasikan sebagai histogram.

Rentang Skor	Frekuensi
0–9	0
10–19	0
20–29	0
30–39	0
40–49	2
50–59	5
60–69	8
70–79	14
80–89	25
90–100	6
Total	**60**



Gambar 5.1: Tabel frekuensi dan histogram untuk distribusi nilai mahasiswa

Histogram mengubah daftar yang berantakan menjadi struktur yang dapat dilihat. Histogram menampilkan:

- Rentang nilai (pada sumbu horizontal)
- Frekuensi pengamatan (pada sumbu vertikal)

Untuk kumpulan data berskala kecil, tabel mungkin sudah cukup. Namun, untuk kumpulan data berskala besar, visualisasi menjadi sangat penting. Persepsi manusia jauh lebih efisien dalam mendeteksi pola melalui bentuk dibandingkan melalui deretan angka.

Apa yang Diungkapkan oleh Distribusi

Setelah divisualisasikan, distribusi memungkinkan kita untuk segera menjawab pertanyaan-pertanyaan kunci:

1. **Nilai Mana yang Umum atau Jarang Ditemukan?**
   Anda dapat dengan cepat melihat di mana sebagian besar pengamatan terkonsentrasi dan apakah terdapat nilai yang tidak lazim. Sebagai contoh, jika sebagian besar mahasiswa memperoleh nilai di kisaran 80-an, sementara satu mahasiswa memperoleh nilai 20, maka nilai tersebut akan terlihat menonjol sebagai kemungkinan pencilan (*outlier*).
2. **Di Mana Letak Pusatnya?**
   Meskipun kita dapat menghitung nilai rata-rata (*mean*) atau median secara numerik, distribusi memungkinkan kita untuk melihat pusat data tersebut secara langsung. Keseimbangan visual sering kali menunjukkan apakah nilai rata-rata merupakan ringkasan yang dapat diandalkan.
3. **Bagaimana Sebaran Datanya?**
   Apakah nilai-nilai tersebut terkonsentrasi dengan rapat atau tersebar luas? Apakah sebarannya meluas secara seimbang ke kedua arah, atau salah satu sisinya lebih panjang? Hal ini membawa kita pada salah satu konsep terpenting dalam analisis distribusi: bentuk.

Simetri dan Asimetri

Distribusi simetris memiliki keseimbangan di sekitar pusatnya. Jika dilipat secara vertikal tepat di tengah, kedua sisinya akan saling mencerminkan satu sama lain. Contoh yang paling terkenal adalah distribusi normal, yang sering disebut sebagai "kurva lonceng" (*bell curve*). Dalam distribusi yang benar-benar normal:

- Rata-rata (*Mean*) = Median = Modus
- Sebagian besar nilai terkonsentrasi di dekat pusat
- Ekor kurva mengecil secara merata di kedua sisi

Banyak proses alami yang mendekati bentuk ini, termasuk kesalahan pengukuran dan karakteristik biologis tertentu.

Distribusi Asimetris

Ketika salah satu ekor kurva memanjang lebih jauh dibandingkan sisi lainnya, distribusi tersebut dikatakan menceng (*skewed*).

- Menceng ke kanan (*positively skewed*): Sebagian besar nilai berada di angka rendah, dengan beberapa nilai yang sangat tinggi menarik ekor kurva ke arah kanan.

- Menceng ke kiri (*negatively skewed*): Sebagian besar nilai berada di angka tinggi, dengan beberapa nilai yang sangat rendah menarik ekor kurva ke arah kiri.

Kecenderungan menceng (*skewness*) mengubah interpretasi data karena nilai-nilai ekstrem akan menarik nilai rata-rata ke arah ekor kurva. Bentuk-Bentuk Distribusi yang Umum dan Mengapa Hal Itu Penting Bentuk distribusi merupakan salah satu petunjuk paling penting mengenai bagaimana data berperilaku. Bentuk memengaruhi hampir setiap keputusan statistik:

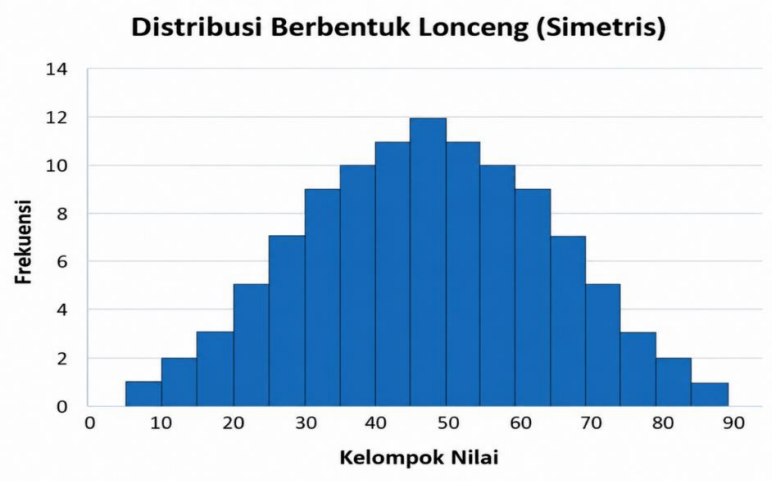
- bagaimana kita meringkas data,
- metode statistik atau uji parametrik apa yang tepat untuk digunakan,
- bagaimana kita menafsirkan hasil,
- apakah transformasi data diperlukan,
- apakah analisis subkelompok secara terpisah diperlukan,
- proses dunia nyata apa yang mungkin menghasilkan data tersebut.

Bentuk data memberi tahu kita apakah data tersebut seimbang atau condong, apakah nilai ekstrem sering muncul atau jarang terjadi, serta apakah nilai rata-rata dapat memberikan informasi yang bermakna atau justru menyesatkan.

Memahami bentuk data membantu analis menghindari asumsi yang keliru, memilih perangkat yang tepat, dan menafsirkan pola dengan penuh keyakinan. Saat Anda mengamati ilustrasi berbagai bentuk distribusi yang umum pada tahap ini, fokuslah pada perspektif strukturalnya (bentuknya). Aspek-aspek yang berkaitan dengan grafik, penyesuaian, gaya, pelabelan, dan teknik lainnya akan dibahas pada bab berikutnya.

1. Distribusi Berbentuk Lonceng (Simetris)

Distribusi ini memiliki satu puncak di bagian tengah dengan sisi-sisi yang merupakan cerminan satu sama lain.



Gambar 5.2: Distribusi berbentuk lonceng

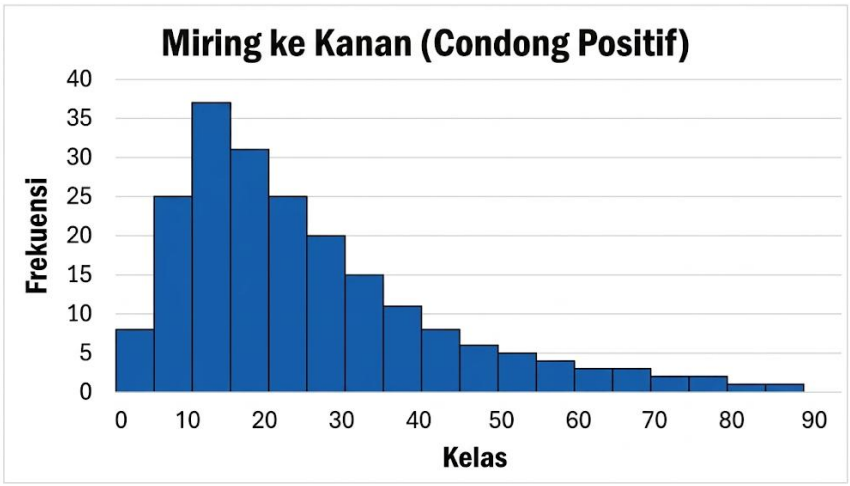
Contoh: Distribusi Normal (misalnya, distribusi usia dalam populasi yang besar). Mengapa ini penting: Dalam distribusi yang simetris sempurna, Rata-rata (*Mean*) ≈ Median ≈ Modus. Hal ini menjadikan nilai rata-rata sebagai nilai "tipikal" yang sangat stabil dan bermakna.

Implikasi:

- Banyak uji statistik tingkat lanjut mengasumsikan bahwa data Anda berdistribusi mendekati normal.
- Pencilan (*outlier*) lebih jarang ditemukan dibandingkan pada distribusi dengan ekor yang tebal (*heavy-tailed*).
- Jika data Anda mengikuti bentuk ini, simpangan baku (standar deviasi) menjadi alat yang sangat ampuh untuk memprediksi di mana titik data di masa mendatang akan berada.
- Berguna untuk memodelkan variasi alami (seperti tinggi badan, kesalahan pengukuran).

2. Distribusi Miring ke Kanan (Miring Positif)

Pada distribusi ini, "ekor" data memanjang ke arah kanan. Sebagian besar pengembangan teknologi masa depan terkonsentrasi di sisi nilai rendah, sementara beberapa nilai ekstrem yang tinggi menarik nilai rata-rata ke arah atas.



Gambar 5.3: Bentuk distribusi yang menceng ke kanan (*right-skewed*)

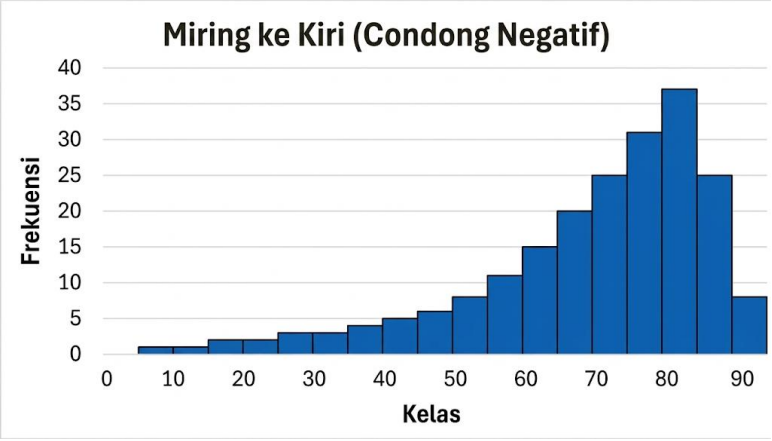
Contoh: Distribusi pendapatan rumah tangga (sebagian besar orang memiliki penghasilan dalam jumlah yang wajar, namun terdapat segelintir miliarder). Mengapa hal ini penting: Nilai rata-rata (*mean*) > nilai tengah (*median*). Pencilan (*outlier*) di sisi atas mendistorsi nilai rata-rata, sehingga membuatnya tampak lebih tinggi daripada nilai yang "tipikal" bagi kebanyakan individu.

Implikasi:

- Nilai tengah (median) biasanya merupakan ukuran yang lebih baik untuk menggambarkan kondisi yang "tipikal".
- Ekor yang panjang di sisi kanan menunjukkan adanya risiko nilai yang sangat tinggi.
- Varians bisa menjadi besar atau tidak stabil.

3. Distribusi yang Menceng ke Kiri (Negatively Skewed)

"Ekor" distribusi memanjang ke arah kiri; artinya, sebagian besar titik data bernilai tinggi, namun terdapat beberapa nilai yang sangat rendah yang menarik nilai rata-rata ke bawah.



Gambar 5.4: Bentuk distribusi yang menceng ke kiri (*negatively skewed*)

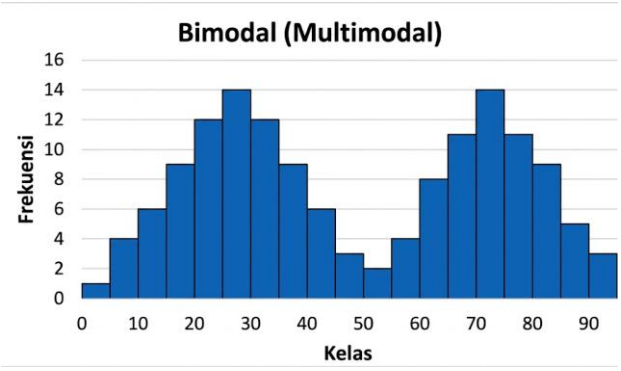
Contoh: Usia pensiun atau nilai ujian yang mudah (kebanyakan orang pensiun di usia lanjut, namun ada segelintir yang pensiun sangat dini; sebagian besar mahasiswa mendapatkan nilai sangat baik dalam ujian yang mudah, sementara segelintir lainnya mengalami kesulitan). Mengapa ini penting: Nilai rata-rata (*mean*) < median. Nilai-nilai yang sangat rendah lebih memengaruhi distribusi dibandingkan nilai-nilai yang tinggi.

Implikasi:

- Bentuk ini sering kali menunjukkan adanya "efek plafon" (*ceiling effect*) atau batas atas alami dalam suatu proses.
- Seperti halnya data yang menceng ke kanan, median lebih stabil (*robust*) daripada rata-rata dalam kondisi ini.
- Ekor kiri yang panjang menunjukkan adanya risiko munculnya nilai-nilai yang sangat rendah.

4. Distribusi Bimodal dan Multimodal

Distribusi ini memiliki dua puncak atau lebih yang terpisah, menyerupai punggung unta.



Gambar 5.5: Bentuk distribusi bimodal

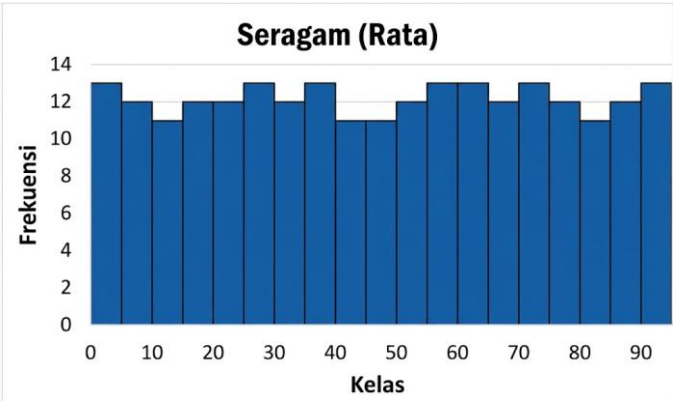
Contoh: Jam sibuk restoran (puncak terjadi pada waktu makan siang dan makan malam) atau tinggi badan populasi campuran (laki-laki dan perempuan). Mengapa ini penting: Meringkas data semacam ini dengan satu nilai rata-rata akan sangat menyesatkan. Nilai rata-rata tersebut mungkin jatuh di "lembah" di antara dua puncak area di mana hampir tidak ada data yang benar-benar muncul.

Implikasi:

- Bentuk bimodal merupakan indikator kuat bahwa Anda sedang melihat dua kelompok berbeda yang digabungkan menjadi satu.
- Diperlukan segmentasi (misalnya, menganalisis waktu makan siang dan makan malam secara terpisah), pengelompokan (*clustering*), atau pengetahuan mendalam mengenai bidang terkait untuk mendapatkan wawasan yang akurat.
- Standar deviasi kehilangan daya interpretasinya.

5. Distribusi Seragam (Uniform/Flat)

Dalam distribusi seragam, setiap nilai yang mungkin memiliki peluang kemunculan yang sama besar. Tidak ada "puncak" yang terlihat. Contoh: Melempar Dadu yang Adil atau Angka Lotre. Mengapa Ini Penting: Tidak ada nilai "khas" sentral. Rata-rata dan median ada, tetapi keduanya tidak mewakili "kelompok" data.



Gambar 5.6: Tahap distribusi uwifior

Implikasi:

- Keseragaman menunjukkan kurangnya preferensi atau proses yang sepenuhnya acak.
- Sering digunakan sebagai "model nol" atau garis dasar untuk melihat apakah pola yang lebih menarik sebenarnya ada.

Wawasan Akhir

Distribusi mengubah daftar nilai yang tidak teratur menjadi narasi visual yang terstruktur. Data tidak tersebar secara acak; data membentuk pola. Pola-pola tersebut mencerminkan kekuatan yang menghasilkan data. Bentuk menentukan apakah kesimpulan stabil atau menyesatkan. Oleh karena itu, memahami distribusi bukan hanya deskriptif; tetapi juga diagnostik. Ini mempersiapkan kita untuk mengajukan pertanyaan yang lebih canggih tentang asimetri, perilaku ekor, dan asumsi statistik topik yang sekarang kita jelajahi melalui kemiringan dan kurtosis.

5.3 KEMIRINGAN DISTRIBUSI (*SKEWNESS*) DAN INTERPRETASINYA

Sementara ukuran tendensi sentral dan variabilitas menggambarkan di mana data berada dan seberapa luas penyebarannya, kemiringan (*skewness*) menggambarkan ke arah mana data tersebut condong. Kemiringan mengukur tingkat asimetri dalam suatu distribusi. Ukuran ini menunjukkan apakah nilai-nilai data lebih menjulur jauh pada satu sisi pusat dibandingkan sisi lainnya. Secara konseptual, ini mengukur seberapa "timpang" struktur data tersebut.

Kemiringan Positif (Distribusi Miring ke Kanan)

Suatu distribusi dikatakan memiliki kemiringan positif jika ekor panjangnya menjulur ke arah kanan. Hal ini terjadi ketika terdapat sejumlah kecil nilai yang sangat besar, yang menarik distribusi ke arah positif.

Karakteristik:

- Sebagian besar pengembangan teknologi masa depan mengelompok di sisi kiri (nilai yang lebih rendah).
- Beberapa nilai ekstrem yang sangat tinggi membentuk ekor panjang di sisi kanan.
- Rata-rata (*Mean*) > Median > Modus.

Karena nilai rata-rata sensitif terhadap nilai ekstrem, maka rata-rata akan tertarik ke arah ekor distribusi. Contoh:Pendapatan rumah tangga. Sebagian besar rumah tangga memiliki pendapatan menengah, namun ada sebagian kecil yang memiliki pendapatan sangat tinggi, sehingga menarik distribusi ke arah kanan.

Kemiringan Negatif (Distribusi Miring ke Kiri)

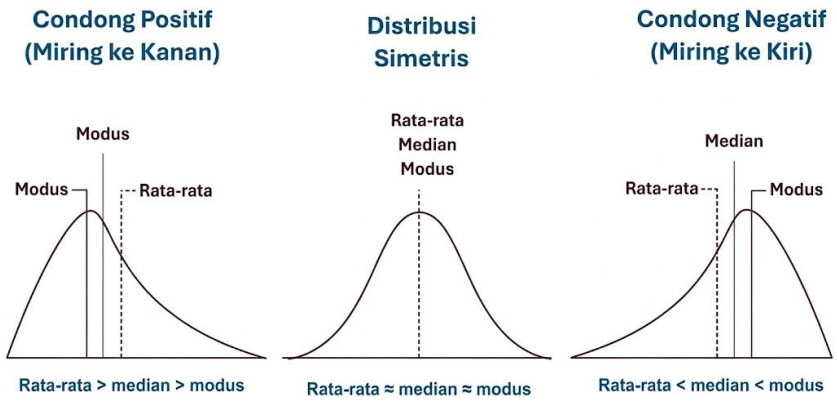
Suatu distribusi dikatakan memiliki kemiringan negatif jika ekor panjangnya menjulur ke arah kiri. Hal ini terjadi ketika terdapat sejumlah kecil nilai yang sangat rendah, yang menarik distribusi ke arah nilai yang lebih kecil (ke arah kiri).

Karakteristik:

- Sebagian besar pengembangan teknologi masa depan mengelompok di sisi kanan (nilai yang lebih tinggi).
- berapa nilai ekstrem yang sangat rendah membentuk ekor panjang di sisi kiri.
- Rata-rata (*Mean*) < Median < Modus.

Sekali lagi, nilai rata-rata akan tertarik ke arah ekor distribusi.

Contoh: Usia saat meninggal di negara-negara maju. Sebagian besar individu hidup hingga usia lanjut, namun sejumlah kecil kasus kematian di usia muda menyebabkan distribusi menjulur ke arah kiri.



Gambar 5.7: Perbandingan Kemiringan Positif, Sistem, dan Kemiringan Negatif dalam Distribusi

Mengapa Kemiringan Penting

Kemiringan memengaruhi interpretasi statistik dalam beberapa cara penting:

- Rata-rata dapat menjadi menyesatkan dalam data yang sangat miring.
- Banyak model statistik mengasumsikan simetri perkiraan.
- Nilai ekstrem dapat mewakili realitas struktural (misalnya, ketidaksetaraan kekayaan) daripada sekadar anomali.

Dalam distribusi miring, median sering memberikan ukuran kecenderungan sentral yang lebih stabil daripada rata-rata.

Menghitung Kemiringan (Skewness) di Excel

Excel menyediakan fungsi langsung untuk menghitung kemiringan sampel (*sample skewness*):

=SKEW(*range*)

Contoh: Jika data Anda berada di sel A1:A100:

=SKEW(A1:A100)

Fungsi SKEW di Excel menghitung kemiringan sampel (yang telah disesuaikan untuk mengurangi bias). Untuk kemiringan populasi (*population skewness*), Excel juga menyediakan fungsi:

=SKEW.P(range)

Menafsirkan Nilai Kemiringan (*Skewness*)

- **Skewness ≈ 0** → Distribusi yang kira-kira simetris
- **Skewness > 0** → Miring ke kanan (kemiringan positif)
- **Skewness < 0** → Miring ke kiri (kemiringan negatif)

Aturan Praktis (Rule of Thumb)

- |Skewness| < 0.5 → Kira-kira simetris
- 0.5 ≤ |Skewness| < 1 → Miring secara moderat (sedang)
- |Skewness| ≥ 1 → Miring secara signifikan (tinggi)

Ambang batas ini hanyalah pedoman, bukan batasan yang kaku. Penafsiran harus selalu mempertimbangkan konteks, ukuran sampel, dan pemeriksaan secara visual.

5.4 KURTOSIS: MENGUKUR BOBOT EKOR DAN KONSENTRASI DISTRIBUSI

Jika tendensi sentral memberi tahu kita di mana data seimbang, variabilitas memberi tahu kita seberapa luas data tersebut tersebar, dan kemiringan (*skewness*) memberi tahu kita ke arah mana data tersebut condong, maka kurtosis memberi tahu kita seberapa ekstrem nilai-nilai ekstrem yang ada. Kurtosis mengukur karakteristik ekor suatu distribusi secara spesifik, seberapa besar probabilitas yang terletak pada bagian ekor dibandingkan dengan distribusi normal. Meskipun kurtosis sering digambarkan sebagai ukuran "puncak" (*peakedness*), interpretasi statistik modern lebih menekankan pada sensitivitasnya terhadap nilai-nilai ekstrem, bukan sekadar ketinggian titik pusatnya. Pada intinya, kurtosis menjawab pertanyaan berikut: Seberapa besar kemungkinan munculnya pengembangan teknologi masa depan ekstrem dibandingkan dengan apa yang diharapkan dalam distribusi normal?

Tolok Ukur Normal

Distribusi normal berfungsi sebagai titik acuan.

- Kurtosis mentah (*raw kurtosis*) dari distribusi normal = 3
- Kurtosis berlebih (*excess kurtosis*) dari distribusi normal = 0

Karena membandingkan segala sesuatu dengan angka 3 bisa terasa kurang intuitif, para analis biasanya menggunakan kurtosis berlebih (*excess kurtosis*), yang diperoleh dengan mengurangi nilai kurtosis mentah sebesar 3.

Kategori Kurtosis

Kurtosis biasanya diklasifikasikan berdasarkan perbandingannya dengan distribusi normal.

1. **Mesokurtik**

   Definisi: Distribusi dengan karakteristik ekor yang serupa dengan distribusi normal.

   Nilai:

   - Kurtosis mentah = 3
   - Kurtosis berlebih = 0

   Interpretasi: Nilai-nilai ekstrem muncul dengan frekuensi yang kurang lebih sama dengan yang diharapkan dalam distribusi normal. Tampilan Visual: Kurva berbentuk lonceng standar.

2. **Leptokurtik**

   Definisi: Distribusi dengan ekor yang lebih tebal (lebih berat) dibandingkan distribusi normal.

   Nilai:

   - Kurtosis mentah > 3
   - Kurtosis berlebih > 0

   Interpretasi:

   - Probabilitas nilai ekstrem lebih besar.
   - Lebih banyak pengembangan teknologi masa depan yang terkonsentrasi di dekat nilai rata-rata (*mean*) dan di bagian ekor.
   - Pencilan (*outlier*) lebih sering terjadi dibandingkan yang diharapkan dalam distribusi normal.

   Tampilan Visual: Puncak yang lebih tajam dengan ekor yang lebih tebal.

3. Platykurtik

Definisi: Distribusi dengan ekor yang lebih tipis (lebih ringan) dibandingkan distribusi normal.

Nilai:

- Kurtosis mentah < 3
- Kurtosis berlebih (*excess kurtosis*) < 0

Interpretasi:

- Lebih sedikit nilai ekstrem
- Pengembangan teknologi masa depan tersebar lebih merata di seluruh rentang
- Kemungkinan munculnya pencilan (*outlier*) lebih rendah

Kesan Visual: Puncak yang lebih datar dengan ekor yang lebih tipis.

Menghitung Kurtosis di Excel

Fungsi KURT() dalam Excel menghasilkan nilai *excess kurtosis* (kurtosis berlebih).

=KURT(rentang_data)

Contoh:

=KURT(A1:A100)

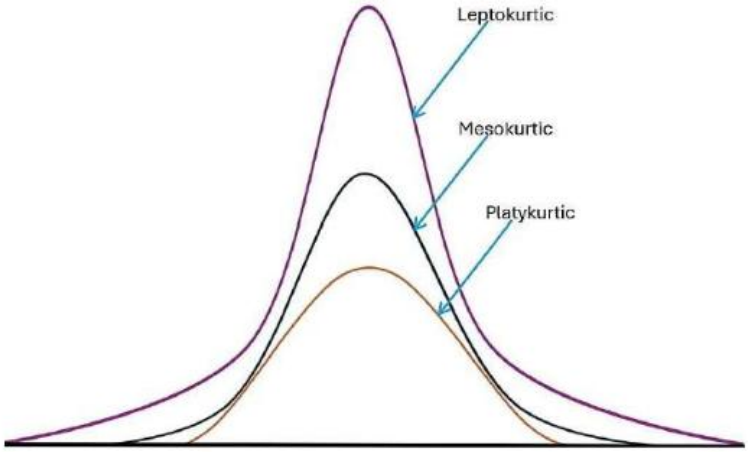
Jika hasilnya:

- Positif → ekor distribusi lebih tebal daripada normal
- Negatif → ekor distribusi lebih tipis daripada normal

Menafsirkan Nilai Kurtosis

- > 0 → Ekor tebal (*leptokurtik*)
- = 0 → Mendekati normal (*mesokurtik*)
- < 0 → Ekor tipis (*platikurtik*)

Karena tidak mengharuskan kita untuk mengurangi angka 3 secara manual, *excess kurtosis* menjadi standar modern yang digunakan saat ini.



Gambar 5.8: Kurva Leptokurtik, Mesokurtik, dan Platikurtik

Sebagian besar perangkat lunak statistik termasuk Excel, Python, dan R secara bawaan menampilkan nilai *excess kurtosis*.

Mengapa Kurtosis Penting dalam Praktik?

Kurtosis bukan sekadar ukuran deskriptif; nilai ini memiliki konsekuensi analitis yang nyata.

1. **Penilaian Risiko (Keuangan)**
   Di pasar keuangan, distribusi imbal hasil (*return*) sering kali bersifat *leptokurtik*. Ekor yang tebal menunjukkan bahwa keuntungan atau kerugian ekstrem terjadi lebih sering daripada yang diprediksi oleh model distribusi normal. Meremehkan kurtosis dapat menyebabkan kesalahan fatal dalam perhitungan risiko.
2. **Asumsi Model**
   Banyak uji statistik mengasumsikan data mendekati distribusi normal. Meskipun kurtosis saja tidak cukup untuk menguji normalitas, nilai ini memberikan indikasi apakah perilaku ekor distribusi menyimpang secara signifikan dari asumsi tersebut. Kurtosis yang tinggi menunjukkan bahwa:
   - Interval kepercayaan (*confidence interval*) mungkin tidak dapat diandalkan.
   - Peristiwa ekstrem lebih mungkin terjadi daripada yang diperkirakan.
3. **Pengendalian Kualitas**
   Dalam manufaktur, kurtosis yang tinggi dapat mengindikasikan munculnya cacat ekstrem sesekali, meskipun proses produksi secara keseluruhan tampak stabil.
4. **Kesehatan Masyarakat dan Epidemiologi**
   Data penyebaran penyakit dapat menunjukkan ekor yang tebal ketika terjadi wabah ekstrem yang jarang namun berdampak besar (misalnya, peristiwa *super-spreader*). Kurtosis membantu mengukur risiko yang berkaitan dengan ekor distribusi tersebut.

Perspektif Konseptual

Skewness (kemiringan) menunjukkan ketidakseimbangan arah, sedangkan kurtosis menunjukkan konsentrasi nilai ekstrem. Dua distribusi bisa saja simetris sempurna (*skewness* ≈ 0), namun memiliki perilaku ekor yang sangat berbeda. Satu distribusi mungkin menghasilkan nilai ekstrem yang jarang terjadi tetapi sangat besar, sementara distribusi lainnya mungkin hampir tidak pernah menghasilkan pencilan (*outlier*). Dengan demikian, kurtosis melengkapi deskripsi struktural suatu distribusi:

- Pusat (*Center*) → titik di mana nilai-nilai mencapai keseimbangan
- Sebaran (*Spread*) → seberapa luas nilai-nilai bervariasi
- Kemiringan (*Skewness*) → ke arah mana ekor distribusi memanjang
- Kurtosis → seberapa tebal atau berat ekor distribusi tersebut

Secara keseluruhan, ukuran-ukuran ini memberikan pemahaman multidimensi mengenai bentuk distribusi, hal yang sangat penting untuk penalaran statistik tingkat lanjut dan interpretasi data yang bertanggung jawab.

5.5 MEMVISUALISASIKAN BENTUK DISTRIBUSI

Memahami konsep kemiringan (*skewness*) dan kurtosis adalah langkah awal yang penting, namun melihatnya secara visual akan memberikan wawasan yang jauh lebih mendalam. Sebelum menjalankan uji statistik formal, analis profesional hampir selalu memeriksa distribusi secara visual. Grafik yang dipilih dengan tepat dapat mengungkap

asimetri, konsentrasi data, karakteristik ekor, pengelompokan (*klaster*), serta potensi pencilan (*outlier*) dalam hitungan detik bahkan sebelum satu pun statistik dihitung. Meskipun Bab 6 secara khusus membahas seni dan teknik dalam menyusun visualisasi yang efektif, bagian ini akan memperkenalkan tiga jenis grafik dasar untuk mendiagnosis bentuk distribusi.

1. Histogram: Pemantauan Struktural terhadap Data

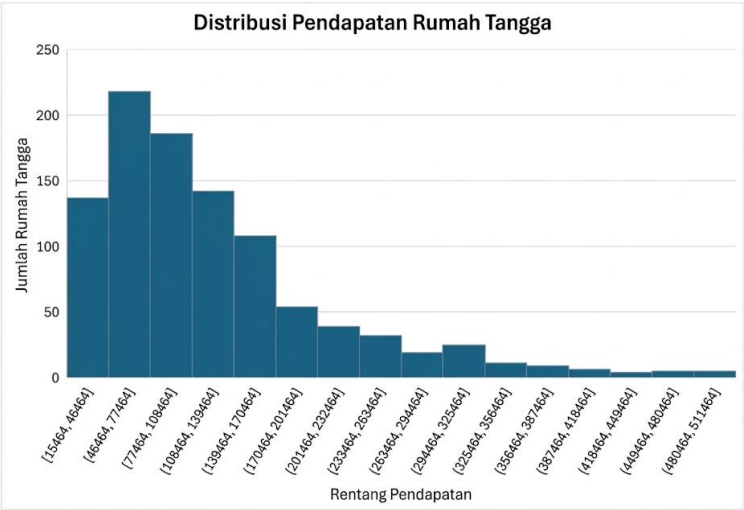
Histogram merupakan alat paling mendasar untuk memahami bentuk distribusi. Grafik ini mengelompokkan nilai-nilai numerik ke dalam interval (bin) dengan lebar yang sama, lalu menampilkan frekuensi data di dalam setiap interval tersebut. Berbeda dengan tabel frekuensi sederhana, histogram mampu memperlihatkan struktur data, seperti:

- Di mana data terkonsentrasi (mengelompok)
- Apakah distribusi bersifat simetris atau miring (*skewed*)
- Apakah terdapat lebih dari satu puncak (multimodal)
- Apakah ekor distribusi cenderung panjang atau pendek

Membuat Histogram Dasar di Excel

1. Pilih data numerik Anda (satu kolom).
2. Buka tab Insert → Statistical Chart → Histogram.
3. Sesuaikan pengaturan bin jika diperlukan (klik kanan pada sumbu horizontal → Format Axis).

Pada tahap ini, fokuslah hanya untuk mendapatkan gambaran struktural yang jelas. Kustomisasi tingkat lanjut, penataan gaya, pemberian label, dan teknik penyampaian cerita (*storytelling*) akan dibahas pada Bab 6.



Gambar 5.9: Histogram dasar

Apa yang Harus Diperiksa dalam Histogram

Saat menafsirkan histogram, perhatikan:

- **Puncak (Modus):** Satu puncak (unimodal), dua puncak (bimodal), atau beberapa puncak (multimodal).
- **Simetri vs. Kemiringan:** Apakah satu ekor lebih panjang daripada yang lain?

- **Sebaran:** Apakah distribusinya sempit atau lebar?
- **Celah:** Apakah ada interval tanpa pengamatan?
- **Pencilan:** Apakah ada batang terisolasi yang jauh dari kelompok utama?

Pedoman praktisnya adalah menggunakan antara 5 dan 20 bin untuk sebagian besar dataset sedang. Terlalu sedikit bin menyembunyikan struktur; terlalu banyak menciptakan noise visual.

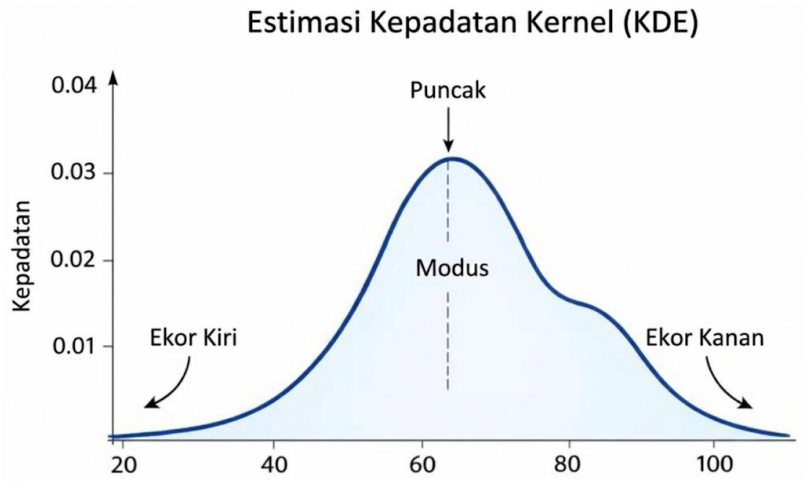
2. Estimasi Kepadatan Kernel (KDE): Perspektif yang Dihaluskan

Histogram bergantung pada lebar bin. Plot KDE menghilangkan ketergantungan ini dengan memperkirakan kurva kepadatan probabilitas yang halus di atas data.

Kurva KDE:

- Menghilangkan batas bin
- Membuat puncak yang halus lebih terlihat
- Memperjelas perilaku ekor
- Memberikan pandangan yang lebih bersih tentang bentuk distribusi

Plot KDE sangat berguna saat membandingkan distribusi.



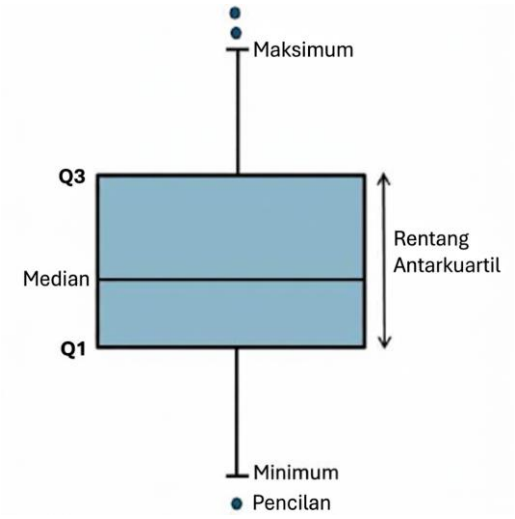
Gambar 5.10: Kurva estimasi kepadatan kernel

Catatan: Excel tidak dapat membuat plot KDE secara bawaan. Plot ini umumnya dibuat menggunakan perangkat lunak statistik seperti Python atau R. Seiring dengan berkembangnya kemampuan analisis Anda, visualisasi KDE akan menjadi alat diagnostik standar. Untuk saat ini, pahamilah KDE secara konseptual sebagai penyempurnaan dari histogram sebuah sarana yang lebih halus untuk melihat struktur distribusi dengan lebih jelas.

3. Boxplot: Diagnostik Bentuk yang Ringkas

Boxplot (atau *box-and-whisker plot*) menyajikan ringkasan bentuk distribusi secara ringkas dengan menggunakan statistik lima angka (*five-number statistics*):

1. Minimum — nilai terkecil yang bukan merupakan pencilan (*outlier*)
2. Q1 (Kuartil Pertama) — 25% data berada di bawah titik ini
3. Median (Q2) — titik tengah dari data
4. Q3 (Kuartil Ketiga) — 75% data berada di bawah titik ini
5. Maksimum — nilai terbesar yang bukan merupakan pencilan (*outlier*)

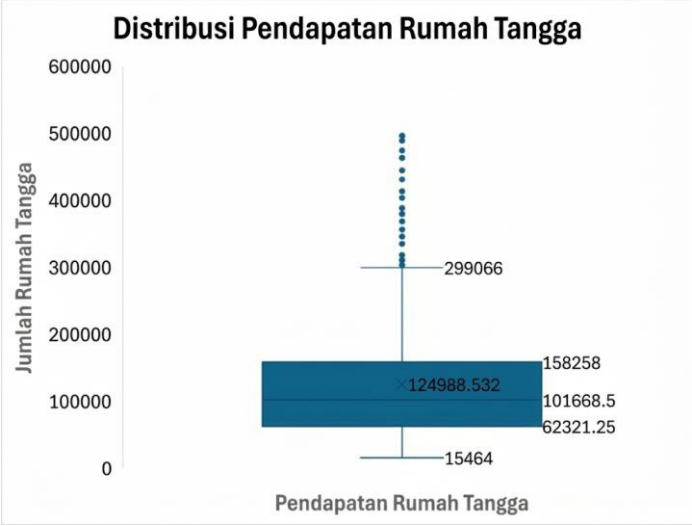


Gambar 5.11 Boxplot yang menampilkan statistik lima angka

Membuat Boxplot Dasar di Excel

1. Pilih data numerik Anda (gunakan kolom terpisah untuk kelompok yang berbeda).
2. Buka menu Insert → Statistical Chart → Box and Whisker.

Sekali lagi, teknik pemformatan dan penyajian secara mendetail akan dibahas pada Bab 6.



Gambar 5.12 Diagram kotak yang menunjukkan distribusi kuartil

Bagaimana Bentuk Mengungkap Karakteristik Distribusi

Diagram kotak tidak hanya menunjukkan angka; diagram ini menunjukkan pola:

- **Kemiringan**
  - Jika median lebih dekat ke Q1 daripada Q3, distribusinya miring ke kanan.
  - Jika median lebih dekat ke Q3 daripada Q1, distribusinya miring ke kiri.
  - Jika garis kumis tidak rata, itu juga menandakan kemiringan.
- **Sebaran (variabilitas)**
  - IQR (Q3 − Q1) adalah lebar kotak.

  - Kotak yang lebar berarti variabilitas tinggi; kotak yang sempit berarti variabilitas rendah.
- **Pencilan**
  - Titik-titik di luar 1,5 × IQR dari kuartil muncul sebagai titik-titik individual.
  - Kehadiran pencilan menandakan pengamatan yang tidak biasa atau ekstrem.
- **Simetri**
  - Median yang terpusat dan garis kumis yang sama menunjukkan distribusi yang kurang lebih simetris.

Mengapa Boxplot Sangat Berguna

Boxplot unggul dalam situasi di mana plot lain kesulitan:

- **Membandingkan beberapa kelompok secara berdampingan**
  Boxplot membuat perbedaan median, sebaran, dan outlier langsung terlihat.
- **Meringkas kumpulan data besar**
  Boxplot mengabaikan noise dan fokus pada struktur.
- **Menyoroti anomali**
  Outlier terlihat jelas tanpa mendistorsi distribusi utama.

Apa yang Tidak Ditunjukkan Boxplot

Boxplot itu ampuh tetapi minimalis. Boxplot tidak menunjukkan:

- Bentuk distribusi yang tepat (misalnya, bimodality)
- Kepadatan nilai dalam kuartil
- Ukuran sampel (kecuali jika diberi anotasi)

Inilah mengapa analis sering memasangkan boxplot dengan histogram atau plot biola. Anda akan bertemu dengan plot biola dan plot tingkat lanjut lainnya ketika Anda mulai menggunakan alat yang lebih canggih seperti Python.

Contoh interpretasi praktis

Jika Anda melihat boxplot di mana:

- Median berada di dekat bagian bawah kotak
- Garis kumis atas panjang
- Beberapa outlier muncul di ujung atas

…Anda sedang melihat distribusi miring ke kanan dengan variabilitas tinggi dan nilai-nilai ekstrem yang tinggi.

Memilih Visual yang Tepat untuk Analisis Bentuk

Tabel 5.1 Memilih visual yang tepat untuk analisis bentuk

Tujuan	Alat Terbaik
Mendiagnosis struktur distribusi secara keseluruhan	Histogram
Memeriksa kepadatan yang mulus (*smooth density*) dan perilaku ekor distribusi	KDE
Membandingkan kelompok-kelompok secara cepat	plot kotak

Setiap visual menawarkan sudut pandang yang berbeda. Secara keseluruhan, visual-visual tersebut memberikan gambaran yang komprehensif mengenai bentuk suatu distribusi.

5.6 STUDI KASUS: ANALISIS DISTRIBUSI PENDAPATAN

Untuk menerapkan konsep ukuran pemusatan dan bentuk distribusi, mari kita analisis kumpulan data nyata. Buku kerja Excel Bab 5 memuat data `household_income` (pendapatan rumah tangga) untuk 1.000 rumah tangga, dengan tingkat pendapatan berkisar antara Rp150 juta hingga Rp5 milyar Mengingat adanya kelompok dengan pendapatan yang sangat tinggi, kita dapat memperkirakan bahwa distribusi data akan menunjukkan kemencengan ke kanan (*right skewness*).

Langkah 1: Statistik Deskriptif Dasar (Excel)

Menggunakan kolom pendapatan (B2:B1001):

Ukuran	**Rumus Excel**	**Hasil**
Rata-rata	=AVERAGE(B2:B1001)	Rp1.249.885.300
Median	=MEDIAN(B2:B1001)	Rp1.016.685.000
Modus	=MODE.MULT(B2:B1001)	Rp836.400.000 dan Rp836.400.000

Pengembangan teknologi masa depan:

Kumpulan data ini memiliki dua nilai yang berulang (lebih dari satu modus). Dari sini, kita dapat melihat: **Mean > Median > Modus**

Pola ini sering kali menunjukkan distribusi yang menceng ke kanan (right-skewed).

Langkah 2: Mengukur Bentuk Distribusi

Ukuran	**Rumus Excel**	**Hasil**
Kemiringan (Skewness)	=SKEW(B2:B1001)	1,6
Kurtosis (Ekses)	=KURT(B2:B1001)	2,8

Interpretasi Ukuran Bentuk: Kemencengan (Skewness) = 1,6

Nilai kemencengan sebesar 1,6 menunjukkan kemencengan positif (ke kanan) yang kuat. Hal ini berarti:

- Sebagian kecil rumah tangga memiliki pendapatan yang sangat tinggi.
- Nilai-nilai ekstrem ini menarik nilai rata-rata (*mean*) ke arah atas.
- Nilai rata-rata (mean) menjadi lebih besar daripada nilai tengah (median).

Kurtosis (Ekses) = 2,8

Fungsi KURT() dalam Excel menghasilkan kurtosis ekses, yang berarti nilai tersebut dilaporkan relatif terhadap nilai 0 pada distribusi normal.

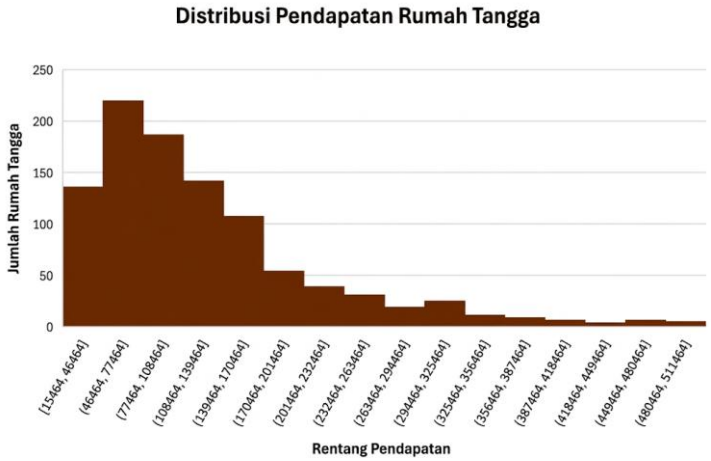
Nilai 2,8 menunjukkan:

- Ekor distribusi yang lebih tebal dibandingkan distribusi normal.
- Nilai pendapatan ekstrem muncul lebih sering daripada yang terjadi pada distribusi normal.
- Distribusi tersebut mengandung pencilan (*outlier*) yang signifikan.

Secara praktis: kelompok dengan pendapatan sangat tinggi tidak hanya sekadar ada, tetapi juga memberikan pengaruh yang nyata terhadap struktur data secara keseluruhan.

Langkah 3: Analisis Visual

Buatlah histogram di Excel dengan menggunakan kategori (*bin*) pendapatan yang sesuai.



Gambar 5.13 Histogram distribusi pendapatan rumah tangga

Histogram tersebut menunjukkan:

- Sebagian besar rumah tangga memiliki pendapatan antara Rp 150 juta dan Rp 17 Triliun.
- Terdapat ekor kanan yang panjang, menjulur ke arah Rp 5 milyar.
- Mayoritas populasi terkonsentrasi pada kisaran pendapatan rendah hingga menengah.

Visualisasi ini mengonfirmasi apa yang ditunjukkan oleh statistik: adanya kemencengan (*skewness*) ke kanan yang jelas, dengan nilai-nilai ekstrem berpendapatan sangat tinggi (*outlier*). Rumah Tangga "Tipikal" Meskipun pendapatan rata-rata (*mean*) adalah Rp 1,249 milyar nilai tengahnya (*median*) hanya Rp 1,16 milyar Hal ini menunjukkan bahwa:

- Pendapatan rata-rata memberikan gambaran yang terlalu tinggi dibandingkan pendapatan rumah tangga yang tipikal.
- Median merupakan ukuran tendensi sentral yang lebih baik untuk data pendapatan yang tidak simetris (menceng).

Jika seorang pembuat kebijakan hanya melaporkan nilai rata-rata, hal tersebut dapat menciptakan kesan yang menyesatkan mengenai tingkat kesejahteraan secara umum.

Ketimpangan Kekayaan

Kombinasi dari:

- Kemencengan ke kanan yang kuat (1,6)
- Kurtosis berlebih (*excess kurtosis*) yang tinggi (2,8)
- Ekor kanan yang panjang

mengindikasikan adanya ketimpangan pendapatan yang signifikan. Sebagian kecil rumah tangga memiliki pendapatan yang sangat tinggi jauh di atas proporsi lainnya sehingga menarik distribusi tersebut menjadi lebih panjang ke arah nilai yang lebih besar.

Implikasi Kebijakan

Wawasan statistik menjadi landasan bagi pengambilan keputusan di dunia nyata. Berdasarkan distribusi ini:

- Gunakan nilai median saat mendeskripsikan "pendapatan rumah tangga tipikal."

- Pertimbangkan ambang batas persentil (misalnya, 10% teratas atau 1% teratas) saat merancang kebijakan pajak.
- Tentukan sasaran program sosial berdasarkan posisi dalam distribusi, bukan sekadar berdasarkan pendapatan rata-rata secara keseluruhan.

Dengan demikian, statistik tidak sekadar meringkas angka statistik mengungkap struktur sosial.

5.7 STATISTIK DESKRIPTIF DENGAN EXCEL DATA ANALYSIS TOOLPAK

Pada Bab 3, 4, dan 5, kita telah mempelajari komponen-komponen dasar statistik deskriptif:

- Ringkasan dalam bentuk tabel (jumlah, total, rasio, persentase)
- Ukuran pemusatan data (rata-rata/mean, median, modus)
- Ukuran penyebaran data (rentang/range, varians, standar deviasi)
- Ukuran bentuk distribusi (kemiringan/skewness dan kurtosis)

Excel memungkinkan kita untuk menghitung semua nilai tersebut dalam satu keluaran (*output*) terpadu menggunakan Data Analysis ToolPak.

Perangkat ini tidak menggantikan pemahaman teoretis, melainkan membantu menerapkannya secara praktis.

Langkah 1: Mengaktifkan Data Analysis ToolPak (Jika Belum Terlihat)

1. Klik File → Options
2. Pilih Add-ins
3. Di bagian bawah, pilih Excel Add-ins → Klik Go
4. Centang Analysis ToolPak
5. Klik OK

Sekarang, Anda akan melihat menu Data Analysis di bawah tab Data.

Langkah 2: Menghasilkan Statistik Deskriptif

1. Buka tab Data
2. Klik Data Analysis
3. Pilih Descriptive Statistics → Klik OK
4. Atur pengaturan berikut:
    - Input Range: Pilih kolom data Anda
    - Centang Labels in first row (jika ada label di baris pertama)
    - Tentukan Output Range (lokasi hasil) atau pilih New Worksheet (lembar kerja baru)
    - Centang Summary Statistics
5. Klik OK

Excel akan menghasilkan laporan statistik terstruktur seperti contoh di bawah ini.

Kolom 1	
Rata-rata	124,988.53
Galat Standar	2,812.19
Median	101,668.50
Modus	83,640.00
Simpangan Baku	88,929.27
Varians Sampel	7,908,415,778.96
Kurtosis	2.83
Kemiringan	1.60
Rentang	480,812.00
Minimum	15,464.00
Maksimum	496,276.00
Jumlah	124,988,532.00
Jumlah Data	1,000.00

Gambar 5.14 Laporan Statistik Deskriptif

Output Statistik Deskriptif dari ToolPak menyajikan ringkasan komprehensif mengenai kumpulan data. Kecuali untuk galat baku (*standard error*) yang akan dibahas pada bab-bab selanjutnya Anda telah mempelajari setiap ukuran yang tercantum dalam laporan ini. Dengan demikian, Anda telah memiliki bekal yang cukup untuk menafsirkan hasil tersebut dengan penuh keyakinan.

5.8 PEMUSATAN DATA DAN MENELAAH BENTUK DISTRIBUSI DATA

Dalam bab ini, kita melangkah lebih jauh dari sekadar ukuran pemusatan data dan mulai menelaah bentuk distribusi data. Meskipun nilai rata-rata (*mean*), median, dan modus dapat menjelaskan di mana pusat data berada, ukuran-ukuran tersebut tidak menunjukkan bagaimana nilai-nilai tersebar, merentang, atau terdistorsi. Untuk berpikir layaknya seorang ilmuwan data, kita harus memahami baik pusat maupun bentuk distribusi data tersebut. Kita telah membahas dua ukuran statistik utama: *Skewness* (kemiringan) dan *Kurtosis* (keruncingan).

Skewness menunjukkan apakah suatu distribusi condong ke kiri atau ke kanan.

- *Skewness* positif (condong ke kanan) terjadi ketika terdapat beberapa nilai besar yang menarik distribusi ke arah kanan (atau ke atas dalam konteks nilai).
- *Skewness* negatif (condong ke kiri) terjadi ketika terdapat beberapa nilai yang sangat kecil yang menarik distribusi ke arah kiri (atau ke bawah dalam konteks nilai).
- Jika nilai *skewness* mendekati nol, maka distribusi tersebut dapat dikatakan simetris.

Kita telah melihat bagaimana *skewness* memengaruhi interpretasi data secara langsung:

- Pada data yang condong ke kanan, nilai rata-rata (*mean*) lebih besar daripada median.
- Pada data yang condong ke kiri, nilai rata-rata (*mean*) lebih kecil daripada median.
- Median sering kali lebih baik dalam merepresentasikan nilai "tipikal" pada distribusi yang tidak simetris (*skewed*). *Skewness* membantu kita mendeteksi distorsi dan distorsi ini dapat memengaruhi pengambilan keputusan.

Kurtosis mengukur bobot pada bagian ekor suatu distribusi. Dengan menggunakan fungsi KURT() di Excel, kita menghitung *excess kurtosis* (kelebihan kurtosis), yang membandingkan data kita dengan distribusi normal.

- *Excess kurtosis* positif menunjukkan ekor yang tebal (*heavy tails*), yang berarti terdapat lebih banyak nilai ekstrem.
- *Excess kurtosis* negatif menunjukkan ekor yang tipis (*lighter tails*), yang berarti lebih sedikit nilai ekstrem.
- Nilai yang mendekati nol menunjukkan karakteristik ekor yang menyerupai distribusi normal.

Kurtosis membantu kita dalam menilai risiko, volatilitas, serta kemungkinan terjadinya hasil yang ekstrem hal-hal yang sangat krusial dalam bidang keuangan, kesehatan masyarakat, ekonomi, dan analitik bisnis. Kita juga melengkapi ukuran numerik tersebut dengan histogram, yang menunjukkan bahwa pola pikir statistik akan menjadi lebih kuat apabila:

- Angka dan visualisasi saling selaras.
- Interpretasi didasarkan pada bukti yang ada.
- Pola dikonfirmasi secara grafis.

Visualisasi bukanlah sekadar dekorasi, melainkan bentuk validasi.

Wawasan Akhir: Data jarang berperilaku dengan sempurna.

Kumpulan data di dunia nyata mengandung asimetri, pencilan (*outliers*), dan nilai-nilai ekstrem. Peran seorang ilmuwan data bukan sekadar menghitung rata-rata, melainkan memahami struktur, mendeteksi distorsi, serta menafsirkan implikasinya.

BAB 6
SENI VISUALISASI DATA

6.1 VISUALISASI DATA DAN PRINSIP ART DALAM ANALISIS DATA

"Sebuah gambar bermakna seribu kata namun visualisasi data yang dirancang dengan baik jauh lebih berharga daripada sekadar seribu titik data." Visualisasi data adalah titik di mana analisis berubah menjadi wawasan. Angka memang dapat mendeskripsikan suatu kumpulan data, tetapi visualisasi mampu mengungkap strukturnya pola, hubungan, dan anomali di dalamnya dengan cara yang jarang bisa dilakukan oleh angka mentah. Grafik yang dirancang dengan baik tidak sekadar menampilkan data; grafik tersebut mengubah informasi menjadi pemahaman.

Pada bab-bab sebelumnya, Anda telah mempelajari cara meringkas data menggunakan ukuran pemusatan, penyebaran, dan bentuk distribusi. Ringkasan statistik ini memberikan deskripsi numerik yang kuat mengenai suatu kumpulan data. Namun, angka saja tidak selalu mampu mengungkapkan keseluruhan cerita. Visualisasi memberikan perspektif pelengkap: memungkinkan Anda untuk melihat distribusi, membandingkan kelompok secara instan, mendeteksi anomali, serta mengomunikasikan wawasan dengan jelas. Dalam praktiknya, banyak penemuan paling penting dalam analisis data justru muncul pertama kali melalui grafik sederhana.

Bab ini memperkenalkan pendekatan berbasis prinsip untuk visualisasi data menggunakan Microsoft Excel. Alih-alih menyajikan grafik sebagai alat yang berdiri sendiri, kami mengembangkan kerangka kerja pengambilan keputusan praktis yang disebut kerangka kerja ART (*ART framework*), yang menjadi panduan dalam membuat dan mengevaluasi visualisasi yang efektif. Kerangka kerja ART menekankan tiga prinsip utama:

- **Akurasi (*Accuracy*):** memastikan bahwa visualisasi merepresentasikan data yang mendasarinya secara tepat dan jujur.
- **Relevansi (*Relevance*):** memilih jenis grafik yang paling tepat untuk menjawab pertanyaan analisis.
- **Transparansi (*Transparency*):** merancang grafik yang jelas, jujur, dan mudah ditafsirkan oleh audiens.

Dengan menggunakan kerangka kerja ini, Anda tidak hanya akan belajar cara membuat grafik di Excel, tetapi juga cara berpikir kritis mengenai pilihan visualisasi. Anda akan mengeksplorasi bagaimana visualisasi mendukung analisis data eksploratif, bagaimana konsep statistik dapat dikomunikasikan secara visual, serta bagaimana perangkat interaktif dalam Excel dapat mengubah grafik statis menjadi dasbor analisis yang dinamis. Di akhir bab ini, Anda akan mampu untuk:

- Menerapkan kerangka kerja ART untuk mengevaluasi dan merancang grafik
- Memilih jenis grafik yang tepat untuk pertanyaan analitis tertentu
- Memformat grafik Excel sesuai dengan prinsip-prinsip visualisasi data profesional

- Menggunakan visualisasi untuk analisis data eksploratif (EDA) sebelum melakukan analisis statistik formal
- Membuat grafik statistik seperti histogram, *box plot*, dan *scatter plot* yang dilengkapi dengan garis tren (*trendline*)
- Membangun visualisasi Excel yang dinamis dan interaktif menggunakan *PivotChart* dan *slicer*
- Menggunakan grafik untuk mengomunikasikan wawasan melalui narasi data yang jelas: konteks → wawasan → tindakan

Menguasai keterampilan-keterampilan ini akan memungkinkan Anda mengubah kumpulan data mentah menjadi cerita visual yang jelas, yang mampu mengungkap wawasan, mendukung analisis, serta memandu pengambilan keputusan.

6.2 PERAN VISUALISASI DALAM SAINS DATA

Visualisasi data memiliki dua peran berbeda dalam alur kerja sains data. Memahami perbedaan ini akan mempertajam keputusan yang Anda ambil saat membuat grafik dan menafsirkan pola-pola visual. Selama fase analisis data eksploratori (EDA), Anda adalah audiens utamanya. Visualisasi menjadi sarana untuk penemuan membantu Anda mengajukan pertanyaan terhadap data sebelum hipotesis formal dirumuskan. Grafik eksploratori tidak perlu dipoles atau disiapkan untuk presentasi.

Tujuannya adalah kecepatan dan wawasan. Histogram sederhana dapat mengungkap distribusi yang tidak simetris (miring) dalam hitungan detik. Diagram sebar (*scatter plot*) sederhana dapat memperlihatkan hubungan yang tidak Anda duga sebelumnya. Bahkan visualisasi kasar pun dapat menyoroti pola, klaster, atau anomali yang mengarahkan investigasi Anda menuju pertanyaan analitis yang paling menjanjikan. Pada fase ini, visualisasi membantu Anda berpikir dengan menggunakan data.

Visualisasi Komunikatif

Dalam fase komunikasi, audiensnya berubah. Alih-alih mengeksplorasi data untuk diri sendiri, Anda menyajikan wawasan kepada pihak lain seperti rekan kerja, manajer, klien, atau masyarakat umum. Oleh karena itu, visualisasi komunikatif harus lebih dari sekadar mengungkap pola; visualisasi tersebut harus menjelaskannya dengan jelas dan meyakinkan. Grafik dalam konteks ini harus dirancang dengan cermat, diberi label yang jelas, dan dibersihkan dari elemen apa pun yang dapat mengalihkan perhatian dari pesan utama.

Tujuannya bukan sekadar menampilkan data, melainkan mengarahkan audiens menuju wawasan yang ingin disampaikan. Karena keputusan mungkin bergantung pada visualisasi ini, standar untuk kejelasan, akurasi, dan desain menjadi jauh lebih tinggi. Grafik yang disusun dengan buruk dapat menyebabkan salah interpretasi, hilangnya kredibilitas, atau pengambilan keputusan yang keliru.

Peran Excel dalam Visualisasi Profesional

Microsoft Excel menempati posisi penting dalam lanskap ini. Meskipun bukan platform visualisasi tercanggih yang tersedia alat seperti Power BI, Tableau, atau pustaka pemrograman seperti seaborn di Python dan ggplot2 di R menawarkan fleksibilitas yang lebih besar Excel

tetap menjadi lingkungan analitis yang paling banyak digunakan dalam dunia bisnis maupun pembelajaran terapan.

Visualisasi dan Tanggung Jawab Analitis

Satu poin krusial yang perlu ditekankan: visualisasi yang buruk bukan sekadar masalah estetika. Grafik yang menyesatkan seperti grafik dengan sumbu yang dipotong, proporsi yang terdistorsi, atau elemen visual yang terlalu ramai dapat menciptakan kesan keliru dan berujung pada kesimpulan yang salah. Kesalahan semacam itu telah berkontribusi pada kekeliruan di dunia nyata, mulai dari bidang kedokteran dan keuangan hingga kebijakan publik.

Oleh karena itu, visualisasi data membawa tanggung jawab etis yang penting. Analis harus memastikan bahwa grafik yang mereka buat mengomunikasikan data yang mendasarinya secara jujur dan jelas. Kerangka kerja yang diperkenalkan pada bagian selanjutnya dirancang untuk mendukung tanggung jawab ini dengan menyediakan serangkaian prinsip untuk mengevaluasi dan merancang visualisasi yang dapat dipercaya.

6.3 KERANGKA KERJA ART UNTUK VISUALISASI

Kerangka kerja ART menyajikan tiga prinsip panduan untuk mengevaluasi dan merancang visualisasi data yang efektif. Alih-alih menganggap prinsip-prinsip ini sebagai daftar periksa yang digunakan setelah sebuah grafik selesai dibuat, pandanglah ART sebagai sebuah lensa yang harus melandasi setiap keputusan dalam visualisasi mulai dari pemilihan jenis grafik hingga penyempurnaan format akhir sebelum dipresentasikan.

Tabel 6.1 Kerangka Kerja ART

Prinsip	Pertanyaan Inti	Hal-hal yang Dicegah
Akurasi	Apakah grafik tersebut merepresentasikan data secara jujur?	Distorsi, sumbu yang terpotong, skala yang menyesatkan
Relevansi	Apakah jenis grafik yang digunakan sesuai dengan pertanyaan analisis yang diajukan?	Penggunaan jenis grafik yang tidak tepat
Transparansi	Apakah audiens dapat langsung memahami apa yang ditampilkan oleh grafik tersebut?	Ambiguitas, label yang hilang, pilihan desain yang tidak disertai penjelasan

Secara keseluruhan, prinsip-prinsip ini membantu memastikan bahwa grafik bersifat jujur, memiliki tujuan yang jelas, dan dapat ditafsirkan dengan baik.

Akurasi

Sebuah grafik dikatakan akurat jika elemen-elemen visualnya merepresentasikan data yang mendasarinya secara tepat. Meskipun hal ini tampak jelas, akurasi sangat mudah untuk dilanggar dan pelanggaran semacam itu tidak selalu dilakukan dengan sengaja. Sebagai contoh, grafik batang yang sumbu y-nya dimulai dari angka di atas nol akan melebih-lebihkan perbedaan antar kategori. Grafik lingkaran (*pie chart*) 3D menciptakan distorsi perspektif yang membuat beberapa bagian tampak lebih besar daripada ukuran aslinya. Grafik area dengan

deret data yang tumpang tindih dapat membuat satu kategori terlihat lebih besar dibandingkan kategori lainnya.

Akurasi adalah fondasi yang tidak bisa ditawar dalam visualisasi yang etis. Kualitas lain seperti relevansi, transparansi, dan bahkan daya tarik estetika menjadi tidak berarti jika grafik tersebut menyajikan data secara keliru. Saat merancang grafik, tanyakan pada diri sendiri: Jika seseorang hanya melihat visualisasinya saja tanpa memeriksa angka-angka yang mendasarinya apakah mereka akan mendapatkan pemahaman yang tepat mengenai data tersebut?

Relevansi

Relevansi berkaitan dengan kesesuaian jenis grafik yang dipilih untuk menjawab pertanyaan analisis yang diajukan. Jenis grafik yang berbeda menyajikan informasi yang berbeda pula:

- Grafik lingkaran (*pie chart*) menunjukkan hubungan bagian terhadap keseluruhan.
- Grafik garis (*line chart*) menggambarkan perubahan dari waktu ke waktu.
- Diagram sebar (*scatter plot*) memperlihatkan hubungan antara dua variabel kontinu.
- Grafik batang (*bar chart*) membandingkan kuantitas antar kategori.

Menggunakan jenis grafik yang salah tidak hanya menghasilkan visualisasi yang kurang efektif, tetapi juga dapat menghalangi audiens untuk memahami data tersebut sama sekali. Perhatikan kesalahan yang sering terjadi: menggunakan grafik lingkaran untuk membandingkan banyak kategori. Persepsi manusia akan kesulitan membandingkan sudut atau panjang busur jika terdapat lebih dari empat atau lima bagian. Data yang sama, jika ditampilkan dalam bentuk grafik batang horizontal yang diurutkan, akan jauh lebih mudah untuk dipahami. Masalah relevansi sering kali tidak disadari oleh pembuat grafik karena mereka sudah memahami pesan yang ingin disampaikan. Namun, audiens mungkin tidak memahaminya.

Transparansi

Sebuah visualisasi dikatakan transparan jika audiens yang tidak memiliki pengetahuan sebelumnya tentang kumpulan data tersebut dapat dengan cepat memahami:

- apa yang ditampilkan oleh grafik,
- bagaimana cara menafsirkannya, dan
- mengapa informasi tersebut penting. Grafik yang transparan biasanya mencakup:
- judul yang jelas dan deskriptif yang menyampaikan poin utama,
- sumbu dengan label yang tepat dan satuan yang sesuai,
- legenda jika diperlukan,
- sumber data jika relevan, dan
- anotasi yang menyoroti wawasan penting.

Transparansi juga mengharuskan kita untuk mengakui apa yang tidak ditampilkan oleh grafik tersebut. Sebagai contoh:

- Jika grafik menyajikan nilai rata-rata, harus diperjelas bahwa mungkin terdapat variasi di dalam data tersebut.
- Jika visualisasi hanya mencakup rentang waktu tertentu, batasan tersebut harus terlihat.

- Jika kategori telah digabungkan atau difilter, keputusan tersebut harus diungkapkan.

Pada akhirnya, transparansi merupakan bentuk penghormatan terhadap audiens, yang memastikan bahwa wawasan disampaikan secara jujur dan jelas.

Diagnosis Singkat

Sebelum menyelesaikan visualisasi apa pun, lakukan diagnosis sederhana berikut:

- **Akurasi**
  Jika angka-angkanya diubah, apakah proporsi visual tetap merepresentasikan data dengan benar?
- **Relevansi**
  Apakah jenis grafik ini menjawab pertanyaan yang diajukan oleh audiens saya secara langsung?
- **Transparansi**
  Bisakah rekan kerja yang tidak familier dengan kumpulan data tersebut memahami grafik ini dalam waktu kurang dari 10 detik?

Jika jawaban untuk salah satu pertanyaan di atas adalah "tidak", revisilah visualisasi tersebut sebelum menyajikannya.

6.4 MEMILIH JENIS GRAFIK YANG TEPAT

Memilih jenis grafik yang sesuai adalah salah satu keputusan terpenting dalam visualisasi data. Grafik yang dipilih dengan kurang tepat tidak hanya mengurangi kejelasan tetapi juga dapat menyamarkan pola, membingungkan audiens, atau bahkan mengarah pada kesimpulan yang keliru. Visualisasi yang efektif dimulai dengan mengajukan satu pertanyaan sederhana:

Pertanyaan analitis apa yang ingin saya jawab?

Sebagian besar visualisasi termasuk ke dalam lima tujuan analitis mendasar. Tabel 6.2 memetakan setiap tujuan ke jenis grafik yang paling sering digunakan dalam Excel.

Tabel 6.2 Pemetaan tujuan analitis ke jenis visualisasi

Tujuan Analitis	Jenis Grafik yang Direkomendasikan	Kapan Digunakan
Perbandingan	Grafik Batang, Grafik Kolom	Membandingkan nilai antar kategori
Distribusi	Histogram, Grafik Kotak	Memahami sebaran, bentuk, dan pencilan
Hubungan	Grafik Sebar, Grafik Gelembung	Mengkaji hubungan antar variabel numerik
Komposisi	Grafik Lingkaran, Grafik Donat, Grafik Batang/Kolom Bertumpuk	Menunjukkan bagaimana suatu keseluruhan dibagi menjadi bagian-bagian
Tren Seiring Waktu	Grafik Garis, Grafik Area	Melacak perubahan dari waktu ke waktu

Subbagian berikut menjelaskan kapan dan mengapa setiap jenis grafik paling tepat untuk digunakan.

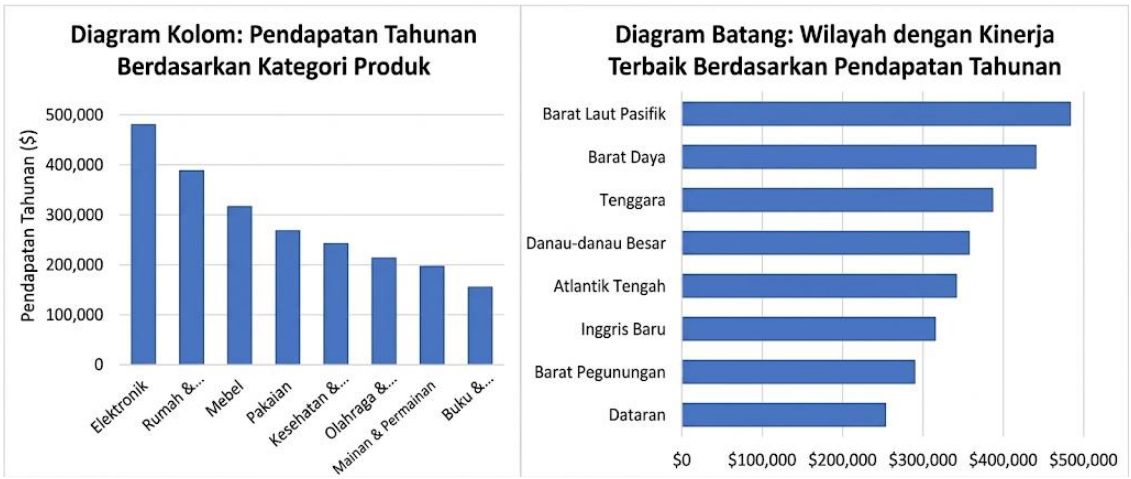
Grafik Perbandingan

Grafik batang merupakan andalan dalam visualisasi data kategorikal. Grafik ini membandingkan nilai di berbagai kategori yang terpisah dan termasuk jenis grafik yang paling intuitif untuk dipahami oleh audiens. Terdapat dua orientasi yang umum digunakan:

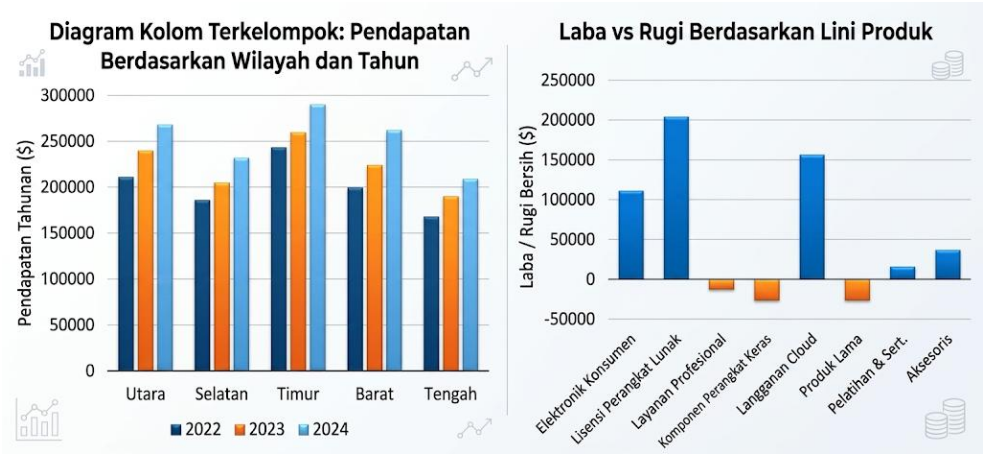
- **Grafik kolom** (batang vertikal)
- **Grafik batang** (batang horizontal)

Keduanya mengkodekan informasi yang sama, tetapi batang horizontal seringkali lebih disukai ketika label kategori panjang atau banyak. Bagan batang sangat berguna untuk:

- Membandingkan total kategori
  Contoh: Pendapatan Tahunan di seluruh kategori produk.
- Memberi peringkat kategori
  Contoh: Wilayah berkinerja terbaik berdasarkan Pendapatan.
- Membandingkan kelompok dalam kategori
  Bagan batang berkelompok memungkinkan perbandingan di seluruh kategori dan subkelompok (misalnya, penjualan berdasarkan wilayah dan tahun).
- Menunjukkan nilai positif dan negatif
  Batang dapat memanjang di atas dan di bawah garis nol, sehingga berguna untuk perbandingan laba vs. rugi.



Gambar 6.1 Bagan Batang dan Bagan Kuster



Gambar 6.2: Diagram batang berkelompok (*clustered bar charts*) dan diagram batang yang menampilkan nilai positif serta negatif

Diagram Batang Bertumpuk (*Stacked Bar Charts*)

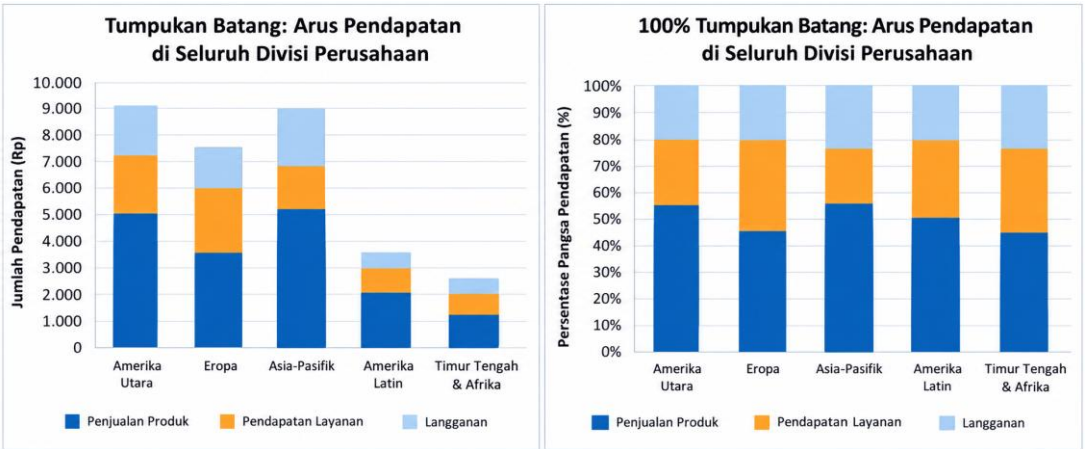
Diagram batang bertumpuk menggambarkan nilai total sekaligus komposisi di dalamnya. Keseluruhan batang mewakili nilai total, sementara segmen-segmen di dalamnya menunjukkan bagaimana total tersebut terbagi ke dalam berbagai komponen. Gunakan diagram batang bertumpuk jika Anda ingin menunjukkan:

- bagaimana kategori-kategori berkontribusi terhadap suatu total
- bagaimana proporsi bervariasi di antara kelompok-kelompok yang berbeda

Diagram batang bertumpuk 100% (100% *stacked bar chart*) berguna untuk membandingkan proporsi relatif, bukan nilai total absolut.

Contoh skenario

Seorang analis ritel ingin membandingkan kontribusi pendapatan di berbagai divisi perusahaan. Diagram batang bertumpuk dapat menyoroti divisi mana yang menghasilkan total pendapatan tertinggi, sementara diagram batang bertumpuk 100% menunjukkan perbandingan proporsi pendapatan setiap divisi relatif terhadap total keseluruhan.

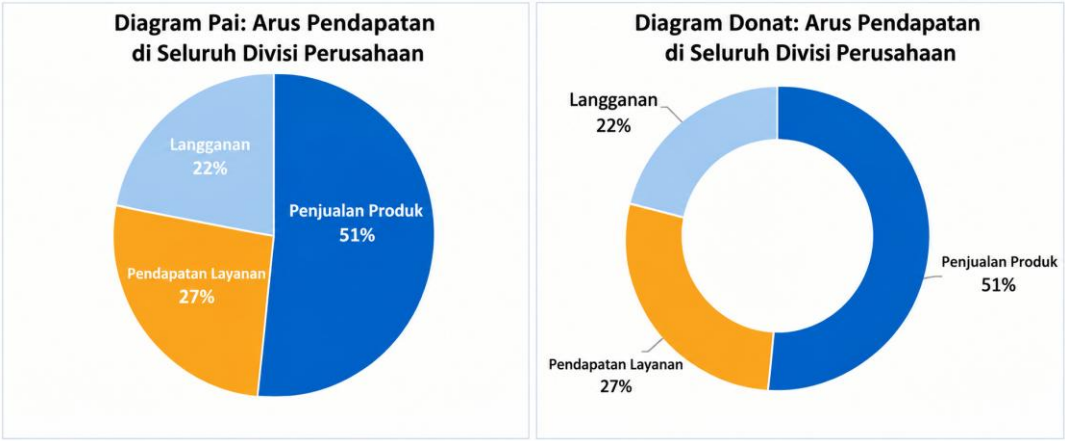


Gambar 6.3 Grafik batang bertumpuk (*stacked bar*) dan grafik batang bertumpuk 100% (100% *stacked bar*)

Grafik Lingkaran (*Pie Chart*) dan Grafik Donat (*Doughnut Chart*)

Grafik lingkaran menggambarkan hubungan bagian terhadap keseluruhan, yang menunjukkan bagaimana suatu total dibagi ke dalam berbagai kategori. Namun, penggunaannya harus dilakukan dengan hati-hati. Persepsi manusia sulit untuk membandingkan sudut dan panjang busur secara akurat, terutama jika terdapat banyak kategori. Grafik lingkaran paling efektif digunakan ketika:

- terdapat empat kategori atau kurang
- hubungan bagian terhadap keseluruhan merupakan inti dari pesan yang ingin disampaikan
- perbandingan yang presisi tidak diperlukan



Gambar 6.4 Diagram lingkaran (*pie chart*) dan diagram donat (*doughnut chart*)

Ketika kategori bertambah banyak atau perbandingan menjadi penting, diagram batang bertumpuk (*stacked bar chart*) biasanya lebih efektif karena memungkinkan pengamat untuk membandingkan panjang batang alih-alih sudut, yang secara perseptual lebih mudah dilakukan.

Diagram Distribusi

Memahami bagaimana nilai data terdistribusi merupakan hal yang esensial untuk analisis statistik. Diagram distribusi menunjukkan sebaran, kemiringan (*skewness*), klaster, dan pencilan (*outlier*).

Histogram

Histogram menampilkan seberapa sering nilai muncul dalam rentang tertentu yang disebut sebagai *bin* (kelompok rentang). Histogram sangat berguna untuk menelaah:

- bentuk distribusi
- pemusatan data
- kemiringan (*skewness*)
- puncak ganda (bimodalitas)

Berbagai pola dapat muncul dalam histogram:

- Distribusi normal tampak berbentuk lonceng dan simetris

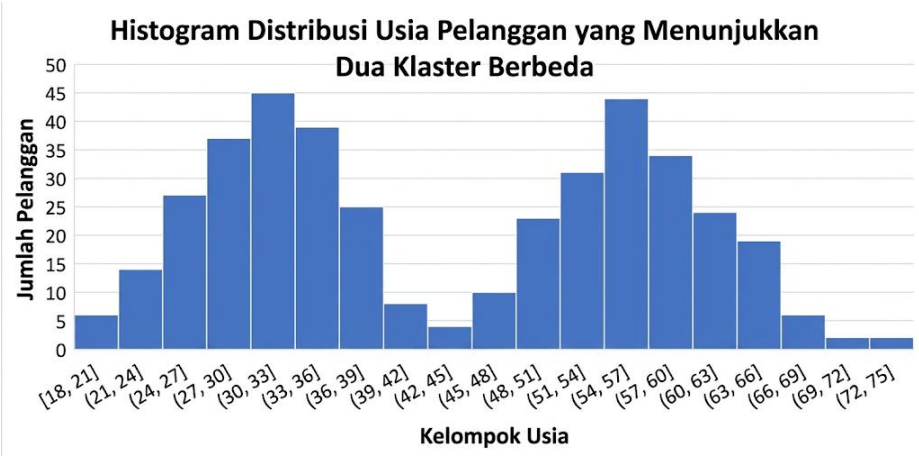
- Distribusi yang miring (*skewed*) menunjukkan ekor yang panjang ke satu arah
- Distribusi bimodal menunjukkan dua puncak
- Distribusi seragam (*uniform*) tampak relatif datar

Pemilihan lebar *bin* sangatlah penting:

- Jumlah *bin* yang terlalu sedikit dapat menyembunyikan pola-pola penting
- Jumlah *bin* yang terlalu banyak dapat memunculkan derau acak (*random noise*)

Contoh skenario

Seorang analis pemasaran mempelajari usia pelanggan. Histogram menunjukkan dua klaster (distribusi bimodal), yang mengindikasikan adanya dua segmen pelanggan yang berbeda.



Gambar 6.5 Grafik Histogram

Plot KDE sebagai Alternatif yang Lebih Mulus dibandingkan Histogram

Meskipun histogram sangat baik untuk memvisualisasikan distribusi keseluruhan suatu variabel, tampilannya sangat bergantung pada lebar dan batas interval (bin). Plot *Kernel Density Estimate* (KDE) menawarkan representasi distribusi yang lebih mulus dan berkelanjutan. Walaupun Excel tidak menyediakan fitur plot KDE secara bawaan, perangkat seperti Python atau R dapat membuatnya. Hal ini membantu analis melihat bentuk distribusi yang sebenarnya tanpa dipengaruhi oleh pemilihan lebar interval.

Plot Kotak dan Kumis (*Box and Whisker Plot*)

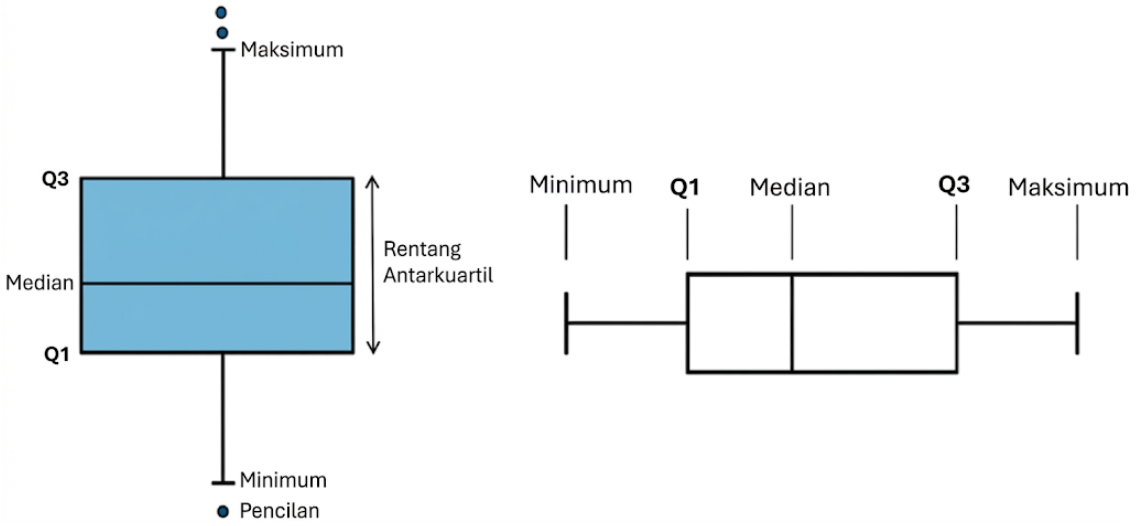
Box plot menyajikan visualisasi ringkas dari ringkasan lima angka (*five-number summary*):

- Nilai minimum
- Kuartil pertama (Q1)
- Median (Q2)
- Kuartil ketiga (Q3)
- Nilai maksimum

Box plot juga menyoroti pencilan (*outliers*), sehingga sangat berguna untuk membandingkan distribusi antar kelompok. Sebuah *box plot* terdiri dari beberapa komponen:

- **Kotak (Q1–Q3):** Mewakili rentang antarkuartil (*interquartile range* atau IQR)—yaitu 50% data di bagian tengah.

- **Garis Median:** Garis di dalam kotak menandai median dan menunjukkan tendensi sentral.
- **Kumis (*whiskers*):** Memanjang hingga ke nilai terkecil dan terbesar dalam rentang yang bukan merupakan pencilan.
- **Pencilan (*outliers*):** Titik-titik yang berada di luar kumis mewakili pengembangan teknologi masa depan yang sangat ekstrem.



Gambar 6.6 Boxplot Vertikal dan Horizontal

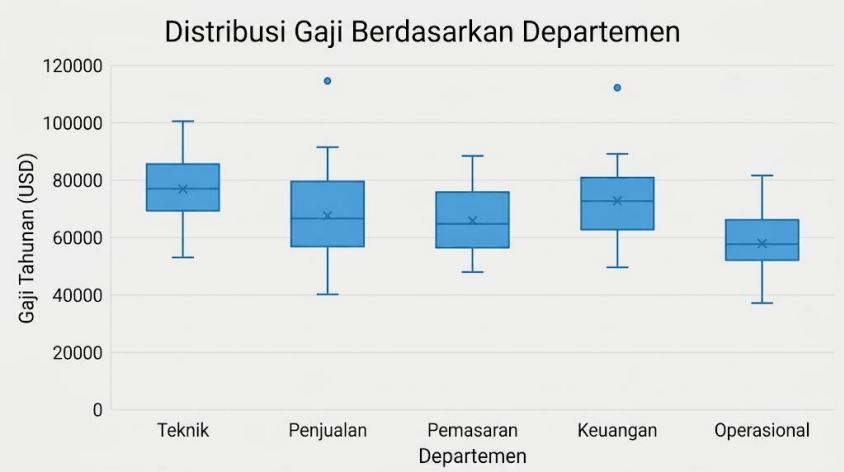
Perbandingan boxplot yang efektif bergantung pada evaluasi sistematis dari penanda statistik inti ini:

Tabel 6.3 Analisis Boxplot Komparatif

Fitur	Hal yang Perlu Dibandingkan	Mengapa Hal Ini Penting
Median	Posisi garis (nilai), tinggi kotak (diagram kotak vertikal)	Perbedaan dalam tendensi sentral. Kelompok mana yang memiliki nilai tipikal lebih tinggi?
IQR (ukuran kotak)	Lebar kotak (diagram kotak horizontal)	Perbedaan dalam variabilitas. IQR yang lebih besar berarti data lebih tersebar.
Garis whisker	Panjang	Perbedaan dalam sebaran di luar rentang pusat. Rentang yang lebih pendek menunjukkan konsistensi yang lebih tinggi.
Pencilan	Jumlah dan posisi	Apakah terdapat kasus ekstrem? Kelompok mana yang memiliki lebih banyak kasus ekstrem?
Kemiringan (*skewness*)	Lokasi median di dalam kotak serta panjang garis (*whisker*)	Mendeteksi asimetri. Berguna untuk memahami perilaku data.

Contoh skenario

Seorang analis SDM membandingkan distribusi gaji di berbagai departemen. Diagram kotak (*box plot*) menunjukkan departemen mana yang memiliki median gaji lebih tinggi dan departemen mana yang menunjukkan variabilitas lebih besar.



Gambar 6.7 Analisis *Box Plot* Komparatif – distribusi gaji berdasarkan departemen

Keterbatasan *box plot*

Meskipun *box plot* memberikan wawasan ringkasan yang berharga, grafik ini tidak memperlihatkan bentuk distribusi yang sebenarnya. *Box plot* menampilkan kuartil dan potensi pencilan (*outlier*), namun tidak menunjukkan bagaimana titik-titik data terdistribusi di dalam kuartil-kuartil tersebut.

Plot biola (*violin plot*) sebagai alternatif

Meskipun Excel tidak mendukung plot biola secara bawaan, perangkat yang lebih canggih seperti Python atau R menyediakannya. Plot biola menggabungkan plot kotak (box plot) dengan kurva kepadatan, sehingga Anda dapat melihat statistik ringkasan sekaligus bentuk distribusi data yang mendasarinya.

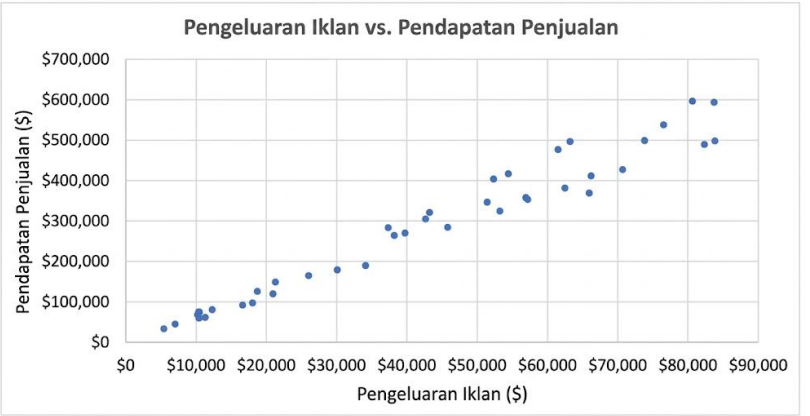
Diagram Sebar (*Scatter Plot*)

Diagram sebar memvisualisasikan hubungan antara dua variabel kontinu dengan menampilkan pengembangan teknologi masa depan sebagai titik-titik pada bidang dua dimensi. Setiap titik mewakili satu pengembangan teknologi masa depan. Diagram sebar dapat menunjukkan:

- Hubungan positif: Kedua variabel meningkat secara bersamaan.
- Hubungan negatif: Satu variabel meningkat sementara variabel lainnya menurun.
- Tidak ada hubungan: Titik-titik tersebar secara acak.
- Hubungan nonlinier: Titik-titik membentuk pola melengkung atau kompleks, bukan garis lurus.
- Pencilan (*outlier*) dan klaster (kelompok data).

Contoh skenario

Sebuah perusahaan meneliti apakah pengeluaran untuk iklan memiliki hubungan dengan pendapatan penjualan. Diagram sebar dapat langsung memperlihatkan apakah peningkatan pengeluaran berkaitan dengan penjualan yang lebih tinggi.



Gambar 6.8 Diagram sebar (*scatter plot*) antara pengeluaran iklan dan pendapatan penjualan

Diagram sebar juga menjadi dasar bagi analisis korelasi dan regresi, yang akan dibahas pada Bab 7.

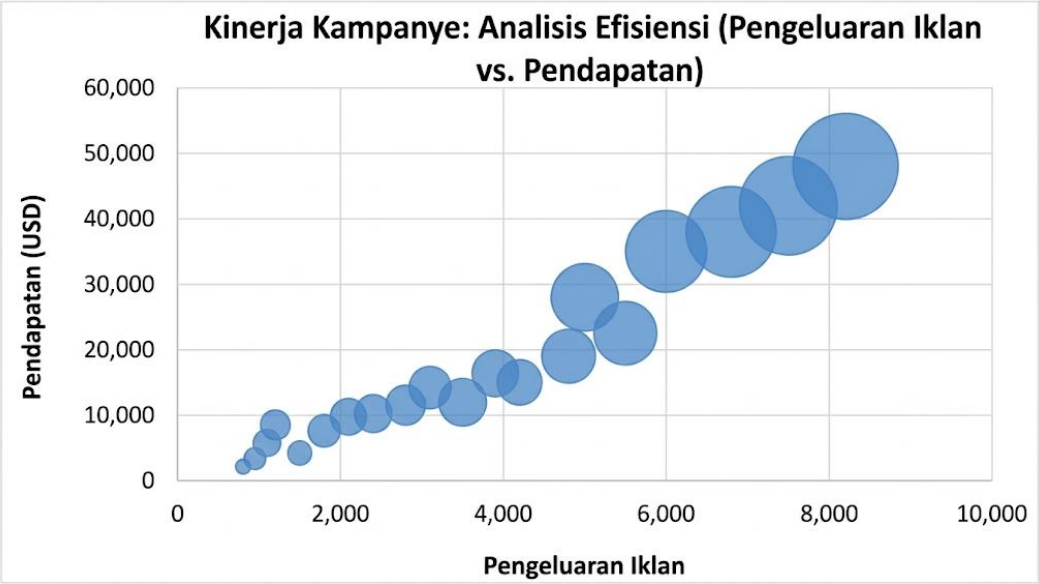
Diagram Gelembung (*Bubble Chart*)

Diagram gelembung merupakan pengembangan dari diagram sebar dengan menambahkan variabel ketiga melalui ukuran gelembung. Hal ini memungkinkan visualisasi tiga variabel secara bersamaan:

- Sumbu X – variabel pertama
- Sumbu Y – variabel kedua
- Ukuran gelembung – variabel ketiga

Contoh skenario

Seorang pemasar digital memetakan pengeluaran iklan (sumbu X) terhadap pendapatan (sumbu Y) untuk mengevaluasi kinerja berbagai kampanye media sosial. Untuk menambahkan dimensi data ketiga, ukuran gelembung digunakan untuk merepresentasikan jumlah konversi (total penjualan yang dihasilkan).

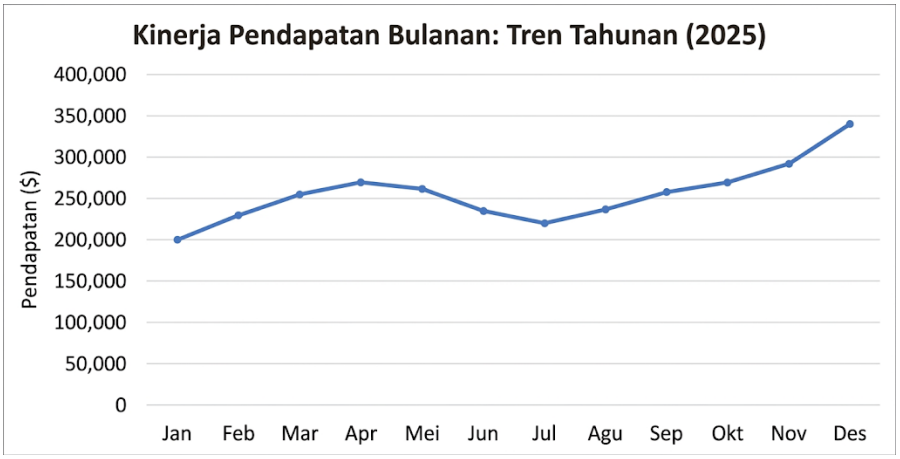


Gambar 6.9 Diagram gelembung (*bubble chart*)

Format ini sangat efektif karena mampu menunjukkan tingkat efisiensi. Gelembung berukuran besar yang terletak di posisi tinggi pada sumbu y dan di posisi rendah pada sumbu x merepresentasikan sebuah "tambang emas" yakni kampanye yang menghasilkan pendapatan tinggi serta banyak konversi dengan pengeluaran yang sangat minim.

Diagram Garis (*Line Chart*)

Diagram garis sangat ideal untuk memvisualisasikan perubahan dari waktu ke waktu atau sepanjang skala yang berkelanjutan. Diagram ini menghubungkan berbagai pengembangan teknologi masa depan secara berurutan, sehingga tren dapat terlihat dengan jelas seketika.



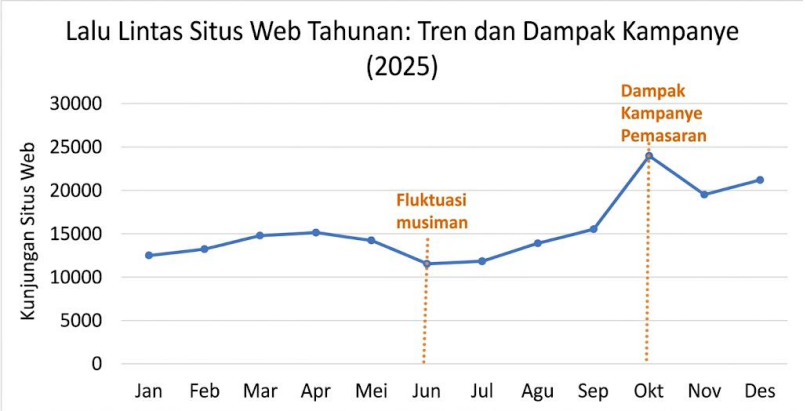
Gambar 6.10 Grafik deret waktu – Tren pendapatan tahunan

Grafik garis sangat baik untuk menunjukkan:

- tren jangka panjang
- pola musiman
- perubahan mendadak

– pergerakan komparatif di antara beberapa rangkaian data

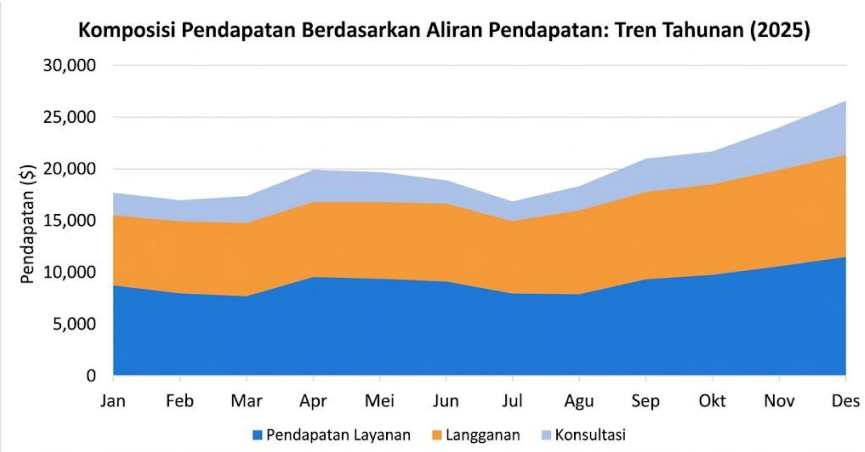
Namun, grafik garis sebaiknya hanya digunakan jika sumbu horizontal mewakili urutan yang bermakna, seperti waktu. Menghubungkan kategori yang tidak saling berkaitan dengan garis akan memberikan kesan keliru mengenai adanya kesinambungan, sehingga melanggar prinsip akurasi visualisasi. Contoh scenario: Sebuah perusahaan melacak lalu lintas situs web bulanan. Grafik garis dapat dengan cepat memperlihatkan tren pertumbuhan, fluktuasi musiman, serta dampak dari kampanye pemasaran.



Gambar 6.11 Pendapatan terus meningkat seiring fluktuasi musiman, dan dampak pasar modal.

Grafik Area

Grafik area mirip dengan grafik garis tetapi mengisi ruang di bawah garis. Grafik ini berguna untuk menekankan total kumulatif atau besaran keseluruhan, terutama ketika membandingkan beberapa komponen dari waktu ke waktu. Grafik area bertumpuk bekerja dengan baik dengan sejumlah kecil kategori; namun, menambahkan terlalu banyak seri dapat menciptakan area yang tumpang tindih yang membuat interpretasi menjadi sulit.



Gambar 6.12 Grafik area yang membandingkan pertumbuhan pendapatan kumulatif.

6.6 JENIS GRAFIK UMUM LAIN YANG PERLU DIKETAHUI

Jenis grafik yang dibahas sebelumnya mewakili alat inti visualisasi data, dan akan menjawab sebagian besar pertanyaan analitis yang ditemui dalam praktik. Namun, analis

sering menghadapi situasi di mana grafik yang lebih khusus memberikan wawasan yang lebih jelas. Seiring bertambahnya ukuran dataset dan pertanyaan analitis menjadi lebih kompleks, jenis visualisasi tambahan dapat membantu mengungkapkan pola yang mungkin tidak ditangkap secara efektif oleh grafik standar. Mengenal grafik-grafik ini memperluas kemampuan Anda untuk mengkomunikasikan hubungan, struktur, dan proses yang kompleks dalam data.

Jenis Grafik Tambahan dan Kegunaannya

Tabel 6.4 Jenis Grafik Tambahan

Jenis Grafik	Apa yang Ditunjukkannya	Kapan Hal Ini Berguna
Peta Panas (*Heatmap*)	Intensitas dengan kode warna dalam dua dimensi	Mengidentifikasi pola, klaster, atau anomali dalam kumpulan data yang besar, seperti matriks korelasi
Peta Pohon (*Treemap*)	Persegi panjang bersarang yang merepresentasikan proporsi hierarkis	Memvisualisasikan hubungan bagian terhadap keseluruhan ketika kategori mencakup beberapa subkategori
Grafik Corong (*Funnel Chart*)	Pengurangan bertahap melalui tahapan-tahapan berurutan	Melacak tahapan konversi, alur penjualan, atau titik penghentian dalam suatu proses
Grafik Air Terjun (*Waterfall Chart*)	Kontribusi positif dan negatif secara berurutan terhadap suatu total	Menjelaskan bagaimana nilai awal berubah langkah demi langkah hingga mencapai hasil akhir
Grafik Radar (*Spider Chart)*	Perbandingan multidimensi di sepanjang sumbu radial	Membandingkan profil kinerja di berbagai variabel (contoh: keterampilan, atribut produk)
Bagan Gantt	Tugas-tugas yang disusun sepanjang garis waktu	Merencanakan dan memantau jadwal, ketergantungan, serta kemajuan proyek
Diagram Sankey	Aliran kuantitas antar tahapan atau kategori	Menunjukkan bagaimana sumber daya, pengguna, atau nilai bergerak melalui suatu sistem
Peta Koroplet	Wilayah geografis berdasarkan nilai-nilai tertentu	Memvisualisasikan pola spasial, seperti kepadatan penduduk atau kinerja wilayah

Grafik-grafik khusus ini memungkinkan analis untuk melangkah lebih jauh daripada sekadar perbandingan atau tren sederhana. Grafik tersebut sangat berguna ketika data melibatkan hierarki, proses, hubungan geografis, atau variabel multidimensi. Banyak platform visualisasi modern termasuk Power BI, Tableau, pustaka Python, dan paket R menyediakan dukungan bawaan untuk jenis-jenis grafik ini. Beberapa di antaranya juga tersedia di Excel atau melalui perangkat tambahan (*add-ins*) dan templat khusus.

Menyatukan Semuanya

Memilih jenis grafik yang tepat merupakan perpaduan antara keputusan teknis dan keterampilan berkomunikasi. Visualisasi yang efektif menyeimbangkan akurasi analitis dengan kejelasan serta kesederhanaan. Saat memilih grafik, mulailah dengan jenis-jenis grafik dasar seperti diagram batang, diagram garis, histogram, dan diagram sebar (*scatter plot*). Seiring dengan meningkatnya kompleksitas analisis Anda, gunakan jenis grafik tambahan yang lebih mampu mencerminkan struktur dan kerumitan data Anda. Pada akhirnya, tujuan visualisasi tetaplah sama: mengubah data mentah menjadi wawasan yang jelas dan bermakna. Grafik yang dipilih dengan tepat memungkinkan audiens untuk menangkap pola, hubungan, dan tren dengan cepat, hal-hal yang mungkin akan tetap tersembunyi jika hanya disajikan dalam bentuk tabel angka.

6.5 ALAT VISUALISASI EXCEL

Meskipun platform khusus seperti Python, R, Power BI, dan Tableau menawarkan kemampuan visualisasi yang luas, Microsoft Excel tetap menjadi salah satu alat yang paling banyak digunakan untuk visualisasi data di lingkungan profesional. Aksesibilitasnya, kemudahan penggunaannya, serta integrasi yang erat dengan data menjadikannya sangat berharga, baik untuk analisis eksploratif maupun untuk mengomunikasikan wawasan kepada audiens non-teknis. Di banyak organisasi, Excel menjadi alat pertama yang digunakan analis untuk memeriksa data dan membuat ringkasan visual. Oleh karena itu, mengembangkan keterampilan visualisasi yang kuat di Excel memberikan landasan praktis yang dapat diterapkan saat menggunakan alat-alat yang lebih canggih.

Keunggulan Visualisasi Excel

Excel sangat efektif untuk tugas-tugas berikut:

- **Analisis eksploratif cepat:** Grafik dapat dibuat dalam hitungan detik, sehingga memungkinkan analis untuk segera mengidentifikasi tren, pola, dan anomali dalam data.
- **Dasbor interaktif:** Fitur seperti *PivotTable*, *PivotChart*, dan *slicer* memungkinkan pengguna untuk mengeksplorasi data secara dinamis melalui penyaringan (*filtering*) dan segmentasi informasi.
- **Presentasi profesional:** Excel menyediakan tema grafik, gaya, dan alat pemformatan bawaan yang membantu menghasilkan grafik dengan tampilan rapi dan profesional, yang cocok untuk laporan maupun presentasi. Aksesibilitas universal: Berkas Excel dapat dibuka dan dipahami oleh berbagai pengguna, sehingga memudahkan pembagian analisis di seluruh tim maupun organisasi.
- **Integrasi data dinamis:** Grafik terhubung langsung dengan data sumbernya. Saat data berubah, grafik akan diperbarui secara otomatis.

Pivot Chart dan Slicer

Dua fitur ini menjadikan Excel sangat berguna untuk eksplorasi data secara interaktif.

- **Pivot Chart:** Pivot Chart adalah representasi grafis dari data yang diringkas dalam PivotTable. Karena grafik ini terhubung dengan PivotTable, setiap perubahan pada struktur atau filter tabel akan langsung memperbarui visualisasinya.
- **Slicer:** Slicer adalah alat penyaring (*filter*) interaktif yang memungkinkan pengguna untuk memilah data dengan cepat di dalam PivotTable, PivotChart, atau tabel Excel. Alih-alih menggunakan filter menu tarik-turun (*dropdown*), pengguna cukup mengeklik tombol yang mewakili kategori tertentu untuk menyaring data secara visual.

Slicer membuat dasbor menjadi jauh lebih intuitif dan sering digunakan dalam pelaporan intelijen bisnis (*business intelligence*).

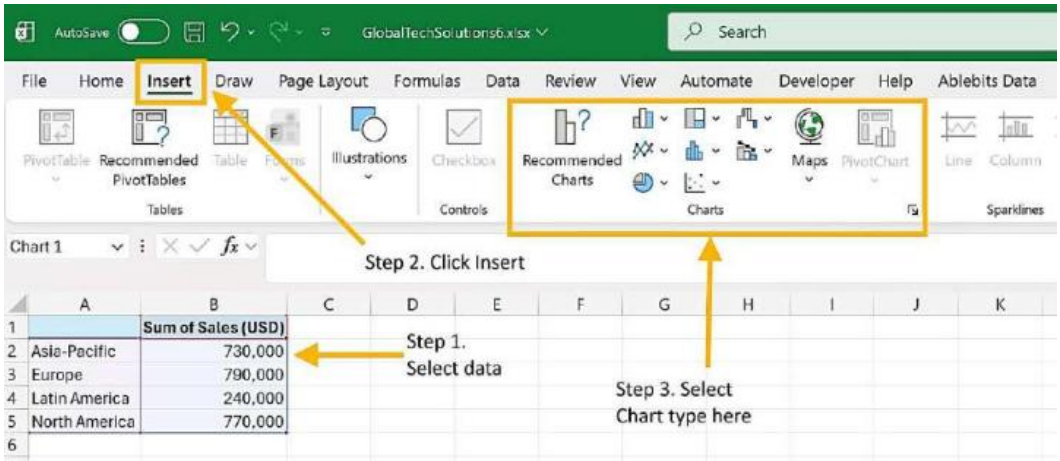
Membuat Grafik Dasar di Excel

Meskipun Excel mendukung banyak jenis grafik, proses pembuatannya mengikuti alur kerja yang konsisten.

Prosedur Umum

Dengan asumsi data Anda telah disusun dalam baris dan kolom yang berurutan serta memiliki judul kolom yang jelas di Excel:

1. Pilih Data: Sorot sel-sel yang berisi data yang ingin Anda visualisasikan.
2. Buka tab "*Insert*" (Sisipkan) pada Ribbon.
3. Pilih jenis grafik yang diinginkan dari grup "*Charts*" (Grafik).
4. Sesuaikan Grafik: Gunakan tab "*Chart Design*" (Desain Grafik) dan "Format" untuk mengatur judul, label, warna, serta tata letak grafik.



Gambar 6.13 Membuat grafik dasar di Excel

Jenis-jenis Grafik Umum dan Lokasi Penemuannya

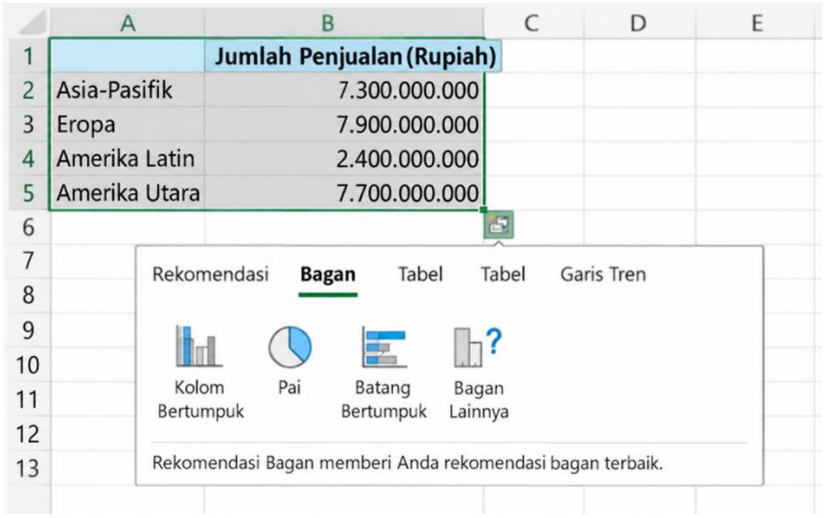
Tabel 6.5 Jalur navigasi Excel untuk membuat grafik

Jenis Grafik	Kegunaan Utama	Jalur Navigasi Excel
Grafik Batang / Kolom	Membandingkan nilai antar kategori	Sisipkan → Bagan → Sisipkan Bagan Kolom atau Batang. Pilihan: Berkelompok, Bertumpuk, Bertumpuk 100%

Grafik Lingkaran / Donat	Menunjukkan bagian dari keseluruhan	Sisipkan → Bagan → Sisipkan Bagan Pai atau Donat → pilih Pai atau Donat
Grafik Garis	Menunjukkan tren sepanjang waktu atau variabel kontinu	Sisipkan → Bagan → Sisipkan Bagan Garis atau Area → pilih Garis
Grafik Area	Menampilkan total kumulatif dari waktu ke waktu	Sisipkan → Bagan → Sisipkan Bagan Garis atau Area → pilih Area
Diagram Sebar (*Scatter Plot*)	Memvisualisasikan hubungan antar variabel numerik	Sisipkan → Bagan → Sisipkan Bagan Sebar (X,Y) atau Gelembung → Sebar
Diagram Gelembung (*Bubble Chart*)	Menunjukkan hubungan antara dua variabel numerik (x, y) dengan menambahkan variabel ketiga sebagai ukuran gelembung (*bubble size*)	Sisipkan → Bagan → Sisipkan Bagan Sebar (X, Y) atau Gelembung → Gelembung
Histogram	Menampilkan distribusi frekuensi	Sisipkan → Bagan → Sisipkan Bagan Statistik → Histogram
Diagram Kotak-Garis (Box & Whisker Plot)	Menunjukkan distribusi, kuartil, dan pencilan (*outlier*)	Sisipkan → Bagan → Sisipkan Bagan Statistik → Kotak & Kumis

Alat Analisis Cepat (*Quick Analysis Tool*)

Excel juga menyediakan jalan pintas untuk visualisasi eksploratif yang cepat. Setelah memilih data Anda, tekan Ctrl + Q untuk membuka alat Analisis Cepat (*Quick Analysis tool*). Sebuah jendela seperti yang terlihat pada gambar di bawah ini akan muncul.



Gambar 6.14 Panduan fitur *Quick Analysis* (Analisis Cepat) di Excel

Secara *default*, opsi *Formatting* (Pemformatan) terpilih. Beralihlah ke *Charts* (Grafik) dan pilih salah satu jenis grafik yang disarankan untuk memvisualisasikan data Anda secara instan. Fitur ini memungkinkan Anda untuk mengeksplorasi hal-hal berikut secara langsung:

- Grafik yang disarankan

- Pemformatan bersyarat (*conditional formatting*)
- Total dan ringkasan
- Tabel dan *pivot table*

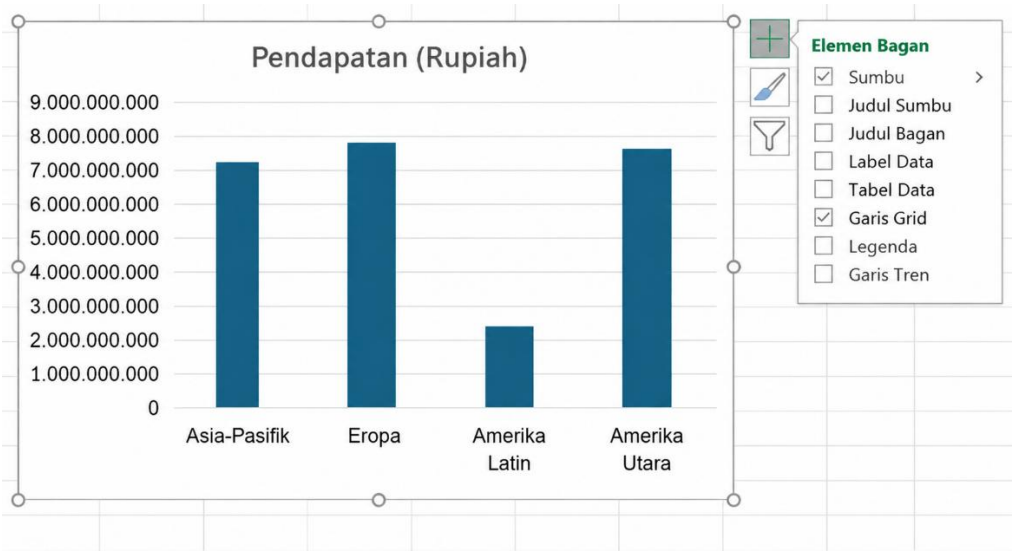
Quick Analysis sangat berguna saat melakukan analisis data eksploratif awal.

Menyesuaikan Visualisasi di Excel

Setelah grafik dibuat, Excel menyediakan beberapa alat untuk meningkatkan kejelasan dan tampilan grafik tersebut dengan cepat. Saat sebuah grafik dipilih, tiga ikon akses cepat akan muncul di sudut kanan atas area grafik. Ikon-ikon ini memungkinkan Anda untuk memodifikasi grafik tanpa harus menelusuri menu pada *ribbon*.

Ikon-ikon tersebut adalah:

1. *Chart Elements* (Elemen Grafik)
2. *Chart Styles* (Gaya Grafik)
3. *Chart Filters* (Filter Grafik)



Gambar 6.15 Alat akses cepat untuk grafik di Excel

Anggaplah alat-alat ini sebagai panel kontrol untuk menyempurnakan visualisasi Anda.

1. Elemen Grafik (Chart Elements)

Tujuan: Menambah, menghapus, atau menyesuaikan komponen struktural grafik. Contohnya meliputi:

- Judul Grafik (*Chart Title*)
- Judul Sumbu (*Axis Titles*)
- Label Data (*Data Labels*)
- Legenda (*Legend*)
- Garis Kisi (*Gridlines*)

Cara penggunaan:

- Klik ikon tanda tambah (+).
- Pilih elemen yang ingin ditampilkan.

- Gunakan tanda panah di samping setiap elemen untuk memilih opsi tata letak.

Praktik terbaik: Selalu sertakan judul grafik dan label sumbu yang jelas agar grafik Anda dapat dipahami tanpa memerlukan penjelasan tambahan.

2. Gaya Grafik (*Chart Styles*)

Tujuan: Menerapkan tema visual untuk meningkatkan tampilan grafik.

Cara penggunaan:

1. Klik ikon kuas.
2. Pilih desain tata letak melalui tab Gaya (*Style*).
3. Pilih palet warna melalui tab Warna (*Color*).

Praktik terbaik: Gunakan warna yang konsisten pada berbagai grafik dalam laporan yang sama untuk menjaga keselarasan visual.

3. Filter Grafik (Chart Filters)

Tujuan: Mengontrol kategori data atau seri mana saja yang ditampilkan dalam grafik.

Cara penggunaan:

1. Klik ikon corong (*filter*).
2. Pilih atau hapus pilihan kategori maupun nilai tertentu.
3. Klik Terapkan (*Apply*) untuk memperbarui grafik.

Fitur ini sangat berguna saat menyajikan data dan berfokus pada kelompok atau periode waktu tertentu.

Penyesuaian Grafik yang Penting

Judul dan Label

- Tambahkan judul grafik deskriptif yang menjelaskan pesan grafik tersebut dengan jelas.
- Sertakan judul sumbu untuk mengidentifikasi satuan pengukuran.
- Tambahkan label data jika nilai yang tepat diperlukan.

Untuk menambahkan label data:

- Gunakan elemen grafik (+), atau
- Klik kanan pada seri data → Tambahkan Label Data (*Add Data Labels*)

Warna dan Pemformatan

Pemilihan warna yang efektif akan meningkatkan keterbacaan. Praktik yang disarankan:

- Gunakan tema grafik bawaan Excel.
- Sorot data penting dengan warna yang kontras.
- Pertahankan skema warna yang konsisten di seluruh grafik.

Hindari kombinasi warna yang bermasalah, seperti:

- Merah vs. Hijau
- Hijau vs. Cokelat
- Biru vs. Ungu
- Warna-warna yang sangat terang sehingga mengurangi kontras

Kombinasi-kombinasi ini mungkin sulit dibedakan, terutama bagi pembaca yang mengalami buta warna.

Sumbu dan Penentuan Skala (*Scaling*)

Penentuan skala sumbu yang tepat memastikan grafik merepresentasikan data secara akurat. Untuk mengubah pengaturan sumbu:

- Klik kanan pada sumbu → *Format Axis* (Format Sumbu)

Anda dapat menyesuaikan:

- Nilai minimum dan maksimum
- Interval tanda skala (*tick mark*)
- Label sumbu

Fitur Grafik Lanjutan

Excel juga mendukung beberapa penyempurnaan visualisasi tingkat lanjut.

- Garis Tren (*Trendlines*): Digunakan untuk menunjukkan pola umum dalam diagram sebar (*scatter plot*) atau grafik deret waktu (*time-series*). Klik kanan pada seri data → *Add Trendline* (Tambahkan Garis Tren)
- Bilah Galat (*Error Bars*): Berguna untuk menampilkan variabilitas atau ketidakpastian dalam pengukuran.
- Tabel Data (*Data Tables*): Menampilkan nilai numerik dasar di bawah grafik.

Sumbu Sekunder (*Secondary Axes*): Memungkinkan perbandingan dua variabel dengan skala berbeda dalam satu grafik yang sama. Gunakan sumbu sekunder dengan hati-hati, karena terkadang dapat menciptakan perbandingan yang menyesatkan.

Opsi Pemformatan Tambahan

Saat Anda memilih grafik, dua tab tambahan akan muncul pada Ribbon:

- *Chart Design* (Desain Grafik)
- Format

Tab-tab ini menyediakan akses ke alat penataan gaya dan pemformatan tingkat lanjut. Untuk menyesuaikan elemen grafik tertentu:

1. Klik elemen grafik tersebut (judul, sumbu, batang, dan lain lain.).
2. Pilih *Format Selection* (Format Pilihan) dari tab Format.

   Atau:

Klik kanan pada elemen grafik mana pun → *Format Chart Area* (Format Area Grafik)

Panel pemformatan akan muncul di sisi kanan layar, yang memberikan kendali mendetail atas tata letak, warna, batas (*border*), dan pemformatan teks.

Membuat Histogram dengan Data Analysis ToolPak

Selain grafik statistik bawaan, Excel dapat membuat histogram menggunakan Data Analysis ToolPak.

Untuk membuat histogram:

Data → *Data Analysis* (Analisis Data) → Histogram

Alat ini memungkinkan Anda untuk menentukan rentang bin (kelompok data) sesuai keinginan serta menghasilkan tabel frekuensi sekaligus grafik histogram.

6.6 MERANCANG UNTUK KEJELASAN DI EXCEL

Memilih jenis grafik yang tepat memenuhi komponen Relevansi (*Relevance*) dari kerangka kerja ART yang diperkenalkan sebelumnya dalam bab ini. Namun, memilih grafik yang tepat hanyalah langkah awal. Grafik dengan format yang buruk tetap dapat menyesatkan pembaca atau mengaburkan wawasan penting. Merancang grafik dengan cermat akan memenuhi dua prinsip lain dari kerangka kerja ART:

- Akurasi (*Accuracy*) — memastikan grafik merepresentasikan data yang mendasarinya secara tepat.
- Transparansi (*Transparency*) — membuat informasi mudah ditafsirkan oleh pembaca.

Bagian ini memperkenalkan beberapa prinsip desain praktis yang dapat mengubah grafik dasar menjadi visualisasi yang jelas, profesional, dan dapat diandalkan secara analitis.

Rasio Data-Tinta (*Data-Ink Ratio*)

Salah satu prinsip paling berpengaruh dalam visualisasi data adalah rasio data-tinta (*data-ink ratio*), sebuah konsep yang diperkenalkan oleh pakar statistik dan perancang informasi, Edward Tufte. Rasio data-tinta merujuk pada proporsi elemen visual dalam grafik yang merepresentasikan data aktual dibandingkan dengan elemen yang hanya berfungsi sebagai dekorasi.

Gagasan utamanya sederhana:

'Maksimalkan proporsi tinta yang digunakan untuk menampilkan data dan minimalkan elemen visual yang tidak penting'. Grafik yang memuat dekorasi berlebihan seperti garis tepi yang tebal, latar belakang yang tidak perlu, atau efek tiga dimensi dapat mengalihkan perhatian pembaca dari informasi yang paling penting. Di Excel, meningkatkan rasio data-tinta biasanya melibatkan langkah-langkah berikut:

- Hapus warna latar belakang (*fill*) dari area grafik (*chart area*) dan area plot (*plot area*) gunakan warna putih atau hilangkan warnanya (*no fill*).
- Hapus garis kisi (*gridlines*) yang tidak perlu, atau gunakan warna abu-abu sangat muda jika memang diperlukan.
- Hapus garis tepi (*border*) di sekeliling grafik dan legenda.
- Hapus legenda jika hanya terdapat satu seri data dan judul grafik sudah menjelaskan data tersebut.
- Hindari efek visual seperti bayangan, format 3D, dan pengisian warna gradasi (*gradient fills*).

Dengan menyederhanakan grafik, perhatian pengamat akan secara alami tertuju pada data itu sendiri, bukan pada elemen-elemen dekoratif.

Judul, Label, dan Anotasi

Elemen teks memainkan peran krusial dalam membuat grafik mudah dipahami. Judul yang ditulis dengan baik dan label sumbu yang jelas memastikan pembaca dapat menafsirkan grafik tanpa memerlukan penjelasan tambahan. Judul grafik yang efektif harus menyampaikan kesimpulan atau poin utama, bukan sekadar mendeskripsikan variabel yang ditampilkan. Perhatikan perbedaan antara contoh-contoh berikut:

JUDUL LEMAH	JUDUL KUAT
PENJUALAN BERDASARKAN WILAYAH	Penjualan Wilayah Barat Unggul 23% pada Kuartal 3
SUHU SEIRING WAKTU	Suhu Harian Rata-rata Naik Stabil dari Maret hingga Juli
DISTRIBUSI USIA PELANGGAN	Sebagian Besar Pelanggan Berusia 25–44 Tahun

Judul yang kuat membantu pembaca memahami cerita di balik data secara langsung.

Label Sumbu

Label sumbu harus selalu mencantumkan satuan pengukuran, seperti:

- Dolar
- Persentase
- Kilogram
- Tahun

Pencantuman satuan mencegah ambiguitas dan memastikan data dapat ditafsirkan dengan tepat.

Label Data

Label data menampilkan nilai secara langsung pada elemen grafik seperti batang, titik, atau penanda garis. Sering kali, label data membuat grafik lebih mudah dibaca karena pengguna tidak perlu memperkirakan nilai berdasarkan skala sumbu.

Anotasi

Anotasi memberikan konteks penjelasan tambahan. Di Excel, anotasi dapat ditambahkan menggunakan kotak teks atau bentuk keterangan (*callout*). Anotasi sangat berguna untuk menyoroti:

- titik data yang tidak biasa
- titik balik yang penting
- kesimpulan utama

Anotasi yang ditempatkan dengan tepat dapat mengarahkan pembaca langsung ke wawasan paling penting dalam grafik.

Penggunaan Warna secara Sengaja

Warna harus memiliki tujuan komunikatif, bukan sekadar sebagai hiasan pada grafik. Untuk grafik yang menampilkan satu seri data, biasanya lebih baik menggunakan satu warna yang konsisten untuk semua elemen. Penggunaan banyak warna yang tidak diperlukan dapat membuat tampilan visual menjadi terlalu ramai atau membingungkan. Kontras warna sebaiknya digunakan hanya jika mampu menyampaikan informasi yang bermakna, seperti:

- menyoroti kategori tertentu
- membedakan beberapa seri data
- menunjukkan nilai di atas atau di bawah ambang batas tertentu
- memisahkan kelompok atau klasifikasi yang berbeda

Pertimbangan Aksesibilitas

Pemilihan warna juga harus mempertimbangkan keterbatasan penglihatan warna. Sekitar:

- 8% pria
- 0,5% wanita

mengalami beberapa bentuk buta warna. Bentuk yang paling umum, yaitu deuteranopia, membuat kombinasi warna merah dan hijau sulit dibedakan. Untuk meningkatkan aksesibilitas:

- Hindari kontras warna merah-hijau
- Gunakan palet warna yang berbasis pada warna biru, oranye, dan abu-abu
- Pastikan adanya kontras yang kuat antar kategori

Tema bawaan Excel menyediakan beberapa kombinasi warna yang aksesibel dan cocok untuk grafik profesional.

Garis Kisi (*Gridlines*), Legenda, dan Batas (*Borders*)

Panduan visual seperti garis kisi dan legenda sebaiknya hanya digunakan jika dapat meningkatkan kejelasan.

Garis Kisi

Garis kisi membantu pembaca memperkirakan nilai, namun dapat dengan mudah mendominasi grafik jika warnanya terlalu gelap atau jumlahnya terlalu banyak. Praktik terbaik meliputi:

- Hapus garis kisi (*gridlines*) jika tidak diperlukan.
- Jika digunakan, tampilkan dengan warna abu-abu sangat muda, bukan hitam.
- Buat tampilannya samar agar tetap berada di latar belakang.

Legenda

Legenda diperlukan ketika terdapat beberapa seri data dalam satu grafik dan seri-seri tersebut tidak dapat diidentifikasi secara langsung. Namun, legenda menjadi tidak perlu jika:

- hanya terdapat satu seri data, atau
- label seri dapat ditempatkan tepat di samping data tersebut.

Jika legenda digunakan:

- tempatkan di dekat konten grafik
- hapus bingkai (*border*) bawaan pada legenda
- pastikan label dalam legenda selaras dengan skema warna grafik secara jelas

Bingkai Grafik

Bingkai di sekeliling grafik maupun area plot jarang memberikan informasi yang berarti dan justru sering kali menambah kerumitan visual. Menghilangkan bingkai-bingkai ini biasanya akan meningkatkan kejelasan tampilan.

Poin Utama

Pilihan desain sangat memengaruhi seberapa efektif sebuah grafik dalam mengomunikasikan informasi. Dengan menerapkan prinsip-prinsip seperti memaksimalkan rasio data terhadap tinta (*data-ink ratio*), menyusun judul yang informatif, menggunakan warna secara terarah, serta meminimalkan elemen grafik yang tidak perlu, analis dapat menghasilkan visualisasi yang akurat sekaligus transparan. Grafik yang dirancang dengan baik tidak sekadar menyajikan data, tetapi juga menuntun audiens menuju pemahaman yang jelas.

6.7 VISUALISASI DATA EKSPLORATORI

Visualisasi Data Eksploratori (*Exploratory Data Visualization* atau EDV) merupakan landasan visual bagi Analisis Data Eksploratori (*Exploratory Data Analysis* atau EDA). EDV mengubah data mentah menjadi pola visual yang memungkinkan analis untuk mendeteksi struktur, anomali, hubungan, dan wawasan awal, jauh sebelum pemodelan statistik formal dimulai. Berbeda dengan grafik untuk presentasi, fokus utama EDV bukanlah menghasilkan grafik yang tampak sempurna atau dipoles dengan detail estetika. Sebaliknya, EDV mendukung pemikiran analitis dan proses penemuan. Melalui pemeriksaan visual yang cepat, analis dapat berinteraksi dengan data mereka, mengenali pola, serta menyusun hipotesis yang akan memandu analisis statistik selanjutnya. Bagian ini menyajikan alur kerja terstruktur untuk analisis data eksploratori, menjelaskan tujuan dari setiap tahapan serta perangkat visual yang membuat proses eksplorasi menjadi efektif.

1. Tujuan Visualisasi Data Eksploratori

Visualisasi data eksploratori membantu analis untuk:

- Memahami distribusi dan struktur variabel
- Mendeteksi pencilan (*outlier*), nilai yang hilang (*missing values*), dan masalah kualitas data
- Mengidentifikasi hubungan antarvariabel
- Menghasilkan hipotesis untuk pengujian statistik
- Memberikan dasar bagi rekayasa fitur (*feature engineering*) dan pemilihan model

EDA pada dasarnya bersifat iteratif (berulang). Setiap visualisasi sering kali memunculkan pertanyaan baru, yang mendorong analis untuk menelaah data dari perspektif lain.

2. Proses EDA dan Visualisasi yang Relevan

Alur kerja EDA yang sistematis membantu analis beralih dari pemahaman dasar mengenai kumpulan data menuju wawasan analitis yang lebih mendalam. Tahapan di bawah ini menguraikan proses eksplorasi yang umum dilakukan, beserta jenis visualisasi yang biasanya digunakan pada setiap langkahnya.

Tahap 1: Pengenalan Data (*Data Familiarization*)

Tujuan: Langkah awal dalam EDA adalah memahami struktur kumpulan data, termasuk jenis variabel yang terkandung di dalamnya serta keberadaan nilai yang hilang atau tidak konsisten.

Pertanyaan Kunci

- Variabel apa saja yang terdapat dalam kumpulan data?
- Apa saja jenis data yang ada (numerik, kategorikal, tanggal/waktu)?
- Apakah terdapat nilai yang hilang atau kesalahan input data yang terlihat jelas?

Visualisasi yang Berguna

- Tabel ringkasan tipe data kolom
- Peta panas (*heatmap*) data yang hilang (menggunakan pemformatan bersyarat/conditional formatting di Excel)
- Tabel frekuensi untuk variabel kategorikal

Contoh

Pertimbangkan sebuah dataset yang berisi 1.000 transaksi ritel dengan variabel-variabel berikut:

- Numerik: Kuantitas, Harga Satuan, Total Penjualan
- Kategorikal: Kategori Produk, Wilayah
- Tanggal: Tanggal Transaksi

Penerapan pemformatan bersyarat untuk menyoroti sel kosong akan dengan cepat mengungkap nilai Harga Satuan yang hilang, yang memerlukan penyelidikan lebih lanjut.

Tahap 2: Eksplorasi Univariat

Tujuan: Analisis univariat memeriksa satu variabel dalam satu waktu untuk memahami distribusi, tendensi sentral, dan variabilitasnya.

Pertanyaan Kunci

- Apakah distribusinya simetris atau menceng (*skewed*)?
- Apakah terdapat nilai ekstrem (pencilan)?
- Di mana letak tendensi sentralnya?

Visualisasi yang Berguna

- Histogram (bentuk distribusi)
- Box plot (sebaran dan pencilan)
- Diagram batang (frekuensi data kategorikal)
- Diagram garis (untuk variabel berbasis waktu)

Contoh

Histogram dari Total Penjualan menunjukkan distribusi yang menceng ke kanan (*right-skewed*), yang mengindikasikan bahwa sebagian besar transaksi menghasilkan nilai penjualan yang relatif rendah, sementara sebagian kecil transaksi menghasilkan nilai penjualan yang sangat tinggi.

Tahap 3: Eksplorasi Bivariat

Tujuan: Eksplorasi bivariat menyelidiki hubungan antara pasangan variabel.

Pertanyaan Kunci

- Apakah dua variabel bergerak bersamaan?
- Apakah terdapat perbedaan antar kelompok?
- Apakah hubungannya tampak linear atau non-linear?

Visualisasi yang Berguna

- Scatter plot (hubungan antara variabel numerik dengan numerik)
- Diagram batang berkelompok (perbandingan antara variabel kategorikal dengan kategorikal/numerik)
- Box plot berdampingan (perbandingan distribusi antar kelompok)

Contoh

Scatter plot antara Harga Satuan dan Kuantitas menunjukkan hubungan negatif, yang mengisyaratkan bahwa produk dengan harga lebih tinggi cenderung dibeli dalam jumlah yang lebih sedikit.

Tahap 4: Eksplorasi Multivariat

Tujuan: Analisis multivariat mengeksplorasi interaksi di antara tiga variabel atau lebih, membantu analis mengidentifikasi pola kompleks dan faktor-faktor yang mungkin memengaruhi hasil.

Pertanyaan Kunci

- Apakah hubungan antarvariabel berubah di berbagai kategori?
- Apakah terdapat klaster atau subkelompok dalam data?
- Variabel mana saja yang secara bersama-sama memengaruhi hasil?

Visualisasi yang Berguna

- Matriks scatter plot
- Diagram gelembung (*bubble chart*)
- Diagram batang berkelompok (*clustered bar chart*)
- Kombinasi PivotChart

Contoh

Diagram gelembung yang menampilkan:

- Sumbu X: Harga Satuan
- Sumbu Y: Kuantitas
- Ukuran gelembung: Total Penjualan
- Warna gelembung: Wilayah

menunjukkan bahwa satu wilayah secara konsisten menghasilkan transaksi dengan nilai lebih tinggi dibandingkan wilayah lainnya.

Tahap 5: Deteksi Anomali dan Pola

Tujuan: Tahap akhir dari EDA berfokus pada identifikasi pengembangan teknologi masa depan yang tidak biasa, pola struktural, dan potensi masalah pada data.

Pertanyaan Kunci

- Apakah terdapat pencilan (*outlier*) yang memerlukan investigasi?
- Apakah deret waktu (*time series*) menunjukkan pola musiman atau siklis?
- Apakah anomali yang muncul disebabkan oleh kesalahan atau peristiwa yang bermakna?

Visualisasi yang Berguna

- Diagram kotak (*box plot*) untuk mendeteksi pencilan
- Grafik deret waktu untuk mengidentifikasi tren dan pola musiman
- Diagram sebar (*scatter plot*) dengan garis tren

Contoh

Diagram kotak untuk "Total Penjualan" menunjukkan beberapa nilai ekstrem. Pemeriksaan lebih lanjut mengungkapkan bahwa nilai-nilai tersebut berasal dari pesanan korporat dalam jumlah besar, yang menandakan transaksi bernilai tinggi yang sah, bukan sekadar kesalahan dalam memasukkan data.

Poin Utama

Visualisasi data eksploratif merupakan langkah krusial dalam alur kerja analisis data. Dengan memeriksa distribusi, hubungan, dan anomali secara visual, analis dapat memperoleh

pemahaman intuitif mengenai kumpulan data sebelum menerapkan teknik statistik formal. Visualisasi data eksploratif bersifat iteratif, fleksibel, dan sangat penting untuk memahami data sebelum melakukan pemodelan. Setiap tahapan dalam analisis data eksploratif (EDA) memerlukan perangkat visual khusus untuk mengungkap struktur, hubungan, dan anomali. Excel menyediakan serangkaian perangkat visualisasi yang tangguh, cocok baik bagi pemula maupun analis berpengalaman. Visualisasi data eksploratif mendukung perumusan hipotesis dan pengambilan keputusan berbasis data.

6.8 VISUALISASI DINAMIS DAN INTERAKTIF

Grafik tradisional menyajikan gambaran statis dari data. Meskipun grafik statis berguna untuk laporan dan presentasi, analis sering kali membutuhkan perangkat yang memungkinkan pengguna untuk berinteraksi langsung dengan data. Visualisasi dinamis dan interaktif memungkinkan audiens untuk mengeksplorasi data dengan cara:

- memfilter kategori
- menyesuaikan rentang waktu
- memeriksa pengelompokan yang berbeda
- menelusuri lebih dalam ke subkelompok

Alih-alih mengharuskan analis untuk membuat banyak grafik secara manual, perangkat interaktif memungkinkan pengguna untuk memanipulasi visualisasi tersebut sendiri. Kemampuan ini sangat berharga untuk analisis eksploratif dan pelaporan melalui dasbor. Microsoft Excel menyertakan fitur-fitur yang mendukung visualisasi dinamis, termasuk PivotChart dan *slicer*.

PivotChart

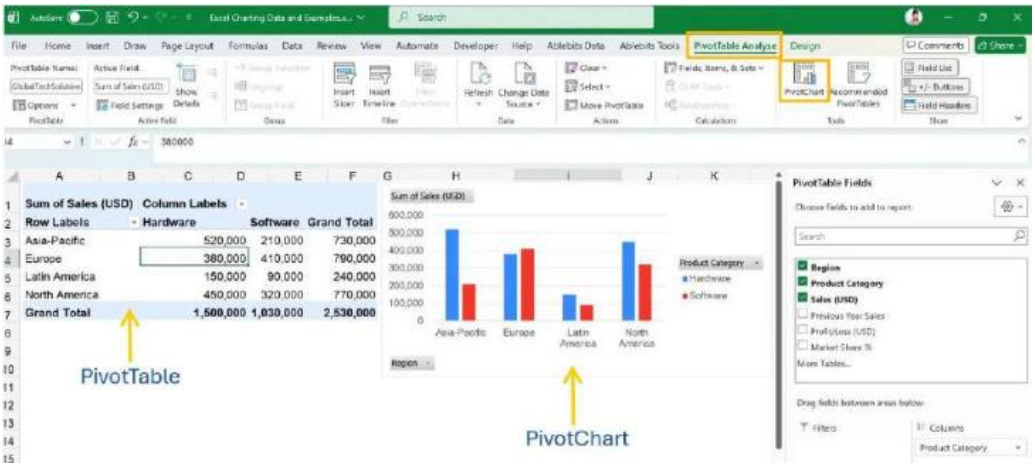
PivotChart adalah grafik yang terhubung langsung ke PivotTable. Berkat hubungan ini, grafik akan diperbarui secara otomatis setiap kali struktur atau filter pada PivotTable diubah. Hubungan dinamis ini menjadikan PivotChart sangat berguna untuk analisis eksploratif. Dengan mengatur ulang kolom (*field*) di dalam PivotTable, analis dapat langsung memvisualisasikan berbagai agregasi dari kumpulan data yang sama. PivotChart mendukung sebagian besar jenis grafik standar di Excel dan dapat disesuaikan menggunakan perangkat pemformatan yang sama dengan yang digunakan untuk grafik biasa.

Membuat PivotChart

Untuk membuat PivotChart:

1. Pertama, buatlah PivotTable dari kumpulan data Anda.
2. Pilih sel mana pun di dalam PivotTable tersebut.
3. Buka tab PivotTable Analyze → PivotChart.
4. Pilih jenis grafik yang diinginkan.

Setelah dibuat, PivotChart akan menampilkan tombol-tombol kolom (*field buttons*) di dalam area grafik. Tombol-tombol ini memungkinkan pengguna untuk memfilter atau mengatur ulang data secara langsung melalui antarmuka grafik. Untuk keperluan presentasi, tombol-tombol ini dapat disembunyikan guna menghasilkan tampilan visual yang lebih bersih setelah filter yang diinginkan diterapkan.



Gambar 6.16 Membuat Pivot Chart di Excel

Slicer

Slicer adalah kontrol filter visual yang menyediakan cara sederhana dan intuitif untuk memfilter PivotTable dan PivotChart. Alih-alih menggunakan menu dropdown, slicer menampilkan nilai filter yang tersedia sebagai tombol yang dapat diklik. Desain ini membuat status filter saat ini langsung terlihat oleh pengguna. Slicer meningkatkan transparansi karena pemirsa dapat dengan jelas melihat kategori mana yang saat ini termasuk dalam analisis.

Menambahkan Slicer

Untuk menyisipkan slicer:

1. Klik di mana saja di dalam PivotTable.
2. Navigasi ke Analisis PivotTable → Sisipkan Slicer.
3. Pilih bidang yang ingin Anda filter.
4. Klik OK.

Beberapa slicer dapat dihubungkan ke beberapa PivotTable dan PivotChart secara bersamaan. Kemampuan ini memungkinkan analis untuk membuat dasbor interaktif di mana pemfilteran satu elemen secara otomatis memperbarui beberapa grafik.

Membangun Dasbor Excel Sederhana

Dasbor Excel menggabungkan berbagai visualisasi dalam satu lembar kerja untuk menyajikan tampilan terpadu dari metrik-metrik utama. Dengan menyertakan *slicer* (alat pemilah) dan *PivotChart*, pengguna dapat mengeksplorasi data secara interaktif sembari tetap mendapatkan gambaran menyeluruh yang koheren.



Gambar 6.17 Sebuah slicer

Dasbor yang dirancang dengan baik biasanya:

- menampilkan beberapa grafik yang saling berkaitan
- menghubungkan grafik-grafik tersebut melalui filter bersama
- termuat dalam satu layar agar mudah dipahami

Langkah-langkah berikut menjelaskan cara menyusun dasbor penjualan sederhana dalam satu halaman.

Langkah 1: Menyiapkan Data

Atur kumpulan data Anda ke dalam satu tabel datar (*flat table*) yang memuat semua kolom relevan. Gunakan kumpulan data penjualan yang tersedia dalam buku kerja Excel di bab ini untuk mengikuti dan menyelesaikan latihan tersebut. Kumpulan data ini memiliki variabel-variabel berikut:

- order_id
- country
- order_date
- month
- product_category
- product
- order_quantity
- sales_amount

Ubah kumpulan data tersebut menjadi Tabel Excel (*Excel Table*) dengan menekan tombol Ctrl + T. Hal ini memastikan bahwa grafik dan *PivotTable* akan diperbarui secara otomatis saat ada data baru yang ditambahkan.

Langkah 2: Membuat *PivotTable*

Buat beberapa *PivotTable* dari sumber data yang sama untuk merangkum berbagai aspek data, contohnya:

- Total pendapatan berdasarkan negara

- Total pendapatan berdasarkan produk
- Total pendapatan berdasarkan bulan
- Komposisi pendapatan berdasarkan produk

Langkah 3: Membuat *PivotChart*

Buat *PivotChart* dari setiap *PivotTable* menggunakan jenis grafik yang sesuai dengan datanya:

- Grafik batang (*bar chart*) untuk pendapatan berdasarkan negara
- Grafik lingkaran (*pie chart*) atau grafik kolom (*column chart*) untuk pendapatan berdasarkan produk
- Grafik garis (*line chart*) untuk tren pendapatan bulanan
- Grafik area bertumpuk (*stacked area chart*) untuk komposisi pendapatan berdasarkan produk

Excel tidak menyediakan jenis grafik khusus berupa "kartu" (*card*) seperti yang ditemukan pada banyak perangkat lunak BI (*Business Intelligence*) khusus. Namun, Anda tetap dapat membuat kartu KPI yang rapi dan professional seperti Total Pendapatan atau Nilai Pembelian Rata-rata dengan menambahkan *PivotTable* yang hanya menampilkan satu nilai ringkasan ke dalam dasbor Anda. Melalui pengaturan format yang cermat, ringkasan ini dapat terlihat menyerupai kartu KPI yang sesungguhnya.

Langkah 4: Menambahkan *Slicer*

Tambahkan *slicer* untuk variabel-variabel penyaring utama, seperti Wilayah (*Region*) dan Produk (*Product*). Hubungkan setiap *slicer* ke semua *PivotTable* yang relevan menggunakan fitur *Report Connections* (Koneksi Laporan), sehingga pemilihan filter akan memperbarui seluruh grafik di dalam dasbor secara bersamaan. Untuk mengubah orientasi *slicer* dari vertikal menjadi horizontal:

1. Pilih *Slicer*: Klik pada *slicer* yang ingin Anda ubah orientasinya.
2. Buka Tab *Slicer*: Cari tab *Slicer* (pada versi Excel yang lebih lama, tab ini diberi label *Options*) di *Ribbon*.
3. Sesuaikan Kolom:
    - Cari grup *Buttons* (Tombol) di bagian paling kanan *Ribbon*.
    - Tingkatkan angka dalam kotak *Columns* (Kolom) agar sesuai dengan jumlah item yang ada di dalam *slicer* Anda.

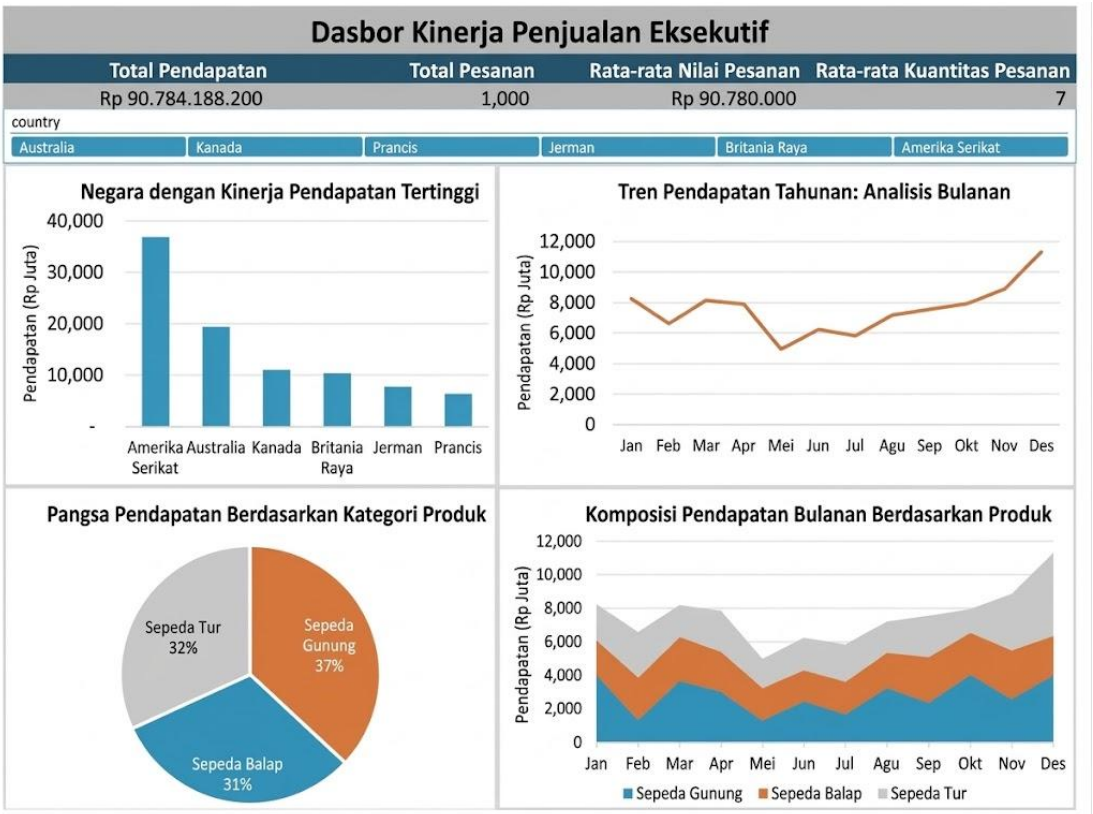
Langkah 5: Atur Tata Letak Dasbor

Tempatkan grafik dan *slicer* pada satu lembar kerja (*worksheet*) dengan tata letak yang rapi. Untuk meningkatkan kejelasan visual:

- Hilangkan garis kisi (*gridlines*) lembar kerja (buka tab *View* → hapus centang pada *Gridlines*).
- Gunakan pewarnaan sel (*cell shading*) yang lembut untuk memisahkan bagian-bagian tertentu.
- Pastikan penyelarasan (*alignment*) antar-grafik tetap konsisten.

Langkah 6: Tambahkan Konteks dan Dokumentasi

Berikan judul dasbor yang jelas dan sertakan detail penting lainnya. Pastikan pengguna dapat dengan cepat memahami cakupan, tujuan, serta periode waktu data yang disajikan.



Gambar 6.18 Contoh Dasbor Excel

Visualisasi dinamis mengubah Excel dari alat pembuatan grafik sederhana menjadi lingkungan analitik interaktif. Dengan menggabungkan PivotTable, PivotChart, dan slicer, analis dapat membuat dasbor yang memungkinkan pengguna untuk menjelajahi data dari berbagai perspektif. Perangkat ini tidak hanya meningkatkan efisiensi analisis, tetapi juga memungkinkan para pemangku kepentingan untuk berinteraksi langsung dengan data, sehingga menghasilkan pemahaman yang lebih mendalam dan pengambilan keputusan yang lebih tepat.

6.9 BERCERITA DENGAN DATA UNTUK MENDUKUNG PENGAMBILAN KEPUTUSAN

Dalam lingkungan profesional, grafik jarang berdiri sendiri. Grafik biasanya muncul dalam laporan yang harus memberikan dasar bagi rekomendasi, presentasi yang perlu menjaga perhatian audiens, dasbor yang dipantau manajer setiap hari, serta sesi pengarahan pengambilan keputusan di mana satu visual yang dirancang dengan buruk dapat merusak kesimpulan yang telah disusun selama berminggu-minggu. Dalam setiap konteks tersebut, kualitas analisis dari sebuah grafik memang penting, namun belum cukup. Grafik tersebut harus ditempatkan dalam struktur naratif yang menjelaskan kepada audiens apa yang ditunjukkan oleh data, mengapa hal itu penting, dan langkah apa yang harus diambil selanjutnya.

Bercerita dengan data (*data storytelling*) bukan berarti memperindah temuan atau membuatnya tampak lebih dramatis daripada kenyataannya. Sebaliknya, pendekatan ini

memastikan bahwa pesan inti yang ditemukan melalui analisis dapat tersampaikan secara utuh kepada audiens sehingga pembaca sampai pada kesimpulan yang sama dengan analis, dengan alasan yang sama, serta tingkat keyakinan yang setara. Bagian ini memperkenalkan struktur naratif yang menjadi landasan bagi presentasi data yang efektif, serta menguraikan teknik-teknik praktis untuk menerapkannya di Excel.

Alur Naratif: Konteks → Wawasan → Tindakan

Setiap cerita berbasis data yang efektif melalui tiga tahapan. Ketiganya membentuk alur naratif yang mencerminkan bagaimana keputusan sebenarnya diambil: suatu situasi harus dipahami, sebuah temuan harus dikenali, dan langkah respons harus ditentukan.

Tabel 6.6 Penjelasan mengenai Alur Naratif

Tahap	Tujuan	Pertanyaan yang Dijawabnya
Konteks	Menetapkan latar agar audiens dapat menafsirkan apa yang akan terjadi selanjutnya	Apa yang sedang kita lihat, dan mengapa hal itu penting?
Wawasan	Sampaikan temuan utama—pola, anomali, atau hubungan yang terungkap dari data tersebut.	Apa yang ditunjukkan oleh data tersebut?
Tindakan	Hubungkan temuan tersebut dengan suatu keputusan atau rekomendasi.	Apa yang sebaiknya kita lakukan secara berbeda sebagai tindak lanjutnya?

Konteks tidak sekadar memperkenalkan grafik. Konteks berfungsi menyelaraskan ekspektasi audiens dan menyediakan titik acuan yang diperlukan untuk menilai apakah suatu hasil memiliki makna. Tanpa konteks, temuan yang mencolok sekalipun bisa tampak abstrak atau kurang meyakinkan.

Tahap wawasan (*insight*) adalah saat grafik menjalankan fungsi analisis utamanya. Namun, wawasan tersebut harus dinyatakan secara eksplisit, bukan sekadar ditampilkan. Pola yang mungkin terlihat jelas bagi analis bisa jadi tidak langsung dipahami oleh audiens yang baru pertama kali melihat data tersebut. Tahap tindakan (*action*) membedakan antara sekadar pengembangan teknologi masa depan analitis dengan komunikasi yang bermanfaat. Presentasi yang berakhir dengan sebuah wawasan tanpa menyertakan langkah tindak lanjut sering kali membuat audiens bersikap pasif, sehingga temuan tersebut pun cepat terlupakan.

Contoh Penerapan: Alur Narasi dalam Praktik

- **Konteks:** Toko daring kami mencatat 48.000 sesi pelanggan pada kuartal lalu di lima kategori produk. Kami menghabiskan Rp 3,2 milyar untuk iklan digital, yang terbagi ke dalam saluran pencarian (*search*), media sosial, dan iklan tampilan (*display*).
- **Wawasan:** Diagram sebar yang membandingkan pengeluaran iklan dengan pendapatan per saluran menunjukkan bahwa iklan pencarian menghasilkan pendapatan sebesar Rp 82 ribu untuk setiap satu dolar yang dibelanjakan lebih dari dua kali lipat dibandingkan imbal hasil dari iklan tampilan (Rp 36 ribu). Hubungan ini bersifat linear kuat ($R^2 = 0{,}91$).

- **Tindakan:** Mengalihkan 20% anggaran iklan tampilan ke iklan pencarian diproyeksikan dapat meningkatkan pendapatan kuartalan sekitar Rp 740 juta, tanpa menambah total pengeluaran pemasaran.

Perhatikan bahwa diagram sebar di atas tidak akan mampu menyampaikan kesimpulan ini tanpa struktur narasi yang menyertainya.

Satu Grafik, Satu Pesan

Salah satu aturan paling andal dalam penyampaian cerita berbasis data (data storytelling) adalah aturan yang paling sederhana: setiap grafik harus menyampaikan tepat satu pesan. Grafik yang mencoba menampilkan berbagai tren, perbandingan, dan perincian secara bersamaan justru tidak menyampaikan lebih banyak informasi, melainkan lebih sedikit. Audiens tidak dapat memproses beberapa gagasan analitis sekaligus, sehingga pesan utama yang ingin disampaikan menjadi kabur. Sebelum menyelesaikan grafik untuk laporan atau presentasi, tulis satu kalimat yang melengkapi pernyataan berikut:

"Bagan ini menunjukkan bahwa _____, yang berarti bahwa _____ _____."

Jika pesan tidak dapat disampaikan dengan jelas dalam satu kalimat, visualisasi tersebut kemungkinan besar perlu disederhanakan atau disusun ulang.

Anotasi dan Penambahan Narasi

Anotasi berfungsi sebagai suara analis di dalam bagan berupa label singkat, keterangan (*callout*), atau kotak teks yang mengarahkan perhatian audiens pada temuan paling penting. Anotasi membantu menjembatani kesenjangan antara apa yang ditampilkan oleh bagan dan apa yang seharusnya disimpulkan oleh audiens. Jika digunakan secara efektif, satu anotasi saja dapat membuat sebuah bagan menjadi hampir bisa dipahami dengan sendirinya. Namun, jika digunakan secara berlebihan, anotasi justru menjadi gangguan visual yang bersaing dengan data itu sendiri. Excel menyediakan beberapa alat untuk menambahkan kejelasan naratif pada bagan:

- Label data *(Data labels*): Dapat ditambahkan dengan mengeklik kanan pada seri data dan memilih *Add Data Labels*. Fitur ini paling efektif jika digunakan untuk menyoroti poin-poin prioritas tertentu, alih-alih melabeli setiap pengembangan teknologi masa depan.
- Kotak teks (Text boxes): Dapat disisipkan melalui menu *Insert* → *Text* → *Text Box*. Kotak teks dapat ditempatkan di mana saja pada bagan dan diformat secara terpisah, menjadikannya alat anotasi yang paling fleksibel.
- Bentuk keterangan (*Callout shapes*): Tersedia melalui menu *Insert* → *Illustrations* → *Shapes* → *Callouts*. Bentuk ini menyertakan penghubung visual yang mengarah langsung ke elemen data yang relevan, sehingga mengurangi ambiguitas.

Bantuan Naratif Tambahan: Garis Referensi

Garis referensi memberikan konteks visual yang membantu audiens menafsirkan kinerja, variasi, atau arah data. Contoh yang umum meliputi:

- Garis rata-rata (*Average line*) – menunjukkan tingkat keseluruhan yang dapat digunakan sebagai pembanding bagi nilai-nilai individual.
- Garis dasar (*Baseline*) – menetapkan titik awal atau ambang batas minimum yang dapat diterima.
- Garis target (*Target line*) – menunjukkan tujuan atau tolok ukur kinerja.
- Garis tren (*Trendline*) – mengungkapkan arah mendasar atau pola jangka panjang dalam data. Elemen-elemen ini mengubah bagan dari sekadar tampilan sederhana menjadi penjelasan visual yang terarah.

Menyesuaikan Visual untuk Audiens yang Berbeda

Kumpulan data yang sama sering kali disajikan kepada audiens dengan kebutuhan informasi yang sangat berbeda. Dalam lingkungan analitik profesional, biasanya terdapat dua jenis audiens.

Tabel 6.7 Menyesuaikan visual untuk dua jenis audiens yang umum

Audiens	**Apa yang Mereka Butuhkan**	**Implikasi untuk Desain Bagan**
Eksekutif / Pengambil keputusan	Kesimpulan, besaran dampak, dan tindakan yang disarankan—secara cepat	Pimpin dengan temuan di judul bagan. Gunakan bagan tunggal yang tidak berantakan. Minimalkan label sumbu teknis. Anotasi harus menyatakan rekomendasi.
Analis / Rekan teknis	Metodologi, asumsi, ketidakpastian, dan distribusi yang mendasarinya	Sertakan label sumbu terperinci, ukuran sampel, dan interval kepercayaan bila perlu. Grafik pendukung mungkin muncul di lampiran atau bagian teknis.

Memenuhi kebutuhan ini tidak mengharuskan pembuatan visualisasi yang sepenuhnya berbeda. Sebaliknya, hal ini memerlukan penyusunan informasi secara berlapis dengan cermat, di mana grafik utama menyampaikan kesimpulan pokok, sementara materi pendukung memberikan kedalaman analisis.

Menyusun Urutan Grafik dalam Presentasi yang Melibatkan Banyak Grafik

Ketika beberapa grafik ditampilkan bersama baik dalam laporan, rangkaian slide, maupun dasbor alur narasi berlaku bagi keseluruhan dokumen sekaligus bagi setiap visual secara individual. Urutan kemunculan grafik sangat memengaruhi cerita yang dibangun oleh audiens. Struktur yang dapat diandalkan untuk presentasi dengan banyak grafik adalah:

- **Awali dengan gambaran umum:** Mulailah dengan grafik yang menetapkan skala dan cakupan topik—sering kali berupa tren dari waktu ke waktu atau perbandingan tingkat tinggi.
- **Perdalam menuju temuan:** Lanjutkan dengan satu atau dua grafik yang menyoroti hubungan, anomali, atau perbedaan spesifik yang menjadi wawasan utama.
- **Akhiri dengan implikasi:** Tutup dengan grafik atau ringkasan beranotasi yang mengukur konsekuensi praktis dari temuan tersebut serta mendukung rekomendasi yang jelas.

Dalam dasbor Excel, urutan ini dapat diterapkan secara alami ke dalam desain tata letak: baris KPI ringkasan di bagian atas (konteks), area grafik analisis utama di tengah (wawasan), serta panel perincian atau perbandingan di bagian bawah (tindakan). Interaktivitas berbasis *slicer* (alat pemilah data) yang diperkenalkan pada bagian sebelumnya memungkinkan dasbor tetap bersifat eksploratif sembari tetap mempertahankan struktur naratif ini.

Ringkasan Bagian: Lima Prinsip Bercerita dengan Data (*Data Storytelling*)

1. **Susun setiap presentasi dengan alur:** Konteks → Wawasan → Tindakan. Setiap tahapan menjawab pertanyaan yang berbeda dan mendukung pengambilan keputusan yang efektif.
2. **Terapkan aturan "Satu Grafik, Satu Pesan"**. Jika pesan dalam sebuah grafik tidak dapat diringkas dalam satu kalimat, kemungkinan besar grafik tersebut perlu direvisi.
3. **Gunakan anotasi untuk menjelaskan temuan**. Soroti satu atau dua elemen yang paling penting.
4. **Rancanglah sesuai dengan audiens Anda**. Para eksekutif lebih mengutamakan kesimpulan dan tindakan, sementara analis lebih berfokus pada metodologi dan bukti.
5. **Susun urutan grafik dengan sengaja**. Pola Gambaran Umum → Wawasan → Implikasi menciptakan alur narasi yang jelas untuk laporan, presentasi, maupun dasbor.

6.10 PRINSIP DAN TEKNIK VISUALISASI DATA UNTUK ANALISIS DAN KOMUNIKASI WAWASAN

Visualisasi data merupakan komponen sentral dalam sains data karena mengubah informasi numerik menjadi bentuk visual yang dapat ditafsirkan dengan cepat dan akurat. Grafik yang dirancang dengan baik memungkinkan analis untuk mendeteksi pola, tren, hubungan, dan anomali yang mungkin tersembunyi dalam tabel data mentah. Pada saat yang sama, visualisasi berfungsi sebagai alat komunikasi yang ampuh, membantu analis menyampaikan wawasan secara jelas kepada pengambil keputusan dan pemangku kepentingan lainnya.

Bab ini memperkenalkan prinsip-prinsip visualisasi data yang efektif serta mendemonstrasikan cara penerapannya menggunakan Excel. Pembahasan dimulai dengan kerangka kerja ART—*Accuracy* (Akurasi), *Relevance* (Relevansi), dan *Transparency* (Transparansi)—sebagai landasan bagi visualisasi yang etis dan bertanggung jawab. Akurasi memastikan bahwa grafik merepresentasikan data secara tepat tanpa distorsi. Relevansi menjaga fokus pada informasi yang mendukung tujuan analisis. Transparansi menuntut penyampaian asumsi, skala, dan sumber data secara jelas agar audiens dapat menafsirkan visualisasi dengan benar.

Berdasarkan landasan tersebut, bab ini mengulas pentingnya memilih jenis grafik yang tepat untuk tugas analisis tertentu. Grafik kolom dan batang mendukung perbandingan antar kategori, grafik garis menunjukkan tren dari waktu ke waktu, *scatter plot* membantu mengeksplorasi hubungan antarvariabel, dan histogram menampilkan distribusi data numerik. Pemilihan bentuk visual yang tepat memastikan bahwa grafik menonjolkan struktur data, bukan justru mengaburkannya.

Selanjutnya, bab ini membahas kemampuan visualisasi dalam Excel. Excel menyediakan berbagai perangkat bawaan seperti templat grafik, alat akses cepat untuk grafik, serta beragam opsi pemformatan yang memungkinkan analis membuat visualisasi yang jelas dan profesional tanpa memerlukan perangkat lunak khusus. Setelah membahas perangkat teknis, diskusi beralih ke prinsip-prinsip desain yang meningkatkan kejelasan dan kemudahan penafsiran. Visualisasi yang efektif menggunakan label yang jelas, skala yang tepat, serta meminimalkan elemen visual yang tidak perlu (*visual clutter*). Judul, legenda, dan label sumbu harus memberikan konteks yang cukup agar pembaca dapat memahami grafik tersebut secara mandiri. Desain yang baik mengarahkan perhatian pada data, bukan pada elemen dekoratif.

Bab ini kemudian mengeksplorasi visualisasi eksploratif, yang memegang peranan krusial pada tahap awal analisis. Dengan bereksperimen menggunakan histogram, *scatter plot*, *box plot*, dan jenis grafik lainnya, analis dapat menemukan pola, tren, dan anomali yang menjadi panduan untuk investigasi lebih mendalam serta pemodelan statistik. Dari tahap eksplorasi, bab ini berlanjut ke pembahasan mengenai visualisasi yang dinamis dan interaktif. Perangkat seperti PivotChart, slicer, dan komponen dasbor memungkinkan pengguna untuk memfilter, memperbarui, dan menelusuri data secara waktu nyata (*real-time*). Elemen-elemen interaktif ini membantu analis dalam menyusun visualisasi yang merespons masukan pengguna, sembari tetap menjaga narasi analitis yang koheren.

Terakhir, bab ini membahas tentang penyampaian cerita melalui data (*data storytelling*), yang mengubah grafik dari sekadar gambar terpisah menjadi narasi analitis yang persuasif. Cerita data yang efektif mengikuti struktur yang jelas: menetapkan konteks, menyajikan wawasan utama, serta mengaitkan wawasan tersebut dengan tindakan atau keputusan tertentu. Teknik-teknik seperti penggunaan anotasi, pengaturan urutan grafik yang cermat, dan prinsip “satu grafik, satu pesan” membantu mengarahkan audiens menuju kesimpulan yang paling penting. Dengan memadukan penalaran analitis yang kuat dan desain visual yang dirancang dengan saksama, analis dapat menciptakan visualisasi yang tidak hanya mengungkap pola dalam data, tetapi juga mengomunikasikan pola tersebut dengan cara yang mendukung pemahaman serta pengambilan keputusan yang tepat.

BAB 7
HUBUNGAN DALAM DATA ANDA – KORELASI DAN SELEBIHNYA

7.1 MEMAHAMI HUBUNGAN ANTARVARIABEL

Sejauh ini, dalam perangkat seorang pencerita data, Anda telah mempelajari cara mendeskripsikan variabel secara individual memeriksa pusat, sebaran, dan distribusinya, serta memvisualisasikannya menggunakan grafik dan statistik ringkasan. Meskipun perangkat ini membantu kita memahami satu variabel secara terpisah, wawasan paling bermakna dalam analisis data sering kali muncul saat kita memeriksa bagaimana variabel-variabel saling berhubungan. Dalam banyak masalah dunia nyata, pertanyaan utamanya bukan sekadar "seperti apa variabel ini?", melainkan "bagaimana variabel ini berubah ketika variabel lain berubah?". Sebagai contoh:

- Apakah kepuasan pelanggan berkaitan dengan frekuensi pembelian?
- Apakah pengalaman karyawan berkaitan dengan produktivitas?
- Apakah waktu belajar berkaitan dengan hasil ujian?

Pertanyaan-pertanyaan seperti ini memperkenalkan konsep asosiasi, yang mengkaji apakah dua variabel bergerak bersama secara sistematis. Memahami hubungan ini memungkinkan analis untuk menemukan pola, menyusun penjelasan, dan membentuk dasar bagi model prediktif. Bab ini memperkenalkan perangkat utama yang digunakan untuk mengeksplorasi hubungan antarvariabel. Karena variabel bisa bersifat kategorikal atau numerik, metode analisis yang tepat bergantung pada jenis variabel yang terlibat. Oleh karena itu, kita akan membahas beberapa skenario umum beserta teknik untuk menganalisisnya secara efektif. Di akhir bab ini, Anda akan mampu:

- Menentukan metode yang tepat untuk menganalisis hubungan berdasarkan jenis variabel yang terlibat.
- Menyusun dan menafsirkan tabel kontingensi serta memvisualisasikan hubungan antarvariabel kategorikal di Excel.
- Membandingkan hasil numerik di berbagai kelompok kategorikal menggunakan statistik deskriptif dan diagram kotak (*box plot*).
- Menafsirkan diagram sebar (*scatter plot*) dan menghitung koefisien korelasi Pearson menggunakan Excel.
- Membedakan antara korelasi dan kausalitas (sebab-akibat) serta memahami keterbatasan dalam analisis hubungan.

Meskipun bab ini berfokus pada hubungan deskriptif yaitu mengidentifikasi pola yang muncul dalam data yang diamati perhatian khusus diberikan pada hubungan antara variabel-variabel numerik. Hubungan ini sering kali diperiksa menggunakan analisis korelasi, sebuah perangkat ampuh untuk mengukur kekuatan dan arah asosiasi antarvariabel. Memahami korelasi memberikan landasan penting bagi analisis regresi dan pemodelan prediktif, yang akan dibahas pada Bab 8.

7.2 ANALISIS HUBUNGAN ANTARVARIABEL: KEKUATAN, ARAH, DAN METODE ANALISIS

Apa Arti "Berhubungan"? Saat kita mengatakan bahwa dua variabel saling berkaitan atau berhubungan, maksudnya adalah mengetahui nilai salah satu variabel dapat memberikan informasi mengenai kemungkinan nilai variabel lainnya. Dengan kata lain, variabel-variabel tersebut tidak bervariasi secara independen; perubahan pada satu variabel cenderung berkaitan dengan perubahan pada variabel lainnya. Perhatikan contoh-contoh berikut:

- **Berkaitan:** Suhu dan penjualan es krim — hari yang lebih hangat biasanya dikaitkan dengan angka penjualan yang lebih tinggi.
- **Tidak berkaitan:** Ukuran sepatu dan kecerdasan — mengetahui ukuran sepatu seseorang tidak memberikan informasi yang berguna mengenai IQ orang tersebut.

Dalam praktiknya, hubungan antarvariabel dapat bervariasi, baik dari segi kekuatan maupun arahnya.

Kekuatan Hubungan

Kekuatan hubungan menggambarkan seberapa kuat variabel-variabel tersebut bergerak bersama.

- **Hubungan kuat:** Mengetahui satu variabel memberikan informasi yang substansial mengenai variabel lainnya.
- **Hubungan lemah:** Mengetahui satu variabel hanya memberikan informasi yang terbatas mengenai variabel lainnya.
- **Tidak ada hubungan:** Mengetahui satu variabel tidak memberikan informasi apa pun mengenai variabel lainnya.

Hubungan yang kuat lebih mudah dideteksi dalam data dan sering kali menghasilkan pola visual yang lebih jelas saat digambarkan dalam grafik.

Arah Hubungan

Untuk banyak hubungan terutama yang melibatkan variabel numerik arah hubungan juga menjadi aspek yang penting.

- **Hubungan positif:** Saat satu variabel meningkat, variabel lainnya cenderung ikut meningkat.
- **Hubungan negatif:** Saat satu variabel meningkat, variabel lainnya cenderung menurun.

Sebagai contoh, hubungan antara durasi belajar dan nilai ujian biasanya bersifat positif, sedangkan hubungan antara harga produk dan jumlah permintaan sering kali bersifat negatif.

Mengapa Hubungan Antarvariabel Itu Penting

Memahami hubungan antarvariabel merupakan tujuan utama dalam analisis data karena hal ini mendukung berbagai tugas analisis yang penting.

- **Pengambilan keputusan:** Jika kepuasan pelanggan berkaitan dengan pembelian ulang, maka meningkatkan kepuasan pelanggan menjadi prioritas strategis.
- **Prediksi:** Jika durasi belajar berkaitan dengan nilai ujian, kita dapat mengestimasi kemungkinan hasil ujian berdasarkan waktu yang dihabiskan untuk belajar.

- **Menyelidiki penyebab:** Meskipun hubungan antarvariabel tidak membuktikan adanya sebab-akibat (kausalitas), mengidentifikasi keterkaitan sering kali menjadi langkah awal untuk memahami mekanisme sebab-akibat yang mungkin terjadi.
- **Alokasi sumber daya:** Jika pengeluaran untuk iklan berkaitan dengan pendapatan penjualan, organisasi dapat menggunakan informasi tersebut untuk mengalokasikan anggaran pemasaran secara lebih efektif.

Dalam setiap kasus tersebut, mengidentifikasi hubungan membantu mengubah data mentah menjadi wawasan yang dapat ditindaklanjuti.

Kerangka Kerja untuk Mengeksplorasi Hubungan

Metode yang tepat untuk menganalisis hubungan bergantung pada jenis variabel yang terlibat. Karena variabel dapat berupa kategorikal maupun numerik, teknik analisis yang berbeda diperlukan untuk setiap kombinasinya.

Tabel 7.1 Kerangka kerja untuk mengeksplorasi hubungan

Tipe Variabel 1	Tipe Variabel 2	Metode Utama	Contoh
Kategorikal	Kategorikal	Tabel Kontingensi	Jenis Kelamin vs. Preferensi Produk
Kategorikal	Numerikal	Perbandingan Kelompok	Departemen vs. Gaji
Numerikal	Numerikal	Analisis Korelasi	Jam Belajar vs. Nilai Tes

Setiap pasangan memerlukan teknik visualisasi dan statistik ringkasan yang berbeda. Dengan mempelajari cara menganalisis ketiga jenis hubungan tersebut, Anda akan memiliki kemampuan untuk mengeksplorasi seluruh ragam pola yang muncul dalam kumpulan data di dunia nyata. Bagian-bagian berikut akan mengulas setiap skenario tersebut secara mendetail serta menunjukkan cara menerapkan metode yang relevan menggunakan Excel. Buku kerja Excel untuk bab ini berisi kumpulan data yang dapat Anda gunakan untuk berlatih di setiap tahapan. Cobalah untuk membuat ulang setiap contoh yang diberikan.

7.3 HUBUNGAN ANTARA VARIABEL KATEGORIKAL

Pada Bab 3, Anda telah mempelajari cara menyusun tabel kontingensi (tabulasi silang) untuk mengatur data yang melibatkan dua variabel kategorikal. Pada Bab 6, Anda telah mengeksplorasi bagaimana diagram batang berkelompok (*grouped bar charts*) dan diagram batang bertumpuk (*stacked bar charts*) dapat memvisualisasikan hubungan antarvariabel kategorikal. Dalam bagian ini, kita akan memperluas gagasan tersebut dengan berfokus pada cara menafsirkan hubungan antara variabel-variabel kategorikal serta cara menentukan apakah terdapat asosiasi di antara variabel-variabel tersebut. Tabel kontingensi merangkum distribusi gabungan dari dua variabel kategorikal dengan menampilkan frekuensi pengembangan teknologi masa depan pada setiap kombinasi kategori.

Contoh: Kepuasan Pelanggan Berdasarkan Jenis Produk

Misalkan sebuah perusahaan menyurvei 300 pelanggan dari tiga lini produk dan mencatat tingkat kepuasan mereka.

Tabel 7.2 Tabel kontingensi – kepuasan pelanggan berdasarkan jenis produk

Kepuasan	Produk A	Produk B	Produk C	Total
Puas	65	85	45	195
Netral	20	10	25	55
Tidak Puas	15	5	30	50
Total	100	100	100	300

Meskipun jumlah mentah memberikan informasi yang berguna, perbandingan antar kategori menjadi lebih mudah ketika kita mengubah tabel tersebut ke dalam bentuk persentase kolom.

Tabel 7.3 Tabel kontingensi – persentase kolom

Kepuasan	Produk A	Produk B	Produk C
Puas	65%	85%	45%
Netral	20%	10%	25%
Tidak puas	15%	5%	30%

Persentase kolom memungkinkan kami membandingkan kepuasan dalam setiap kategori produk, sehingga membuat pola lebih terlihat.

Menafsirkan Hubungan Kategoris

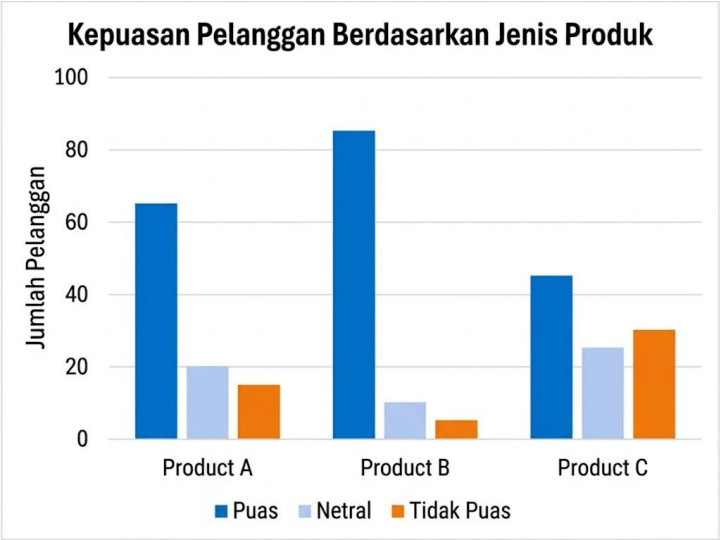
Beberapa pengamatan dapat dilakukan dari tabel persentase.

- **Pengenalan Pola**
  Produk B memiliki tingkat kepuasan tertinggi (85%), sedangkan Produk C memiliki tingkat kepuasan terendah (45%). Distribusi kepuasan jelas berbeda pada ketiga produk tersebut.
- **Kekuatan Asosiasi**
  Perbedaan antar produk sangat besar. Misalnya, terdapat kesenjangan sebesar 40 poin persentase antara tingkat kepuasan Produk B dan Produk C. Perbedaan besar tersebut menunjukkan adanya hubungan yang kuat antara jenis produk dan kepuasan pelanggan.
- **Arah Asosiasi**
  Meskipun variabel kategori tidak memiliki "arah" seperti hubungan numerik, kita masih dapat mengamati kategori mana yang dikaitkan dengan hasil tertentu.
  Dalam contoh ini, Produk B sangat terkait dengan kepuasan pelanggan, sedangkan Produk C lebih sering dikaitkan dengan ketidakpuasan.

- **Prinsip Utama: Kemandirian vs. Persatuan**
  Dua variabel bersifat independen ketika mengetahui nilai satu variabel tidak memberikan informasi tentang variabel lainnya. Mereka terkait ketika pengetahuan tentang satu variabel mengubah ekspektasi kita terhadap variabel lainnya.
- **Contoh Kemerdekaan**
  Jika Anda melempar koin dan melempar sebuah dadu, hasil pelemparan koin tidak mempengaruhi hasil pelemparan dadu. Peristiwa-peristiwa ini bersifat independen.
- **Contoh Asosiasi**
  Jika kita menelaah tingkat pendidikan dan pendapatan, mengetahui bahwa seseorang mempunyai gelar sarjana mungkin akan mengubah ekspektasi kita terhadap pendapatannya. Oleh karena itu, variabel-variabel ini saling terkait. Memahami apakah variabel independen atau terkait merupakan langkah penting dalam analisis data eksplorasi.
- **Memvisualisasikan Hubungan Kategoris**
  Seperti yang diperkenalkan di Bab 6, beberapa tipe bagan efektif untuk memvisualisasikan hubungan kategoris.

Bagan Batang yang Dikelompokkan

Menampilkan kategori dari satu variabel secara berdampingan dalam setiap kategori variabel lainnya. Format ini membuat perbandingan antar kategori menjadi mudah.



Gambar 7.1 Diagram batang berkelompok – kepuasan pelanggan berdasarkan jenis produk

Diagram Batang Bertumpuk 100% (100% *Stacked Bar Chart*)

Menampilkan distribusi proporsional dalam setiap kategori, yang memungkinkan pengguna untuk membandingkan perbedaan persentase antar kelompok.

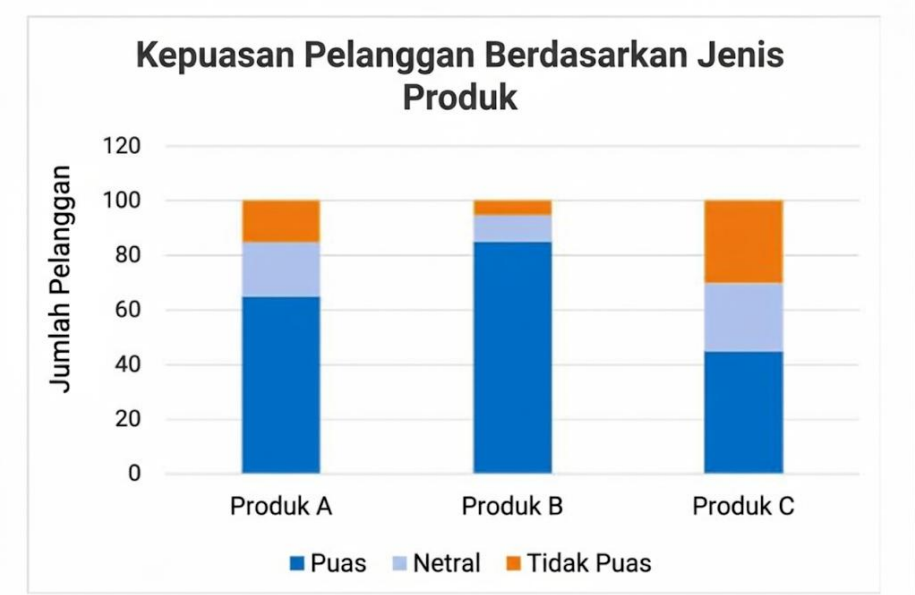
Memilih Visualisasi yang Tepat

- Gunakan grafik batang berkelompok saat membandingkan jumlah absolut.
- Gunakan grafik batang bertumpuk 100% saat membandingkan distribusi persentase.
- Pilih visualisasi yang paling jelas mengkomunikasikan pola tersebut.

Aplikasi Praktis di Excel

Excel menyediakan cara efisien untuk membuat tabel kontingensi menggunakan PivotTable. Pada tahap ini dalam kursus, Anda telah membuat beberapa PivotTable, jadi prosedur di bawah ini berfungsi sebagai ringkasan singkat:

1. Pilih dataset Anda.
2. Buka Sisipkan → *PivotTable*.
3. Seret satu variabel kategorikal ke area Baris.
4. Seret variabel kategorikal kedua ke area Kolom.
5. Tempatkan variabel yang sama di area Nilai dan ringkas menggunakan Hitung.



Gambar 7.2 Grafik batang bertumpuk – kepuasan pelanggan berdasarkan jenis produk

Untuk menampilkan persentase:

- Klik kanan nilai-nilai di PivotTable.
- Pilih Tampilkan Nilai Sebagai → % dari Total Kolom.

Pemeriksaan Cepat untuk Hubungan

Saat memeriksa tabel kontingensi:

- Jika persentase berbeda secara substansial di seluruh kolom (atau baris), variabel-variabel tersebut kemungkinan terkait.
- Jika persentase serupa di seluruh kolom, variabel-variabel tersebut mungkin independen.
- Perbedaan yang lebih besar biasanya menunjukkan hubungan yang lebih kuat.

Di bagian selanjutnya, kita akan memperluas gagasan ini ke situasi di mana satu variabel bersifat kategorikal dan yang lainnya numerik, yang membutuhkan pendekatan analitis yang berbeda.

7.4 HUBUNGAN ANTARA VARIABEL KATEGORIKAL DAN NUMERIK

Di bagian ini, kita akan memeriksa situasi di mana satu variabel bersifat kategorikal dan yang lainnya numerik. Pertanyaan utama dalam jenis analisis ini adalah:

Apakah variabel kategorikal berbeda di seluruh kategori?

Contohnya meliputi:

- Apakah gaji berbeda menurut departemen?
- Apakah nilai ujian berbeda berdasarkan metode pengajaran?
- Apakah pengeluaran pelanggan berbeda berdasarkan tingkat loyalitas?

Pada Bab 3, Anda telah mempelajari cara menggunakan *PivotTable* untuk menghitung statistik ringkasan berdasarkan kelompok. Pada Bab 6, Anda telah mempelajari cara memvisualisasikan distribusi di seluruh kategori menggunakan boxplot dan grafik kelompok. Sekarang kita akan menggunakan alat-alat tersebut untuk menafsirkan hubungan antara variabel kategorikal dan numerik.

Membandingkan Rata-Rata: Pendekatan Praktis

Saat menganalisis apakah variabel kategorikal memengaruhi variabel numerik, pendekatan terstruktur membantu memandu interpretasi.

Langkah 1: Hitung Statistik Deskriptif

Untuk setiap kategori, hitung:

- Rata-rata dan median (kecenderungan sentral)
- Deviasi standar (sebaran)
- Ukuran sampel (konteks untuk reliabilitas)

Langkah 2: Bandingkan Rata-Rata

Tanyakan:

- Seberapa besar perbedaan antara rata-rata kelompok?
- Apakah perbedaan tersebut bermakna secara praktis, bukan hanya berbeda secara numerik?

Langkah 3: Memeriksa Variabilitas

Deviasi standar yang besar menunjukkan tumpang tindih yang lebih besar antara distribusi kelompok. Deviasi standar yang lebih kecil menunjukkan pemisahan yang lebih jelas antara kelompok.

Contoh: Gaji di Seluruh Departemen

Misalkan sebuah perusahaan memeriksa gaji tahunan untuk 150 karyawan di lima departemen. Setelah menghitung statistik kecenderungan sentral dan variabilitas per departemen, mereka memperoleh ringkasan berikut.

Tabel 7.4 Statistik deskriptif perbandingan kelompok

Departemen	Gaji Rata-rata (Rp)	Gaji Median (Rp)	Simpangan Baku (Rp)	Ukuran Sampel
Teknik	Rp768.320.000	Rp770.560.000	Rp116.030.000	35

Keuangan	Rp728.000.000	Rp727.020.000	Rp125.790.000	30
Pemasaran	Rp658.210.000	Rp647.890.000	Rp113.450.000	25
Operasional	Rp580.360.000	Rp577.430.000	Rp97.130.000	20
Penjualan	Rp674.870.000	Rp665.530.000	Rp156.760.000	40

Menginterpretasikan Perbandingan Antar Kelompok

1. **Apakah Terdapat Perbedaan?**
   Ya. Rata-rata gaji departemen berkisar dari Rp 580.360.000 (Operasi) hingga Rp 768.320.000 (Teknik), selisih sebesar Rp 187.960.000. Variasi ini menunjukkan bahwa gaji dan departemen saling terkait.
2. **Seberapa Besar Perbedaannya?**
   Besarnya perbedaan memberikan konteks penting:
   - Gaji Teknik sekitar 32% lebih tinggi daripada gaji Operasi.
   - Kesenjangan antara rata-rata departemen tertinggi dan terendah cukup besar.
   - Perbedaan ini kemungkinan besar mewakili pola organisasi yang bermakna daripada fluktuasi kecil.
3. **Pola Apa yang Muncul?**
   Beberapa pola terlihat:
   - Departemen teknis seperti Teknik dan Keuangan memiliki gaji yang lebih tinggi.
   - Peran operasional cenderung memiliki rata-rata gaji yang lebih rendah.

   Penjualan menunjukkan deviasi standar terbesar, menunjukkan variabilitas yang lebih besar dalam pembayaran mungkin mencerminkan kompensasi berbasis komisi atau perbedaan kinerja.

Menggunakan PivotTable Excel untuk Ringkasan Kategorikal-Numerik

PivotTable menyediakan cara efisien untuk meringkas nilai numerik berdasarkan kategori.

1. Sisipkan → PivotTable.
2. Tempatkan Departemen di area Baris.
3. Tempatkan Gaji di area Nilai.

Untuk menghitung beberapa statistik:

1. Tambahkan Gaji beberapa kali ke area Nilai.
2. Ubah statistik ringkasan setiap bidang.

Tabel 7.5 Pivot table – ringkasan kategorikal-numerik

Label Baris	Rata-rata Gaji	Deviasi Standar Gaji	Jumlah Data Gaji
Teknik (*Engineering*)	76.832,34	11.603,00	35
Keuangan (*Finance*)	72.799,53	12.579,36	30
Pemasaran (*Marketing*)	65.821,48	11.345,44	25
Operasional (*Operations*)	58.035,60	9.712,85	20
Penjualan (*Sales*)	67.486,98	15.676,38	40
Total Keseluruhan (*Grand* Total)	**69.192,31**	**13.940,29**	**150**

Excel PivotTable tidak secara bawaan menyertakan median di antara statistik ringkasan untuk kolom-kolomnya.

Memvisualisasikan Hubungan antara Variabel Kategorikal dan Numerik

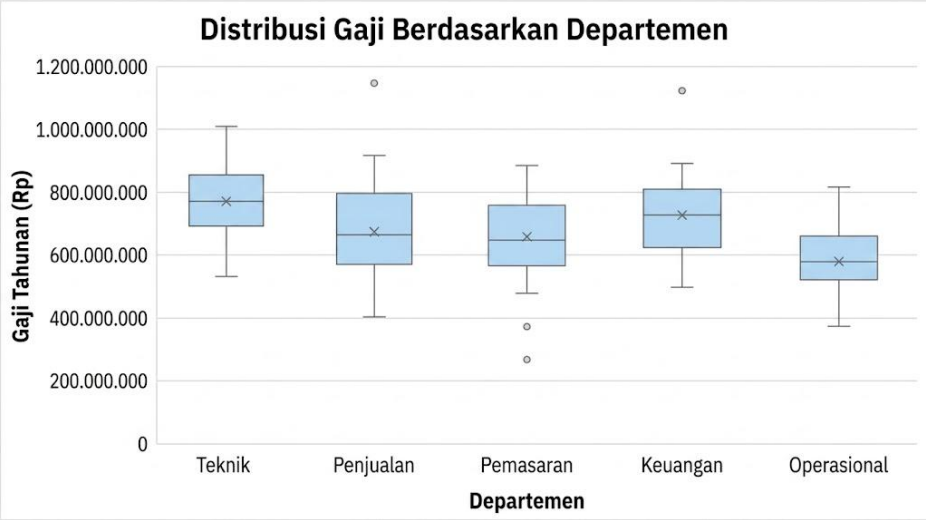
Visualisasi yang paling efektif untuk membandingkan variabel numerik di berbagai kategori adalah *box plot* berdampingan (*side-by-side boxplot*), yang telah diperkenalkan pada Bab 6. Setiap *box plot* merangkum karakteristik utama dari suatu distribusi, yaitu: median, rentang interkuartil (IQR), pencilan (*outliers*), dan sebaran data secara keseluruhan.

Manfaat Boxplot

Boxplot memungkinkan analis untuk:

- Membandingkan bentuk distribusi antar kelompok
- Mengidentifikasi perbedaan variabilitas
- Mendeteksi outlier dalam kategori tertentu
- Mengevaluasi tumpang tindih antara distribusi kelompok

Karena boxplot menampilkan seluruh distribusi, bukan hanya satu statistik, boxplot sering kali mengungkapkan pola yang tidak dapat ditunjukkan oleh tabel ringkasan saja.



Gambar 7.3 categoricat-wuericat retatiowstip - boxplot berdampingan

Membuat Boxplot Berdampingan di Excel

Untuk memastikan Excel mengenali data Anda sebagai kategori terpisah untuk plot berdampingan, data Anda harus diatur dalam "Format Panjang" tertentu.

1. **Siapkan Data Anda**
   - Kolom A (Kategori) misalnya, "Teknik", "Penjualan", dan lain lain.
   - Kolom B (Nilai) titik data numerik yang sesuai untuk setiap kategori.

   Catatan: Pastikan kategori berada dalam satu kolom dan semua nilai berada di kolom berikutnya.
2. **Sisipkan Bagan**
   - Pilih Data: Sorot kedua kolom, termasuk header.
   - Buka Sisipkan → Sisipkan Bagan Statistik → Kotak dan Kumis

Catatan: Kecuali versi sebelum pembaruan terbaru mungkin tidak menyediakan bagan kotak berdampingan.

Pertimbangan Praktis

- Sampel kecil (misalnya, kurang dari 10 pengamatan) menghasilkan estimasi yang tidak stabil.
- Ukuran kelompok yang tidak sama dapat diterima tetapi harus diakui.
- Sampel yang lebih besar meningkatkan kepercayaan pada pola yang diamati.

Pencilan Dapat Mengubah Hasil

- Nilai ekstrem dapat menarik rata-rata ke atas atau ke bawah.
- Median lebih tahan terhadap pencilan.
- Selalu periksa visualisasi untuk nilai yang tidak biasa.

Konteks Sangat Penting

Perbedaan statistik harus diinterpretasikan dalam konteksnya.

- Perbedaan gaji Rp 50 juta mungkin signifikan di beberapa industri tetapi kecil di industri lain.
- Pertimbangkan perbedaan absolut dan perbedaan persentase.
- Fokus pada signifikansi praktis, bukan hanya variasi numerik.

7.5 HUBUNGAN ANTAR VARIABEL NUMERIK – ANALISIS KORELASI

Sekarang kita sampai pada inti dari bab ini: memahami hubungan antara dua variabel numerik. Pada tahap ini, kita memperkenalkan analisis korelasi, alat statistik yang ampuh yang digunakan untuk mengukur bagaimana dua variabel bergerak bersama. Tidak seperti hubungan yang melibatkan variabel kategorikal yang biasanya kita gambarkan secara kualitatif atau hubungan kategorikal-numerik yang kita bandingkan menggunakan rata-rata kelompok hubungan antara variabel numerik dapat diukur secara tepat. Ketepatan ini menjadikan analisis korelasi sebagai dasar penting untuk prediksi, yang akan kita bahas lebih lanjut di Bab 8 tentang analisis regresi.

Kemungkinan Hubungan Antara Variabel Numerik

Saat mengeksplorasi hubungan antara dua variabel numerik, beberapa pola berbeda dapat muncul dalam diagram pencar (*scatterplot*). Mengenali pola-pola ini sangat penting karena korelasi hanya mengukur hubungan yang bersifat linear (garis lurus). Hubungan linear terjadi ketika perubahan pada satu variabel dikaitkan dengan perubahan proporsional pada variabel lainnya. Dalam diagram pencar, titik-titik data cenderung mengelompok di sekitar garis lurus, baik yang miring ke atas (positif) maupun ke bawah (negatif). Analisis korelasi dirancang khusus untuk mengukur kekuatan dan arah dari hubungan yang berbentuk garis lurus ini. Gagasan utama: Jika pola dalam diagram pencar kira-kira mengikuti garis lurus, maka korelasi merupakan alat analisis yang tepat.

Hubungan Nonlinear

Variabel-variabel mungkin tetap memiliki hubungan meskipun hubungan tersebut tidak bersifat linear. Dalam kasus seperti ini, diagram pencar mungkin menunjukkan pola yang melengkung atau lebih kompleks. Contohnya meliputi:

- Hubungan berbentuk huruf U
- Pola pertumbuhan eksponensial
- Hasil yang semakin menurun (*diminishing returns*)

Hubungan nonlinear yang kuat bisa tampak lemah atau bahkan tidak ada sama sekali jika diukur menggunakan korelasi, karena korelasi hanya mendeteksi pola garis lurus. Mengenali sifat nonlinear sejak awal membantu mencegah kesalahan interpretasi, di mana korelasi yang lemah dianggap sebagai hubungan yang lemah pula.

Tidak Ada Hubungan

Dalam beberapa situasi, dua variabel mungkin bergerak secara independen satu sama lain. Jika ini terjadi, diagram pencar akan menampilkan sebaran titik-titik acak tanpa pola yang jelas tidak ada tren naik, turun, maupun melengkung.

Langkah Awal Memahami Korelasi: Diagram Pencar (*Scatterplot*)

Sebelum menghitung statistik apa pun, selalu visualisasikan hubungan antara variabel-variabel numerik menggunakan diagram pencar.

Contoh: Jam Belajar dan Nilai Ujian

Gunakan kumpulan data `*hours_studied_vs_exam_score*` yang tersedia dalam buku kerja Excel bab ini untuk mengikuti pembahasan dan menyelesaikan latihan. Enam puluh mahasiswa melaporkan jumlah jam yang mereka habiskan untuk belajar menghadapi ujian beserta nilai ujian yang mereka peroleh:

Tabel 7.6 Beberapa data awal dari kumpulan data `hours_studied_vs_exam_score`.

nomor_mahasiswa	jam_belajar	skor_ujian
1	6	79
2	5	75
3	7	75
4	9	83
5	5	82
6	5	86
7	9	93
8	7	93
9	5	76
10	7	78
... (50 mahasiswa lagi)		

... (50 mahasiswa lainnya)

Membuat Scatterplot di Excel

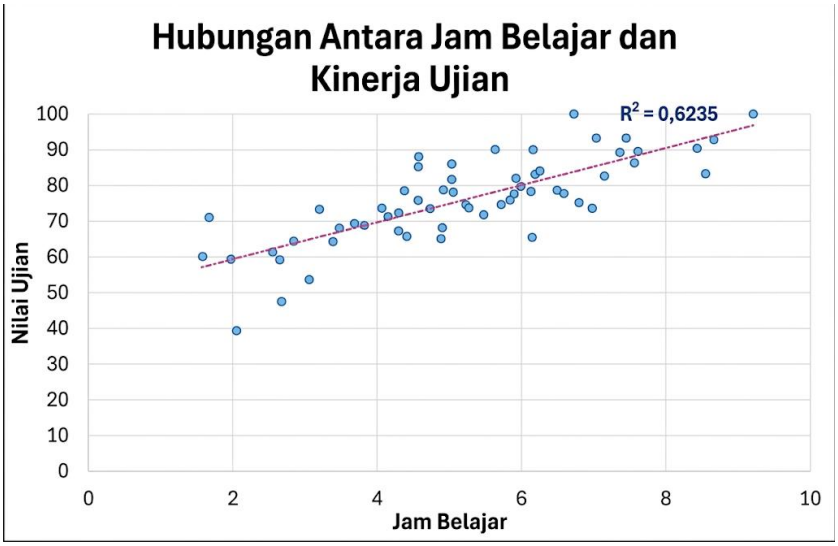
Untuk membuat scatterplot (diagram sebar) di Excel, kita mengikuti langkah-langkah berikut:

1. Pilih dua kolom data (jam_belajar dan nilai_ujian).
2. Buka menu Insert → Scatter → Scatter with Only Markers (Scatter hanya dengan penanda).
3. Tambahkan judul sumbu dan judul grafik.
4. Sesuaikan format agar lebih jelas.

Menambahkan garis tren (*trendline*) dan nilai R^2:

1. Klik kanan pada salah satu titik data.
2. Pilih Add Trendline (Tambahkan Garis Tren).
3. Pilih Linear.
4. Centang opsi Display R-squared value on chart (Tampilkan nilai R-squared pada grafik).

Hasilnya akan menyerupai grafik pada gambar 7.4.



Gambar 7.4 Scatterplot – hubungan antara jam belajar dan performa ujian

Hal yang Perlu Diperhatikan dalam Scatterplot

Saat memeriksa diagram pencar (*scatterplot*), perhatikan karakteristik berikut:

- **Linearitas:** Apakah titik-titik tersebut membentuk pola garis lurus, atau justru melengkung?
- **Arah:** Apakah titik-titik tersebut cenderung mengarah ke atas (hubungan positif) atau ke bawah (hubungan negatif)?
- **Kekuatan:** Apakah titik-titik tersebut berkumpul rapat di sekitar garis, atau tersebar luas?
- **Pencilan (*Outlier*):** Apakah terdapat pengembangan teknologi masa depan yang letaknya jauh dari pola umum?

Contoh Interpretasi

Dalam contoh mengenai jam belajar, diagram pencar menunjukkan pola linear yang jelas mengarah ke atas: titik-titik tersebut umumnya bergerak dari kiri bawah menuju kanan atas. Hal ini mengindikasikan hubungan positif; mahasiswa yang belajar lebih lama cenderung memperoleh nilai ujian yang lebih tinggi.

Memahami Korelasi

Korelasi mengukur kekuatan dan arah hubungan linear antara dua variabel numerik. Korelasi menjawab pertanyaan: Seberapa erat variabel-variabel ini bergerak bersama dalam pola garis lurus?

Koefisien Korelasi Pearson (r)

Ukuran korelasi yang paling banyak digunakan adalah koefisien korelasi Pearson, yang umumnya dilambangkan dengan r. Statistik ini merangkum hubungan antara dua variabel menggunakan satu angka saja.

1. **Rentang: Nilai r selalu berada di antara –1 dan +1.**
2. **Tanda menunjukkan arah**
   - r > 0: Hubungan positif
   - r < 0: Hubungan negatif
   - r = 0: Tidak ada hubungan linear
3. **Nilai mutlak menunjukkan kekuatan**

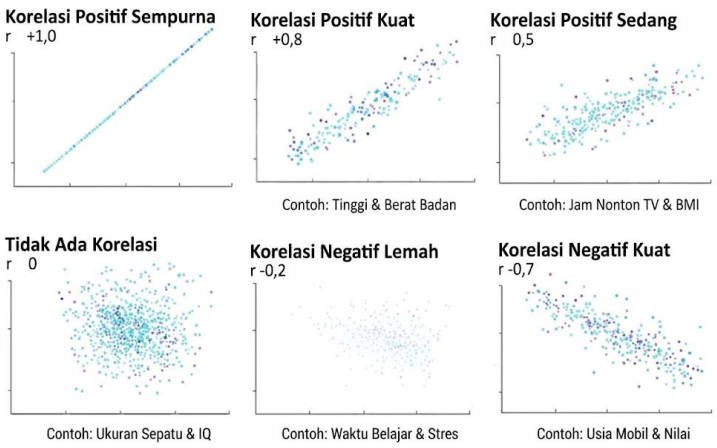
   Nilai mutlak dari r menunjukkan seberapa kuat hubungan tersebut.
   - |r| mendekati 1: Hubungan kuat
   - |r| sekitar 0,5: Hubungan moderat (sedang)
   - |r| mendekati 0: Hubungan lemah atau tidak ada hubungan sama sekali

Pola Visual r

Memahami korelasi memerlukan kemampuan untuk mengenali pola dalam diagram pencar:

- **Korelasi Positif Sempurna (r = +1,0):** Semua titik terletak tepat pada garis lurus yang mengarah ke atas.
- **Jarang ditemukan dalam data nyata. Korelasi Positif Kuat (r ≈ +0,8):** Titik-titik berkumpul rapat di sekitar garis yang mengarah ke atas, namun tetap menunjukkan sedikit sebaran.
- **Korelasi Positif Moderat (r ≈ +0,5):** Terdapat tren umum ke arah atas, namun dengan sebaran yang cukup signifikan.
- **Korelasi Positif Lemah (r ≈ +0,2):** Terdapat sedikit tren meningkat, namun titik-titik data tersebar luas. Pola sulit terlihat tanpa konfirmasi statistik.

Tidak Ada Korelasi (r ≈ 0): Titik-titik data membentuk sebaran acak tanpa pola yang jelas. Korelasi Negatif (r < 0): Pola serupa, tetapi dengan kemiringan ke arah bawah. Contoh: Usia mobil dan nilainya (r ≈ -0,7).



Gambar 7.5 Pola visual korelasi

Angka memberi tahu kita seberapa kuat hubungan antarvariabel, namun visualisasi membantu kita memahami sifat hubungan tersebut serta mengidentifikasi pola yang mungkin tidak tertangkap oleh koefisien korelasi.

Menghitung Korelasi di Excel

1. Menggunakan Fungsi CORREL

Excel menyediakan fungsi bawaan, CORREL, yang menghitung koefisien korelasi Pearson antara dua variabel numerik.

Sintaks:

$$= CORREL(rentang1, rentang2)$$

Contoh:

$$= CORREL(B2:B61, C2:C61)$$

Rumus ini menghitung korelasi antara jumlah jam belajar dan nilai ujian. Misalkan hasilnya adalah:

$$r = 0{,}79$$

2. Menggunakan Data Analysis ToolPak

Metode Analysis ToolPak, yang akan dibahas lebih lanjut di bab ini, menyediakan cara alternatif untuk menghitung korelasi. Berbeda dengan fungsi CORREL, ToolPak dapat menghasilkan matriks korelasi lengkap, yang memungkinkan Anda untuk memeriksa korelasi berpasangan di antara beberapa variabel numerik secara sekaligus.

Menafsirkan Korelasi

Untuk kumpulan data jam belajar:

$$r = 0{,}79$$

Ini menunjukkan korelasi positif yang kuat antara jumlah jam belajar dan nilai ujian. Artinya:

- Mahasiswa yang belajar lebih lama cenderung mendapatkan nilai yang lebih tinggi.
- Titik-titik pada diagram pencar (*scatterplot*) mengelompok cukup rapat di sekitar garis yang mengarah ke atas.

Namun, korelasi saja tidak sepenuhnya menggambarkan hubungan tersebut. Untuk memahami seberapa besar variasi dalam nilai ujian yang dikaitkan dengan waktu belajar, kita perlu memeriksa r^2.

Memahami r²: Koefisien Determinasi

Statistik r^2 (r-kuadrat) memberikan wawasan tambahan mengenai hubungan antarvariabel. Jika r mengukur kekuatan dan arah hubungan, maka r^2 mengukur proporsi variasi dalam satu variabel yang dikaitkan dengan variasi dalam variabel lainnya.

Rumus

$$r^2 = r \times r$$

Contoh Jika:

$$\mathrm{r} = 0.79$$

Kemudian:

$$\mathrm{r}^2 = 0.79^2 = 0.62$$

Ini berarti 62% variasi dalam nilai ujian terkait dengan variasi jam belajar. Sisa 38% variasi berasal dari faktor lain, seperti:

- pengetahuan sebelumnya
- metode belajar
- tidur
- kecemasan ujian
- variasi acak

Mengapa r² Penting

Koefisien determinasi penting karena:

- Lebih mudah diinterpretasikan daripada r
- Mengukur seberapa banyak informasi prediktif yang diberikan satu variabel tentang variabel lain
- terhubung langsung dengan analisis regresi, yang akan kita bahas di bab berikutnya. Tidak seperti r, r^2 selalu positif, berkisar dari 0 hingga 1.

Tabel 7.7 Contoh Interpretasi r^2

r	r²	Interpretasi
0,9	0,81	81% variasi dapat dijelaskan
0,7	0,49	49% dapat dijelaskan
0,5	0,25	25% dapat dijelaskan
0,3	0,09	Hanya 9% yang dapat dijelaskan

Bahkan korelasi yang moderat pun bisa menyisakan sebagian besar variasi yang tidak terjelaskan, yang menunjukkan kompleksitas fenomena di dunia nyata.

Memahami Hubungan: Interpretasi r dan r²

Memahami korelasi memerlukan lebih dari sekadar menghitung suatu angka. Analis harus menafsirkan koefisien korelasi (r) maupun koefisien determinasi (r^2) dalam konteks data yang ada. Contoh-contoh berikut mengilustrasikan bagaimana analisis korelasi diinterpretasikan dalam praktik.

Contoh 1: Periklanan dan Penjualan

Seorang pengecer melacak pengeluaran iklan bulanan dan pendapatan penjualan selama periode 24 bulan.

Tabel 7.8 4 catatan pertama dari sebuah dataset

Periklanan (dalam ribuan Rp)	Penjualan (dalam ribuan Rp)
12	185
15	203
10	172
18	228
…	…

Analisis

- Diagram pencar (scatterplot) menunjukkan tren naik yang jelas.
- $r = +0{,}68$, yang menunjukkan korelasi positif moderat (sedang).
- $r^2 = 0{,}46$, yang berarti 46% variasi dalam penjualan berkaitan dengan variasi dalam pengeluaran iklan.

Interpretasi

Iklan dan penjualan memiliki korelasi moderat. Pengeluaran iklan yang lebih tinggi cenderung dikaitkan dengan penjualan yang lebih tinggi. Namun, iklan hanya menjelaskan kurang dari separuh variasi dalam penjualan; ini menunjukkan bahwa faktor lain seperti musiman, strategi penetapan harga, persaingan, dan kondisi ekonomi juga memengaruhi pendapatan.

Contoh 2: Suhu dan Biaya Pemanas Ruangan

Seorang pemilik rumah mencatat suhu luar ruangan harian dan biaya pemanas ruangan selama 30 hari.

Analisis

- Diagram pencar menunjukkan tren turun yang jelas.
- $r = -0{,}82$, yang menunjukkan korelasi negatif yang kuat.
- $r^2 = 0{,}67$, yang berarti 67% variasi dalam biaya pemanas ruangan berkaitan dengan perubahan suhu.

Interpretasi

Saat suhu turun, biaya pemanas ruangan meningkat. Hubungan ini kuat, yang menunjukkan bahwa suhu luar ruangan merupakan faktor utama yang memengaruhi pengeluaran untuk pemanas ruangan. Meskipun demikian, sekitar 33% variasi dalam biaya pemanas ruangan dijelaskan oleh faktor lain, seperti kualitas isolasi, pengaturan termostat, dan ukuran rumah.

Contoh 3: Usia dan Waktu Reaksi

Para peneliti mengukur waktu reaksi (dalam milidetik) pada 50 orang dewasa berusia 20–70 tahun.

Analisis

1. Diagram pencar menunjukkan tren naik yang tipis.
2. $r = +0{,}31$, yang menunjukkan korelasi positif yang lemah.
3. $r^2 = 0{,}10$, yang berarti hanya 10% variasi dalam waktu reaksi yang berkaitan dengan usia.

Interpretasi

Orang dewasa yang lebih tua cenderung menunjukkan waktu reaksi yang sedikit lebih lambat, namun hubungannya lemah. Usia hanya menjelaskan sebagian kecil dari variasi tersebut; ini menunjukkan bahwa perbedaan individu, kesehatan, kebugaran fisik, dan pelatihan kognitif memainkan peran yang jauh lebih besar.

Matriks Korelasi dan Peta Panas (*Heatmap*)

Saat bekerja dengan kumpulan data yang memuat banyak variabel numerik, analis sering kali menghitung korelasi untuk setiap pasangan variabel secara bersamaan. Proses ini menghasilkan matriks korelasi. Matriks korelasi adalah tabel yang menampilkan koefisien korelasi antara setiap pasangan variabel dalam kumpulan data. Meskipun Excel tidak membuat peta panas (*heatmap*) secara otomatis, fitur pemformatan bersyarat (*conditional formatting*) dapat digunakan untuk menghasilkan efek visual yang serupa.

Membuat Matriks Korelasi di Excel

1. Buka Data → Data Analysis → Correlation.
2. Pilih rentang input (Input Range) yang mencakup semua variabel numerik yang ingin dianalisis.
3. Jika data Anda menyertakan judul kolom, centang opsi Labels in First Row.
4. Pada bagian Output Options, tentukan lokasi untuk menampilkan matriks tersebut.
5. Klik OK.

Excel akan menghasilkan tabel yang menunjukkan korelasi antara semua pasangan variabel.

Memvisualisasikan Matriks dengan Pemformatan Bersyarat

Untuk menyoroti pola secara visual:

1. Pilih matriks korelasi tersebut.
2. Buka Home → *Conditional Formatting* → *Color Scales*.
3. Pilih skala warna, misalnya Merah–Kuning–Hijau (*Red–Yellow–Green*).

Untuk pengaturan warna yang lebih presisi:

1. Pilih *Conditional Formatting* → *More Rules*.
2. Pilih 3-Color Scale.
3. Atur nilai sebagai berikut:
   - Minimum: –1 (Merah)
   - Titik tengah (Midpoint): 0 (Kuning)
   - Maksimum: +1 (Hijau)

Langkah ini menghasilkan visualisasi bergaya peta panas (*heatmap*), di mana:

- korelasi positif yang kuat akan tampak berwarna hijau,
- korelasi negatif yang kuat akan tampak berwarna merah, dan
- korelasi yang lemah akan tampak berwarna kuning.

Visualisasi semacam ini memungkinkan analis untuk mengidentifikasi hubungan penting dalam kumpulan data yang besar dengan cepat. Memahami keterbatasan korelasi sama pentingnya dengan memahami cara menghitungnya.

1. Korelasi Tidak Menyiratkan Kausalitas (Sebab-Akibat)

Ini adalah salah satu prinsip terpenting dalam statistika. Jika dua variabel berkorelasi, ada beberapa kemungkinan penjelasan:

- X menyebabkan Y
- Y menyebabkan X
- Variabel ketiga (Z) memengaruhi keduanya
- Hubungan tersebut hanyalah kebetulan

Contoh: Penjualan Es Krim dan Kasus Tenggelam

Penjualan es krim dan jumlah kematian akibat tenggelam memiliki korelasi yang kuat. Apakah es krim menyebabkan orang tenggelam? Tentu saja tidak. Penjelasan yang sebenarnya adalah suhu:

- Cuaca hangat meningkatkan konsumsi es krim
- Cuaca hangat juga meningkatkan aktivitas berenang

Suhu berperan sebagai variabel pengganggu (*confounding variable*) yang memengaruhi kedua hasil tersebut.

2. Korelasi Hanya Mengukur Hubungan Linear

Koefisien korelasi Pearson mendeteksi pola garis lurus. Jika variabel-variabel mengikuti hubungan melengkung, korelasi mungkin tampak lemah meskipun sebenarnya terdapat hubungan yang kuat.

Contoh: Hubungan antara kecemasan dan kinerja sering kali berbentuk huruf U terbalik tingkat kecemasan moderat dapat meningkatkan kinerja, sementara tingkat kecemasan yang terlalu rendah atau terlalu tinggi justru menurunkannya. Meskipun terdapat hubungan non-linear yang kuat, nilai korelasinya bisa mendekati nol. Praktik terbaik: Selalu periksa diagram pencar (*scatterplot*) sebelum menafsirkan koefisien korelasi.

3. Pencilan (*Outlier*) Dapat Mendistorsi Korelasi

Satu pengamatan ekstrem saja dapat mengubah koefisien korelasi secara drastis. Sebagai contoh, bayangkan 19 mahasiswa tidak menunjukkan hubungan apa pun antara waktu belajar dan nilai ujian. Jika satu mahasiswa belajar selama 20 jam dan memperoleh nilai 100%, nilai korelasi bisa melonjak dari mendekati nol menjadi 0,6. Praktik terbaik: Periksa diagram pencar untuk melihat adanya pencilan, dan pertimbangkan untuk mengevaluasi korelasi dengan menyertakan maupun tidak menyertakan nilai-nilai ekstrem tersebut.

4. Rentang yang Terbatas Dapat Menyembunyikan Hubungan

Jika data hanya mencakup rentang nilai yang sempit, korelasi mungkin tampak lebih lemah daripada yang sebenarnya. Sebagai contoh, pada kelompok orang dewasa berusia 25–30 tahun, korelasi antara pendidikan dan pendapatan mungkin sebesar $r = 0{,}30$. Namun,

jika dilihat pada seluruh populasi orang dewasa (usia 18–65 tahun), korelasi tersebut bisa meningkat menjadi $r = 0{,}60$ karena terdapat variasi yang lebih besar pada kedua variabel tersebut.

5. Korelasi Tidak Menunjukkan Kemiringan (*Slope*)

Dua hubungan dapat memiliki koefisien korelasi yang sama namun dengan kemiringan (*slope*) yang sangat berbeda. Sebagai contoh:

- Hubungan antara jam belajar dan nilai ujian mungkin menunjukkan peningkatan 5 poin per jam.
- Hubungan antara tinggi badan dan berat badan mungkin menunjukkan peningkatan 2 pon per inci.

Kedua hubungan tersebut bisa saja menghasilkan nilai $r = 0{,}8$, meskipun tingkat perubahannya berbeda. Korelasi mengukur kekuatan asosiasi, bukan tingkat kecuraman hubungan tersebut. Korelasi memberi tahu kita apakah dua variabel saling berhubungan dan seberapa kuat hubungan tersebut. Namun, korelasi tidak menjawab beberapa pertanyaan penting, seperti:

- Persamaan apa yang mendeskripsikan hubungan tersebut?
- Bagaimana cara memprediksi Y berdasarkan X?
- Seberapa besar perubahan pada Y ketika X meningkat sebesar satu satuan?

Pertanyaan-pertanyaan ini dijawab melalui analisis regresi, yang akan kita bahas pada Bab 8. Anda dapat memahami hubungan ini sebagai berikut:

- Korelasi menetapkan bahwa suatu hubungan memang ada.
- Regresi memodelkan hubungan tersebut secara matematis.
- Keduanya bersama-sama menyediakan perangkat yang ampuh untuk memahami dan memprediksi hubungan numerik.

Korelasi adalah fondasinya; regresi adalah struktur yang dibangun di atas fondasi tersebut.

Ukuran Sampel Itu Penting

Sampel yang kecil menghasilkan korelasi yang tidak stabil. Pada kumpulan data yang sangat kecil ($n < 10$), korelasi yang besar bisa muncul hanya karena kebetulan. Sampel yang lebih besar ($n \geq 30$) umumnya menghasilkan estimasi yang lebih dapat diandalkan. Apa yang dianggap sebagai korelasi "kuat" bergantung pada disiplin ilmunya.

Bidang Interpretasi Umum

Berikut adalah data dari gambar yang telah diubah menjadi tabel dan diterjemahkan ke dalam bahasa Indonesia:

Bidang	**Interpretasi Umum**
Fisika / Kimia	r > 0,9 sering diharapkan
Ilmu Sosial	r > 0,6 dianggap kuat
Kedokteran	r ≈ 0,4 mungkin bermakna

Variasi ini mencerminkan kompleksitas dari sistem yang berbeda. Perilaku manusia dan sistem biologis sering kali melibatkan banyak faktor yang saling berinteraksi, sehingga korelasi yang sangat tinggi jarang ditemukan.

7.6 MENYATUKAN SEMUANYA – KERANGKA KERJA PENGAMBILAN KEPUTUSAN

Sepanjang bab ini, Anda telah mempelajari beberapa metode untuk memeriksa hubungan antarvariabel. Metode yang tepat terutama bergantung pada jenis variabel yang terlibat. Sebelum memulai analisis apa pun, langkah pertama yang harus dilakukan adalah mengidentifikasi apakah setiap variabel bersifat kategorikal atau numerik. Setelah mengetahui jenis variabelnya, pemilihan pendekatan analisis yang tepat menjadi lebih mudah.

Pohon Keputusan:

Apakah kedua variabel bersifat kategorikal?
├ YA → Gunakan tabel kontingensi dan diagram batang berkelompok/bertumpuk
└ TIDAK → Lanjutkan

Apakah satu variabel kategorikal dan satu numerik?
├ YA → Bandingkan rata-rata antar kelompok menggunakan tabel pivot dan diagram kotak
└ TIDAK → Lanjutkan

Apakah kedua variabel bersifat numerik?
└ YA → Hitung korelasi dan buat diagram pencar

Kerangka kerja ini membantu memandu proses analisis dengan mencocokkan jenis variabel dengan metode statistik dan visualisasi yang sesuai.

Tabel 7.9 Tabel Perbandingan

Jenis Variabel	Metode	Visual	Statistik Utama	Interpretasi
Kategorikal × Kategorikal	Tabel kontingensi dengan persentase	Diagram batang berkelompok atau bertumpuk	Bandingkan persentase antar kategori	Perhatikan perbedaan persentase
Kategorikal × Numerikal	Rerata, median, dan simpangan baku kelompok	Diagram kotak (*boxplot*) berdampingan	Bandingkan rata-rata dan distribusi	Perhatikan perbedaan serta tumpang tindih (*overlap*) pada nilai rata-rata (*mean*) atau median
Numerikal × Numerikal	Koefisien korelasi	Diagram sebar (*scatterplot*)	Koefisien korelasi Pearson r dan r^2	Interpretasikan kekuatan dan arahnya

Tabel ini merangkum perangkat utama untuk mengeksplorasi hubungan dalam analisis data eksploratif (*exploratory data analysis*).

Contoh Komprehensif: Analisis Karyawan

Untuk mengilustrasikan bagaimana metode-metode ini bekerja bersama, perhatikan skenario berikut. Sebuah perusahaan ingin memahami faktor-faktor yang memengaruhi kinerja karyawan. Mereka mengumpulkan data dari 200 karyawan, yang mencakup:

- Departemen (kategorikal): Penjualan (*Sales*), Teknik (*Engineering*), Pemasaran (*Marketing*)
- Lama Pengalaman (numerik): 0–25 tahun
- Jam Pelatihan (numerik): 0–100 jam per tahun
- Peringkat Kinerja (numerik): skala 1–10

Tabel 7.10 Sampel dari kumpulan data analisis kinerja karyawan

Pengalaman Kerja	Jam_Pelatihan	Departemen	Peringkat_Kinerja
16,8	99	Teknik	10
11	99	Penjualan	9,2
19	98	Penjualan	9,4
20	97	Teknik	10
15	97	Pemasaran	10
...	...	...	...

Gunakan dataset employee_performance_analysis yang disediakan dalam buku kerja Excel bab ini untuk mengikuti dan mereplikasi analisis. Mari kita jelajahi beberapa pertanyaan analitis menggunakan kerangka kerja ini.

Pertanyaan 1: Apakah kinerja berbeda menurut departemen?

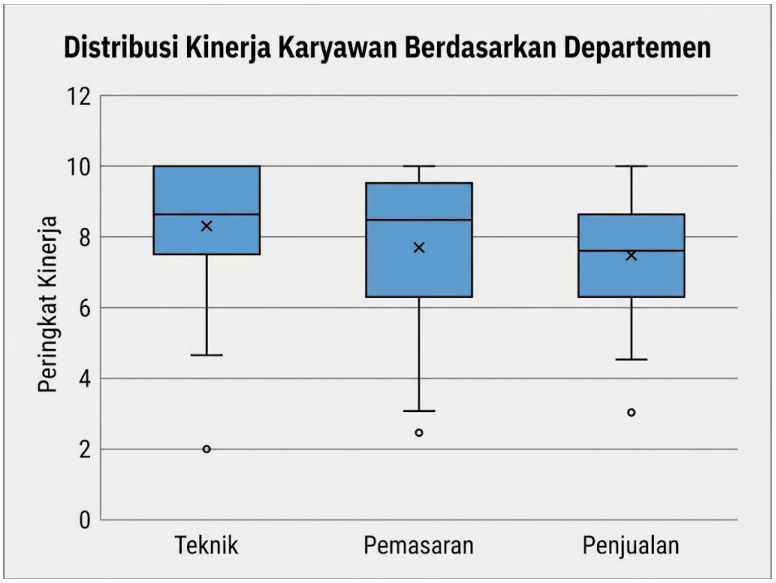
Tipe Variabel: Kategorikal (Departemen) × Numerik (Kinerja)

Metode: Hitung rata-rata, median, dan deviasi standar dari peringkat kinerja untuk setiap departemen dan/atau visualisasikan distribusinya menggunakan boxplot berdampingan.

Hasil

Tabel 7.11 Statistik deskriptif kelompok perbandingan

Label Baris	Peringkat Kinerja Rata-rata	Median	Standar Deviasi dari Peringkat Kinerja
Teknik	8,3	8,6	1,7
Pemasaran	7,7	8,45	2,1
Penjualan	7,5	7,6	1,4
Total Keseluruhan	7,9	8,35	1,8



Gambar 7.6 Diagram kotak (*box plot*) berdampingan untuk distribusi kinerja karyawan

Interpretasi

Karyawan di departemen Teknik (*Engineering*) menunjukkan peringkat kinerja rata-rata tertinggi, diikuti oleh departemen Pemasaran (*Marketing*), dan kemudian Penjualan (*Sales*). Meskipun perbedaannya tidak terlalu besar, hal ini mengindikasikan bahwa kinerja mungkin sedikit bervariasi antar departemen.

Pertanyaan 2: Apakah pengalaman dan jumlah pelatihan memengaruhi kinerja?

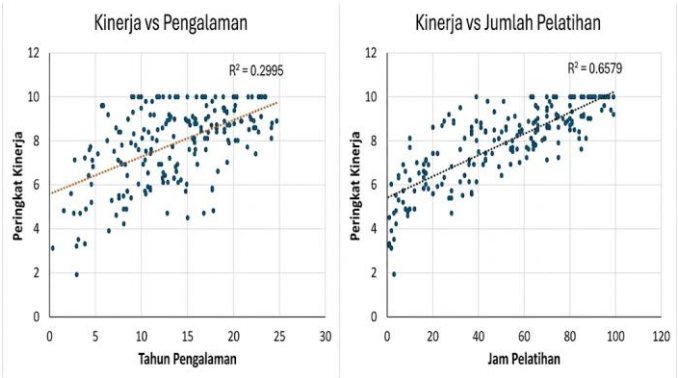
Jenis Variabel:

- Numerik (Lama Pengalaman) × Numerik (Kinerja)
- Numerik (Jam Pelatihan) × Numerik (Kinerja)

Metode:

- Diagram pencar (*scatter plot*) yang menampilkan nilai R^2 (linearitas, arah, dan kekuatan hubungan)
- Fungsi korelasi dan R^2 di Excel (arah dan kekuatan hubungan)
- Matriks korelasi (arah dan kekuatan hubungan)

Diagram Pencar (*Scatter Plot*)



Gambar 7.7 Diagram pencar kinerja vs pengalaman dan kinerja vs jumlah pelatihan yang menampilkan nilai R^2

Fungsi Korelasi (r) dan R^2 di Excel

Menghitung korelasi menggunakan fungsi (=CORREL()) dan menguadratkan nilai korelasi tersebut untuk memperoleh nilai R^2.

Variabel	Korelasi	r-kuadrat
Pengalaman vs Kinerja	0,55	0,30
Pelatihan vs Kinerja	0,81	0,66

Tidak ada cara untuk menghitung nilai *R-squared*, karena sudah ditampilkan di awal; kami menghitungnya di sini untuk latihan.

Peta Panas Matriks Korelasi

Tabel 7.12 Peta panas korelasi

	Pengalaman Bertahun-tahun	Jam Pelatihan	Penilaian Kinerja
Pengalaman (Tahun)	1		
Jam Pelatihan	0,38	1	
Peringkat Kinerja	0,55	0,81	1

Matriks ini menggunakan gradien berkode warna untuk memvisualisasikan bagaimana variabel-variabel saling berhubungan:

- Hijau Tua menunjukkan hubungan positif yang kuat
- Merah Tua menunjukkan hubungan negatif (berlawanan) yang kuat
- Kuning menunjukkan hubungan yang lemah atau hampir tidak ada

Sel-sel diagonal selalu menunjukkan nilai 1,0 (berwarna hijau pekat) karena setiap variabel memiliki hubungan yang sempurna dengan dirinya sendiri. Ini merupakan persyaratan matematis standar dari matriks tersebut dan dapat diabaikan saat Anda melakukan analisis data yang sesungguhnya.

Pengalaman vs. Kinerja

Kinerja cenderung meningkat seiring dengan bertambahnya tahun pengalaman; namun, pengalaman hanya mencakup 30% dari variasi kinerja. Hal ini menunjukkan bahwa pengalaman saja bukanlah prediktor yang dapat diandalkan untuk kinerja tinggi, dan terdapat banyak faktor lain yang berkontribusi terhadap hasil kinerja. Sebaliknya, peningkatan jam pelatihan dikaitkan dengan perbaikan kinerja yang lebih konsisten dan dapat diprediksi. Pelatihan menjelaskan 66% dari variasi kinerja, menjadikannya faktor yang jauh lebih efektif untuk meningkatkan kinerja.

7.7 MENGANALISIS DAN MENAFSIRKAN HUBUNGAN ANTARVARIABEL

Dalam bab ini, Anda telah mempelajari cara menganalisis dan menafsirkan hubungan antarvariabel sebuah langkah penting dalam analisis data eksploratif dan penalaran statistik. Memahami bagaimana variabel-variabel bergerak bersama membantu analis dalam

menemukan pola, menghasilkan wawasan, dan mendukung pengambilan keputusan. Anda memulai dengan mengkaji hubungan antara variabel kategorikal menggunakan tabel kontingensi dan perbandingan persentase. Dengan mengubah jumlah (frekuensi) menjadi persentase, Anda belajar menentukan apakah distribusi suatu variabel berubah di berbagai kategori variabel lainnya. Jika persentase menunjukkan perbedaan yang signifikan antar kategori, kemungkinan besar variabel-variabel tersebut saling berhubungan (berasosiasi) dan tidak independen. Selanjutnya, Anda mengeksplorasi hubungan antara variabel kategorikal dan variabel numerik. Dalam kasus ini, analisis berfokus pada perbandingan hasil numerik di berbagai kategori. Alat seperti rata-rata kelompok, standar deviasi, dan *box plot* yang disandingkan membantu mengungkapkan apakah kelompok tertentu cenderung memiliki nilai yang lebih tinggi atau lebih rendah, serta apakah distribusi datanya berbeda.

Kemudian, Anda mempelajari hubungan antara variabel-variabel numerik menggunakan *scatter plot* (diagram pencar) dan analisis korelasi. *Scatter plot* memungkinkan Anda untuk mengidentifikasi pola secara visual, sementara koefisien korelasi (r) mengukur kekuatan dan arah hubungan linear. Anda juga mempelajari bagaimana r^2 (koefisien determinasi) menunjukkan proporsi variasi dalam satu variabel yang dikaitkan dengan variasi dalam variabel lainnya. Bab ini juga menyoroti keterbatasan penting dari korelasi. Anda telah mempelajari bahwa korelasi tidak menyiratkan sebab-akibat, bahwa korelasi hanya mengukur hubungan linear, serta bahwa faktor-faktor seperti pencilan (*outliers*) atau rentang data yang terbatas dapat mendistorsi hasil.

Peringatan-peringatan ini menegaskan pentingnya menggabungkan ukuran statistik dengan pemeriksaan visual yang cermat dan pemikiran kritis. Anda juga telah mempelajari cara menganalisis berbagai hubungan secara bersamaan menggunakan matriks korelasi dan peta panas (*heatmaps*), yang menyajikan gambaran ringkas mengenai bagaimana beberapa variabel numerik saling berhubungan. Terakhir, bab ini merangkum gagasan-gagasan tersebut melalui kerangka kerja pengambilan keputusan yang membantu analis dalam memilih metode yang tepat berdasarkan jenis variabelnya.

Baik saat menganalisis variabel kategorikal, numerik, maupun campuran, pemilihan pendekatan yang tepat akan memastikan bahwa hubungan antarvariabel dapat dieksplorasi secara efektif dan diinterpretasikan dengan benar. Di akhir bab ini, Anda telah memperoleh perangkat praktis untuk mengidentifikasi dan menafsirkan hubungan dalam data. Keterampilan ini menjadi landasan krusial untuk tahap analisis berikutnya. Pada bab selanjutnya, Anda akan mengembangkan pemahaman ini lebih jauh dengan mempelajari analisis regresi, yang memodelkan hubungan secara matematis dan memungkinkan Anda untuk membuat prediksi berdasarkan data.

BAB 8
PENGANTAR ANALISIS REGRESI

8.1 JEMBATAN ANTARA DESKRIPSI DAN PREDIKSI

Pada bab sebelumnya, kita telah mempelajari cara mengukur dan menafsirkan korelasi yaitu kekuatan dan arah hubungan antara variabel-variabel numerik. Dengan menggunakan alat seperti koefisien korelasi Pearson (r), Anda telah melihat bagaimana dua variabel bergerak bersama secara sistematis. Namun, menemukan adanya hubungan secara alami akan memunculkan pertanyaan yang lebih mendalam: Jika dua variabel saling berhubungan, bisakah kita menggunakan hubungan tersebut untuk membuat prediksi?

Pertanyaan ini menandai titik balik penting dalam perjalanan Anda mendalami data. Meskipun korelasi menunjukkan bahwa suatu hubungan memang ada, korelasi tidak menyediakan mekanisme untuk mengestimasi nilai yang belum diketahui. Di sinilah analisis regresi berperan. Anggaplah korelasi sebagai upaya untuk mengidentifikasi adanya keterkaitan; sedangkan regresi mendeskripsikan keterkaitan tersebut melalui persamaan matematis. Setelah memiliki persamaan ini, kita dapat menggunakannya untuk memprediksi nilai suatu variabel berdasarkan variabel lainnya.

Catatan tentang Posisi Kita dalam Pembelajaran Ini

Anda mungkin bertanya-tanya mengapa regresi dibahas di sini dalam bagian statistika deskriptif dan bukan nanti pada topik-topik yang lebih "lanjutan". Alasannya sederhana: regresi memiliki dua peran yang berbeda.

1. Deskriptif & Prediktif: Regresi membantu kita mendeskripsikan pola matematis dalam data yang sudah ada. Kita menarik sebuah garis secara harfiah yang paling tepat menggambarkan tren tersebut, sehingga kita dapat membuat prediksi mengenai pengembangan teknologi masa depan baru. Inilah fokus kita untuk saat ini.
2. Inferensial: Nantinya, kita akan membahas pertanyaan yang lebih mendalam, seperti:
   - Apakah hubungan yang teramati signifikan secara statistik?
   - Apakah hubungan tersebut kemungkinan besar juga ada dalam populasi yang lebih luas?
   - Seberapa yakin kita terhadap prediksi yang dibuat?

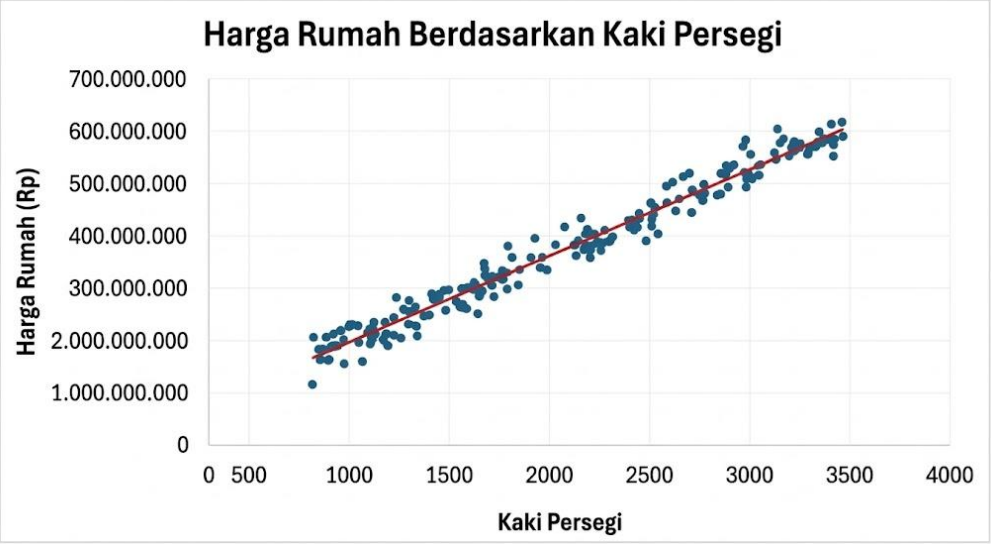
Pertanyaan-pertanyaan tersebut melibatkan inferensi statistik, di mana kita beralih dari data sampel untuk menarik kesimpulan mengenai populasi. Kita akan kembali membahas regresi dari perspektif tersebut pada Bab 14. Untuk saat ini, kesampingkan dulu kekhawatiran mengenai *p-value*, pengujian hipotesis, dan parameter populasi. Dalam bab ini, tujuan kita bersifat intuitif: menggunakan regresi sebagai alat untuk mendeskripsikan hubungan secara presisi dan membuat prediksi yang berdasar.

8.2 DARI KORELASI MENUJU PREDIKSI: MENARIK GARIS

Bayangkan Anda adalah seorang agen properti yang sedang mempelajari data perumahan. Anda memperhatikan bahwa rumah yang lebih besar cenderung terjual dengan harga yang lebih tinggi, dan perhitungan korelasi Anda mengonfirmasi adanya hubungan positif yang kuat. Kini, seorang klien mengajukan pertanyaan spesifik: "Berapa harga yang kira-kira harus saya bayar untuk sebuah rumah seluas 2.500 kaki persegi?" Korelasi tidak dapat menjawab pertanyaan ini. Korelasi hanya menunjukkan bahwa ukuran dan harga bergerak searah, tetapi tidak menjelaskan "tingkat perubahannya." Di sinilah analisis regresi menjadi penting. Analisis ini melampaui sekadar mengidentifikasi hubungan, dengan menyediakan model matematis untuk mengestimasi nilai-nilai tertentu.

Transisi Visual

Saat kita memplot ukuran dan harga rumah pada diagram pencar (*scatterplot*), titik-titik data tersebut biasanya membentuk "awan" yang mengarah ke atas. Secara alami, mata kita cenderung ingin menarik garis menembus "awan" tersebut untuk merepresentasikan tren yang ada.



Gambar 8.1 Diagram sebar harga properti berdasarkan luas persegi yang memvisualisasikan garis yang paling sesuai

Tetapi apa yang membuat sebuah garis menjadi representasi "terbaik"? Haruskah garis tersebut melewati titik tertinggi, terendah, atau tepat di tengah? Meskipun mata kita dapat memperkirakan tren, analisis regresi menyediakan metode matematika yang tepat metode Kuadrat Terkecil Biasa (*Ordinary Least Squares*/OLS) untuk mengidentifikasi garis regresi. Garis ini adalah dasar dari semua pemodelan prediktif.

8.3 MEMAHAMI ANALISIS REGRESI

Analisis regresi adalah landasan ilmu data yang memungkinkan kita untuk mengukur hubungan antar variabel. Sementara diagram sebar memungkinkan kita untuk memvisualisasikan suatu hubungan, regresi mengubah intuisi visual tersebut menjadi

persamaan matematika yang tepat. Pada intinya, regresi menjawab pertanyaan mendasar: "Sejauh mana satu variabel memengaruhi variabel lain?"

Aplikasi di Dunia Nyata:

- **Real Estat:** Bagaimana setiap tambahan satu kaki persegi ruang memengaruhi harga pasar rumah?
- **Sumber Daya Manusia:** Seberapa besar peningkatan gaji seorang profesional setelah satu tahun pengalaman?
- **Pemasaran:** Bagaimana peningkatan pengeluaran iklan sebesar Rp 10.000.000 dapat diterjemahkan ke dalam total volume penjualan?

Tiga Pilar Regresi

Regresi membantu kita mencapai tiga tujuan utama:

1. Menggambarkan: Mengidentifikasi kekuatan dan arah (positif atau negatif) suatu hubungan.
2. Memprediksi: Memperkirakan hasil di masa depan berdasarkan data historis yang diketahui.
3. Mengkuantifikasi: Mengukur dampak pasti dari faktor-faktor tertentu pada hasil target.

Istilah Kunci: Variabel

Setiap model regresi dibangun berdasarkan dua jenis variabel:

- Variabel Independen (Prediktor / X): Ini adalah variabel yang digunakan untuk menjelaskan atau memprediksi perubahan. Ini adalah "input". Dalam contoh kita, luas bangunan adalah variabel independent (X) karena kita menggunakannya untuk memperkirakan harga.
- Variabel Dependen (Hasil / Y): Ini adalah variabel yang ingin kita prediksi. Nilainya "bergantung" pada input. Di sini, harga rumah adalah variabel dependen (Y). Nilainya "bergantung" pada luas bangunan.

Apa itu Regresi Linier?

Bentuk analisis regresi yang paling banyak digunakan adalah Regresi Linier. Regresi ini memodelkan hubungan garis lurus antara variabel numerik. Jika Anda memplot data Anda pada diagram sebar, regresi linier mengidentifikasi "Garis Kecocokan Terbaik" garis lurus tunggal yang sedekat mungkin dengan semua titik data.

Jenis-Jenis Regresi Linier

Tergantung pada kompleksitas data Anda dan jumlah prediktor yang terlibat, Anda akan menggunakan salah satu jenis berikut:

1. **Regresi Linier Sederhana:**
   Menggunakan satu variabel independen untuk memprediksi satu variabel dependen. Contoh: Hanya menggunakan Luas Bangunan untuk memprediksi Harga Rumah.
2. **Regresi Linier Berganda:**
   Menggunakan dua atau lebih variabel independen untuk memprediksi satu variabel dependen. Ini memperhitungkan kenyataan bahwa hasil jarang didorong hanya oleh satu faktor.

Contoh: Menggunakan Luas Bangunan, Jumlah Kamar Tidur, dan Skor Lokasi untuk memprediksi Harga Rumah.

3. **Regresi Polinomial:**
   Regresi polinomial memperluas regresi linier untuk memodelkan hubungan lengkung antar variabel. Meskipun hubungan antar variabel mungkin tampak melengkung, model tetap linier dalam koefisiennya. Hal ini dicapai dengan menyertakan suku tambahan seperti suku kuadrat (x^2) atau suku kubik (x^3) untuk menangkap lekukan pada data.
   Contoh: Memodelkan bagaimana kecepatan mobil memengaruhi efisiensi bahan bakar, di mana efisiensi menurun tajam pada kecepatan yang sangat tinggi.

8.4 REGRESI LINEAR SEDERHANA: KISAH TENTANG GARIS LURUS

Regresi linear sederhana mengasumsikan bahwa nilai dari satu variabel (dilambangkan dengan X) dapat membantu menjelaskan atau memprediksi nilai dari variabel lain (dilambangkan dengan Y). Metode ini mengkaji hubungan linear antara dua variabel numerik dengan menyesuaikan sebuah garis lurus yang paling tepat menggambarkan bagaimana Y cenderung berubah seiring dengan perubahan X. Garis lurus ini, yang dikenal sebagai garis regresi, memungkinkan kita untuk mendeskripsikan hubungan tersebut secara matematis dan membuat prediksi mengenai Y berdasarkan nilai X. Seperti halnya garis lurus pada umumnya, garis regresi dapat dinyatakan dalam bentuk aljabar yang sudah familier:

$$Y = \beta_0 + \beta_1 X$$

Dalam dunia yang ideal, setiap titik data akan tepat berada pada garis ini. Namun, data di dunia nyata jarang berperilaku sesederhana itu. Observasi akan menyimpang dari garis tersebut karena adanya:

- variabilitas alami dalam fenomena yang diteliti
- kesalahan pengukuran
- faktor-faktor yang tidak diamati atau dihilangkan
- keacakan yang melekat dalam sistem dunia nyata

Untuk memperhitungkan variabilitas yang tidak dapat dihindari ini, para ahli statistik memperkenalkan suku galat acak (*random error term*), yang dilambangkan dengan ε (*epsilon*). Suku ini merepresentasikan bagian dari Y yang tidak dapat dijelaskan oleh X saja; inilah alasan mengapa titik-titik data menyimpang dari garis regresi. Dengan demikian, model regresi populasi yang lengkap persamaan yang menggambarkan bagaimana hasil yang sebenarnya berperilaku di dunia nyata menjadi:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Di mana:

- Y = hasil aktual yang diamati

- X = variabel penjelas (independen)
- β_0 (intersep) = nilai harapan Y ketika X = 0
- β_1 (kemiringan/*slope*) = perubahan rata-rata pada Y untuk setiap peningkatan satu unit pada X
- ε (*epsilon*) = galat acak yang merepresentasikan bagian dari Y yang tidak dapat dijelaskan oleh X

Persamaan ini menggambarkan bagaimana hasil yang sebenarnya berperilaku dalam kenyataan, termasuk unsur keacakan yang tidak dapat ditangkap sepenuhnya oleh model apa pun.

Garis Regresi: Prediksi Terbaik Kita

Dalam praktiknya, kita hampir tidak pernah memiliki akses terhadap data untuk seluruh populasi, sehingga nilai parameter yang sebenarnya, yaitu β_0 dan β_1, tidak diketahui. Sebagai gantinya, kita menggunakan data sampel untuk menghitung estimasi, yang secara konvensional dilambangkan dengan b_0 dan b_1. Dengan menggunakan estimasi-estimasi tersebut, kita menyusun garis regresi estimasi:

$$\hat{Y} = b_0 + b_1 X$$

Di mana:

- $\hat{y}$ (y-hat) = nilai prediksi Y
- b_0 = estimasi intersep (titik potong)
- b_1 = estimasi kemiringan (*slope*)

Penting untuk dicatat bahwa persamaan ini menghasilkan nilai prediksi, bukan nilai observasi. Persamaan ini menyatakan bahwa untuk setiap nilai X, nilai prediksi y adalah $b_0 + b_1 X$. Suku galat (*error term*) ε tidak disertakan karena, menurut definisinya, nilai ini tidak dapat diketahui untuk setiap observasi individu, kita hanya dapat mengestimasi bagian sistematis dari hubungan tersebut. Perbedaan ini menjelaskan mengapa, dalam diagram pencar (*scatterplot*), titik-titik data berkumpul di sekitar garis regresi alih-alih tepat berada di atas garis tersebut: garis regresi menangkap pola sistematis dalam data, sementara ε mencakup segala hal yang tidak dapat dijelaskan oleh model.

Logika Garis "Kecocokan Terbaik" (*Best Fit*)

Saat kita menarik garis regresi melalui data kita, kita menginginkan garis tersebut sedekat mungkin dengan titik-titik data yang ada. Namun, apa sebenarnya arti dari "kecocokan terbaik" (*best fit*)? Dan bagaimana cara menentukan nilai b_0 serta b_1 yang menghasilkan garis "kecocokan terbaik" tersebut? Jawabannya terletak pada upaya meminimalkan galat (*error*) antara nilai y observasi dan nilai y prediksi. Galat ini, yang dikenal sebagai residu, mengukur selisih antara nilai observasi aktual dengan nilai prediksi ($\hat{y}$).

Residu: Mengukur Galat Prediksi

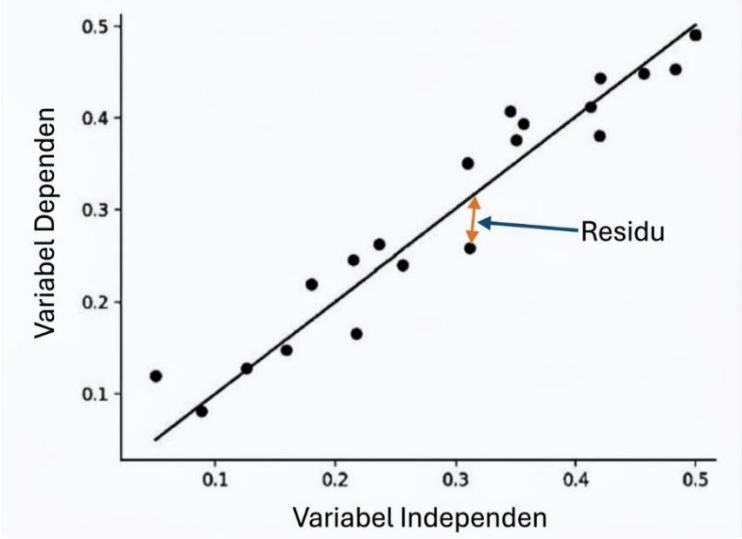
Misalkan kita memprediksi harga rumah berdasarkan luas bangunannya (dalam kaki persegi). Untuk setiap rumah, model regresi menghasilkan harga prediksi $\hat{Y}$, sementara harga jual aktualnya adalah Y. Selisih di antara keduanya disebut sebagai residu:

$$Residu = Y - \hat{Y}$$

Residu merepresentasikan galat prediksi untuk setiap observasi.

- Residu positif berarti model memberikan estimasi yang lebih rendah daripada nilai sebenarnya.
- Residu negatif berarti model memberikan estimasi yang lebih tinggi daripada nilai sebenarnya.

Secara grafis, residu merupakan jarak vertikal antara setiap titik data dengan garis regresi.



Gambar 8.2 Diagram pencar yang memvisualisasikan garis "kecocokan terbaik" (*best-fit*) dan residu

Jika setiap titik jatuh tepat pada garis regresi, maka semua residu akan bernilai nol. Namun, karena data nyata mengandung unsur keacakan dan pengaruh-pengaruh yang tidak teramati, prediksi yang sempurna hampir tidak mungkin dicapai. Oleh karena itu, karena hampir tidak mungkin bagi setiap titik untuk jatuh tepat pada garis regresi, tujuan dari algoritma regresi adalah menemukan garis yang membuat residu-residu tersebut sekecil mungkin di seluruh titik data, inilah yang disebut sebagai garis kecocokan terbaik (*best-fit line*). Algoritma ini memilih nilai b_0 dan b_1 sedemikian rupa sehingga jumlah kuadrat residu menjadi sekecil mungkin.

Metode Kuadrat Terkecil

Metode yang paling banyak digunakan untuk mengestimasi garis regresi kecocokan terbaik disebut dengan *Ordinary Least Squares* (OLS) atau Metode Kuadrat Terkecil Biasa. Metode ini memilih nilai b_0 dan b_1 yang meminimalkan total kuadrat residu, yang didefinisikan sebagai berikut:

$$\sum (Y - \hat{Y})^2$$

Mengapa residual dikuadratkan?

Ada dua alasan penting:

1. Penguadratan menghilangkan tanda negatif, sehingga kesalahan positif dan negatif tidak saling meniadakan.
2. Kesalahan yang besar diberi penalti (bobot hukuman) yang lebih berat, memastikan bahwa garis yang tidak cocok dengan titik-titik data tertentu akan mendapatkan penalti yang lebih besar.

Garis regresi yang dihasilkan oleh OLS secara matematis dijamin sebagai garis yang meminimalkan keseluruhan kesalahan prediksi di seluruh kumpulan data.

Fungsi Biaya (*Cost Function*)

Metode kuadrat terkecil (*least squares*) menggunakan fungsi biaya untuk mengukur seberapa baik garis tersebut mencocoki data. Salah satu fungsi biaya yang paling umum digunakan adalah *Mean Squared Error* (MSE).

$$MSE = \frac{1}{n}\sum\left(Y_i - \hat{Y}_i\right)^2$$

Di mana:

- MSE = *Mean Squared Error* (Rata-rata Kuadrat Galat)
- n = jumlah observasi
- Y_i = nilai observasi aktual
- $\hat{Y}_i$ = nilai prediksi

Perhitungan ini mengikuti tiga langkah:

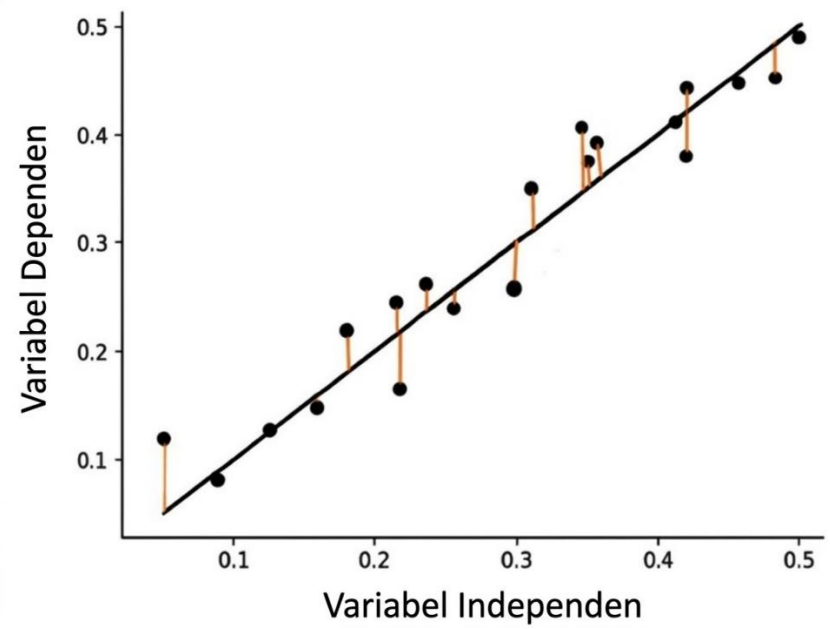
1. Hitung residual (selisih antara nilai observasi dan nilai prediksi).
2. Kuadratkan setiap residu.
3. Hitung rata-rata dari residu yang telah dikuadratkan tersebut.

Garis regresi kecocokan terbaik adalah garis yang meminimalkan MSE, sehingga menghasilkan kesalahan prediksi keseluruhan yang paling kecil.

Merangkum Semuanya

Tabel 8.1 Referensi cepat regresi linear

Konsep	**Makna**
$Y = \beta_0 + \beta_1 X + \varepsilon$	Model populasi yang sebenarnya, mencakup galat acak
$\hat{y} = b_0 + b_1 X$	Persamaan prediksi yang diestimasi dari data sampel
Titik-titik scatterplot (Y)	Observasi nyata yang mengandung variabilitas
Garis regresi ($\hat{y}$)	Pendekatan matematis terbaik terhadap tren
Residual	Selisih antara nilai observasi dan nilai prediksi
Kuadrat Terkecil	Metode yang digunakan untuk menemukan garis yang paling sesuai (*best-fitting line*)
Mengapa titik-titik tersebar	Data nyata mengandung unsur keacakan dan pengaruh-pengaruh yang tidak teramati



Gambar 8.3 Diagram pencar (*scatter plot*) yang memvisualisasikan garis kecocokan terbaik (*best-fit line*) dan residu

Penerapan Regresi Linear Sederhana Dengan Excel

Mari kita terapkan konsep-konsep ini dalam praktik. Pemeriksaan data modern jarang memerlukan perhitungan parameter regresi secara manual. Sebagai gantinya, kita menggunakan perangkat lunak seperti Excel, paket statistik, atau bahasa pemrograman seperti Python. Excel menyediakan beberapa cara untuk melakukan pemeriksaan regresi mulai dari alat visual hingga keluaran statistik lengkap.

Memvisualisasikan Garis Regresi: *Trendline*

Diagram pencar (*scatterplot*) adalah alat utama untuk memvisualisasikan hubungan antara variabel numerik. Setelah memplot data Anda (X pada sumbu horizontal dan Y pada sumbu vertikal), Anda dapat menambahkan garis regresi untuk memvisualisasikan garis yang paling sesuai (*best-fit line*).

Langkah-langkah Menambahkan Garis Regresi

1. Buat diagram pencar dari data Anda.
2. Klik kanan pada salah satu titik data.
3. Pilih *Add Trendline* (Tambahkan Garis Tren).
4. Pada panel Format Trendline, pilih Linear.

Menampilkan Persamaan Regresi

Untuk melihat persamaan regresi dan koefisien determinasi:

1. Centang opsi *Display Equation on Chart* (Tampilkan Persamaan pada Grafik).
2. Centang opsi *Display R-squared Value on Chart* (Tampilkan Nilai *R-squared* pada Grafik).

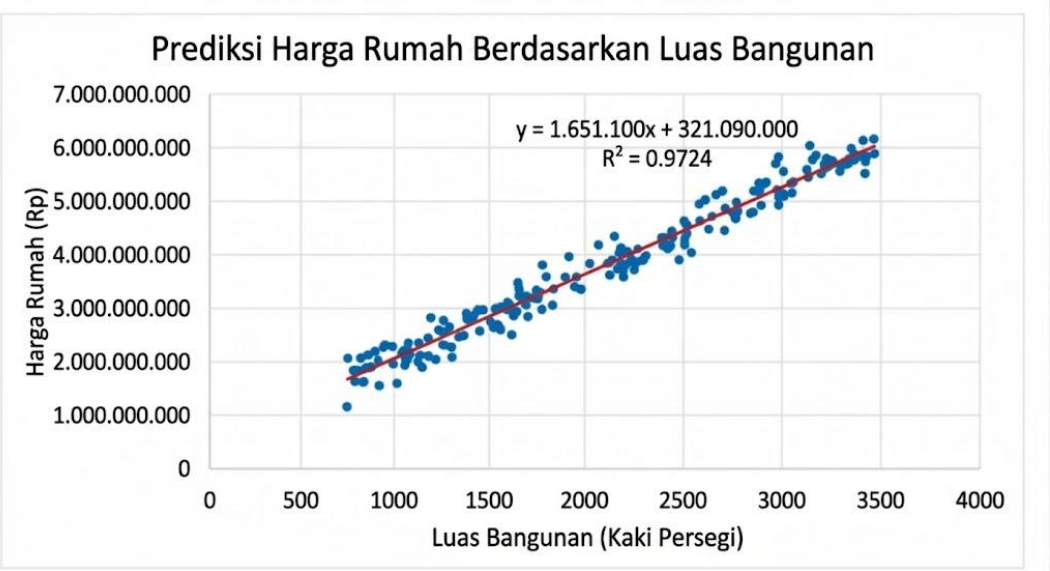
Opsi-opsi ini akan menampilkan model regresi yang diestimasi langsung pada grafik.

Memformat Garis

Agar garis regresi lebih jelas:

1. Buka Format *Trendline*.

2. Pilih *Solid Line* (Garis Utuh) alih-alih garis putus-putus.
3. Pilih warna yang kontras (seperti merah).



Gambar 8.4 Diagram pencar (*scatter plot*) yang menampilkan garis kecocokan terbaik (*best-fit line*), persamaan regresi, dan R^2.

Misalkan Excel menampilkan persamaan berikut:

$$y = 165.11x + 32109$$

Ini menunjukkan kepada kita:

- Kemiringan (b_1) = 165,11
- Intersep (b_0) = 32.109 Interpretasi:

Setiap penambahan satu kaki persegi (square foot) meningkatkan prediksi harga rumah sebesar kurang lebih 1.650 juta. Selain itu, berdasarkan nilai R^2 (0,9724), luas bangunan (dalam kaki persegi) menjelaskan sekitar 97% variasi harga rumah dalam sampel tersebut.

Menghitung Parameter dengan Fungsi Excel

Jika Anda memerlukan nilai numerik secara langsung, Excel menyediakan fungsi-fungsi bawaan.

Tabel 8.2 Fungsi regresi linear

Parameter	**Fungsi Excel**
Kemiringan	= *SLOPE*(known_y's, known_x's)
Intersep	=*INTERCEPT*(known_y's, known_x's)
R^2	=*RSQ*(known_y's, known_x's)
Korelasi	=*CORREL*(array1, array2)

Penting: Selalu pilih nilai-nilai Y terlebih dahulu, diikuti oleh nilai-nilai X.

Terapkan fungsi-fungsi Excel pada kumpulan data yang sama. Kemudian, gunakan nilai kemiringan (*slope*) dan perpotongan (*intercept*) untuk menyusun persamaan (model tersebut). Hasilnya:

$$\hat{Y} = 32109 + 165.11X$$

Ini adalah persamaan yang sama yang ditampilkan pada garis tren.

Menggunakan Alat Pemeriksaan DataPak

Untuk pemeriksaan regresi yang lebih lengkap:

1. Buka Data → Pemeriksaan Data → Regresi
2. Pilih Input Y Range (variabel terikat)
3. Pilih Input X Range (variabel independen)
4. Pilih lokasi keluaran
5. Secara opsional pilih Residual dan Residual Plot
6. Klik Oke

Excel akan menghasilkan Output Ringkasan Regresi, yang merupakan format standar yang digunakan dalam pemeriksaan statistik.

	A	B	C	D	E	F	G	H	I
1	RINGKASAN OUTPUT								
2									
3	*Statistik Regresi*								
4	R Majemuk	0.9861							
5	R Kuadrat	0.9724							
6	R Kuadrat Disesuaikan	0.9723							
7	Kesalahan Standa	22187.2648							
8	Pengamatan	200							
9									
10	ANOVA								
11		*dk*	*JK*	*RKT*	*F*	*Signifikansi F*			
12	Regresi	1	3.43909E+12	3.4391E+12	6986.1274	2.2279E-156			
13	Sisa	198	97470394140	492274718					
14	Total	199	3.53656E+12						
15									
16		*Koefisien*	*Kesalahan Stan*	*t Stat*	*p-value*	*Bawah 95%*	*Atas 95%*	*Bawah 95,0%*	*Atas 95,0%*
17	Intersep	32109.31	4447.75	7.22	1.0949E-11	23338.27	40880.36	23338.27	40880.36
18	LuasBangunan	165.11	1.98	83.58	2.228E-156	161.22	169.01	161.22	169.01

Gambar 8.5 *Output summary* regresi Excet – regresi rata-rata saat ini

Untuk saat ini, fokuslah pada hasil utama:

Statistik Regresi

- Kelipatan R → Koefisien korelasi (r)
- R Square → Koefisien determinasi (r^2)
- Adjusted R Square → dibahas nanti
- Kesalahan Standar → dibahas nanti

Tabel Koefisien

- Intersepsi → diperkirakan b_0
- SquareFootage → kemiringan b_1

Membuat Prediksi dengan Regresi

Kami sekarang siap menjawab pertanyaan jutaan dolar kami: Jika klien meminta Anda memperkirakan harga rumah seluas 2.500 kaki persegi, bagaimana tanggapan Anda? Menggunakan persamaan regresi yang diturunkan sebelumnya:

$$\hat{Y} = 32109 + 165{,}11X$$

Dengan menyubstitusikan X = 2.500:

$$\hat{Y} = 32109 + 165{,}11(2.500)$$
$$\hat{Y} = 444.884$$

Ini berarti model kita memprediksi bahwa sebuah rumah dengan luas 2.500 kaki persegi akan terjual dengan harga sekitar 4.448.840 juta.

Menggunakan Excel untuk Membuat Prediksi

Excel menyediakan fungsi bawaan untuk menghasilkan prediksi secara langsung dari data.

Fungsi Forecast

$$= FORECAST.LINEAR(x, known_y's, known_x's)$$

Fungsi ini memprediksi nilai Y untuk nilai X tertentu. Fungsi ini menggunakan regresi linear sederhana di baliknya.

Contoh:

$$= FORECAST.LINEAR(2500, D2{:}D201, A2{:}A201)$$

Hasil:

$$4.448.920\ Juta$$

Fungsi Trend

$$= TREND(known_y's, known_x's, new_x's)$$

Menghitung satu atau beberapa nilai prediksi. Fungsi ini dapat menghasilkan serangkaian nilai Y prediksi untuk rentang nilai X yang baru. Perbedaan kecil antara Rp 4.448.840.000 dan Rp 4.448.920.000 terjadi karena persamaan regresi yang ditampilkan pada grafik telah dibulatkan, sedangkan fungsi Excel melakukan perhitungan dengan menggunakan presisi penuh dari koefisien yang diestimasi.

Interpolasi vs. Ekstrapolasi

Saat menggunakan regresi untuk melakukan prediksi, penting untuk membedakan antara interpolasi dan ekstrapolasi.

Interpolasi

Interpolasi merujuk pada pembuatan prediksi di dalam rentang data yang diamati. Dalam kumpulan data perumahan, misalnya, anggaplah luas bangunan yang diamati berkisar antara 815 hingga 3.465 kaki persegi. Memprediksi harga untuk luas 2.500 kaki persegi termasuk dalam rentang ini, sehingga disebut sebagai interpolasi. Prediksi yang dibuat melalui interpolasi umumnya lebih dapat diandalkan karena berada dalam domain data yang telah diamati.

Ekstrapolasi

Ekstrapolasi merujuk pada prediksi nilai di luar rentang data yang diamati. Sebagai contoh, memprediksi harga sebuah rumah mewah seluas 7.000 kaki persegi menggunakan model kita merupakan bentuk ekstrapolasi. Hal ini bisa berisiko karena hubungan yang teramati dalam data mungkin tidak berlanjut di luar rentang aslinya. Dinamika pasar untuk rumah mewah, misalnya, mungkin mengikuti pola yang berbeda dibandingkan dengan rumah berukuran menengah. Aturan umum: Semakin jauh prediksi Anda berada di luar rentang data yang diamati, maka prediksi tersebut akan menjadi semakin kurang dapat diandalkan.

Mengevaluasi Kecocokan Model Regresi

Tidak semua model regresi dapat menangkap hubungan dengan tingkat akurasi yang sama. Terkadang, titik-titik data berkumpul dengan rapat di sekitar garis regresi, yang menunjukkan hubungan yang kuat. Dalam kasus lain titik-titik tersebut tersebar luas, sehingga membuat prediksi menjadi kurang dapat diandalkan. Untuk mengevaluasi kualitas model regresi, para ahli statistik menggunakan beberapa metrik kecocokan model (*goodness-of-fit*).

Koefisien Determinasi (R^2)

Salah satu ukuran kecocokan model yang paling banyak digunakan adalah R^2 (R-kuadrat), yang juga dikenal sebagai koefisien determinasi. R^2 menunjukkan proporsi variasi dalam variabel dependen (Y) yang dapat dijelaskan oleh variabel independen (X). Sebagaimana dibahas dalam Bab 7, dalam regresi linear sederhana, terdapat hubungan langsung antara korelasi dan R^2: koefisien determinasi merupakan kuadrat dari koefisien korelasi.

$$R^2 = r^2$$

Ini berarti koefisien determinasi hanyalah kuadrat dari koefisien korelasi.

Definisi R^2 Berdasarkan Varians

R^2 juga dapat dinyatakan menggunakan jumlah kuadrat:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Di mana:

- SS_{res} (Jumlah Kuadrat Residual): $\sum\left(Y_i - \hat{Y}_i\right)^2$
  Variasi yang tidak dapat dijelaskan—seberapa jauh nilai-nilai aktual menyimpang dari garis regresi.

- SS_{tot} (Jumlah Kuadrat Total): $\sum(Y_i - \bar{Y})^2$
  Variasi total dalam data—seberapa jauh nilai-nilai aktual menyimpang dari nilai rata-rata Y.

Rumus ini dimulai dengan angka 1 (yang merepresentasikan 100% varians dalam Y) kemudian dikurangi dengan proporsi yang tidak dapat dijelaskan oleh model. Hasil sisanya merupakan proporsi yang dijelaskan oleh garis regresi.

Menafsirkan R²

Nilai R^2	**Interpretasi**
0	X tidak menjelaskan sama sekali variasi dalam Y
1	X menjelaskan seluruh variasi dalam Y (*perfect fit* / kecocokan sempurna)
0.64	64% dari variasi dalam Y dijelaskan oleh X

Nilai R^2 sebesar 0,64 berarti 64% variasi dalam variabel y dapat dijelaskan oleh variabel x, sementara 36% sisanya disebabkan oleh faktor-faktor lain yang tidak dimasukkan ke dalam model.

Pemahaman Visual tentang R²

R^2 juga memiliki interpretasi visual yang jelas.

- Jika titik-titik dalam diagram pencar (*scatterplot*) tersebar secara acak, nilai R^2 akan mendekati 0.
- Jika titik-titik tersebut mengelompok rapat di sekitar garis lurus yang jelas, nilai R^2 akan mendekati 1.

Pada intinya, R^2 mengukur seberapa baik garis regresi menangkap pola dalam data.

R² yang Disesuaikan (*Adjusted* R²): Memperhitungkan Kompleksitas Model

Meskipun R^2 bersifat intuitif, terdapat keterbatasan penting: nilainya tidak akan pernah berkurang saat prediktor tambahan dimasukkan ke dalam model. Hal ini dapat menyesatkan dalam regresi berganda, di mana penambahan variabel yang tidak relevan dapat meningkatkan nilai R^2 secara semu. Untuk mengatasi masalah ini, para ahli statistik menggunakan R^2 yang Disesuaikan (*Adjusted* R^2), yang memberikan penalti terhadap kompleksitas yang tidak diperlukan.

$$Adjusted R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Di mana:

- n = jumlah observasi
- k = jumlah variabel independen

Berbeda dengan R^2, Adjusted R^2 dapat menurun ketika prediktor yang tidak relevan ditambahkan, menjadikannya ukuran yang lebih andal saat membandingkan model dengan jumlah prediktor yang berbeda.

Standar Error Estimasi (SE)

Ukuran kecocokan model (*model fit*) lainnya yang berguna adalah Standar Error Estimasi, yang sering disebut sebagai Standar Error Regresi. Statistik ini mengukur jarak rata-rata antara titik data observasi dan garis regresi.

Interpretasi

- SE yang lebih kecil → titik-titik data berada lebih dekat dengan garis regresi → kecocokan model lebih kuat
- SE yang lebih besar → titik-titik data lebih tersebar → akurasi prediksi lebih lemah

Berbeda dengan MSE, standar error dinyatakan dalam satuan yang sama dengan variabel dependen (Y). Rumusnya adalah:

$$SE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

Penyebutnya adalah $n - 2$ karena estimasi intersep dan kemiringan (*slope*) menghabiskan dua derajat kebebasan.

Metrik Kesalahan Lain yang Berguna

Beberapa metrik tambahan dapat membantu mengevaluasi kinerja model regresi.

***Mean Squared Error* (MSE)**

MSE mengukur rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual.

$$MSE = \frac{1}{n}\sum(Y - \hat{Y})^2$$

MSE yang lebih rendah menunjukkan akurasi prediksi yang lebih baik.

***Root Mean Squared Error* (RMSE)**

RMSE hanyalah akar kuadrat dari MSE:

$$RMSE = \sqrt{MSE}$$

Karena menghasilkan nilai kesalahan dalam satuan yang sama dengan variabel Y, RMSE lebih mudah untuk diinterpretasikan.

***Mean Absolute Error* (MAE)**

MAE mengukur rata-rata selisih mutlak antara nilai prediksi dan nilai aktual:

$$MAE = \frac{1}{n}\sum|Y - \hat{Y}|$$

Tidak seperti MSE dan RMSE, MAE tidak mengkuadratkan kesalahan, sehingga kurang sensitif terhadap outlier ekstrem.

Ringkasan: Menginterpretasikan Kinerja Model

Bersama-sama, metrik-metrik ini memberikan pandangan yang seimbang tentang kinerja model.

Tabel 8.3 Metrik Kinerja Model

Metrik	Apa yang ditunjukkannya
R^2	Kekuatan hubungan
R^2 yang Disesuaikan (*Adjusted* R^2)	Kualitas model yang disesuaikan dengan kompleksitas
SE	Jarak rata-rata titik data dari garis
MSE / RMSE	Kesalahan prediksi secara keseluruhan
MAE	Rata-rata kesalahan prediksi absolut

Penggunaan berbagai metrik memungkinkan analis untuk memahami kekuatan hubungan sekaligus akurasi prediksi.

8.5 REGRESI LINEAR BERGANDA

Dalam dunia nyata, suatu hasil jarang sekali dipengaruhi oleh satu faktor saja. Sebagai contoh, harga rumah tidak hanya bergantung pada luas bangunan. Harga tersebut juga dipengaruhi oleh banyak variabel lain seperti:

- Lokasi
- Jumlah kamar tidur
- Usia bangunan rumah
- Kualitas sekolah di sekitar lokasi
- Jarak ke pusat kota
- Fasilitas di lingkungan sekitar

Regresi linear sederhana memungkinkan kita untuk mempelajari hubungan antara satu prediktor dan satu hasil. Akan tetapi, fenomena di dunia nyata sering kali dibentuk oleh berbagai pengaruh secara bersamaan. Regresi linear berganda memperluas model regresi sederhana dengan menyertakan dua prediktor atau lebih. Bentuk umum dari model regresi berganda adalah:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Di mana:

- Y = variabel dependen (hasil)
- $X_1, X_2, \ldots, X_k$ = variabel independen (prediktor)
- β_o = intersep (konstanta)
- $\beta_1, \beta_2, \ldots, \beta_k$ = koefisien regresi
- ε = suku galat acak (*random error term*)

Contoh: Pemodelan Harga Rumah

Dalam kumpulan data perumahan, model regresi berganda dapat terlihat seperti ini:

$$Harga = \beta_0 + \beta_1(LuasBangunan) + \beta_2(JumlahKamarTidur) + \beta_3(KedekatanKeKota) + \varepsilon$$

Model ini berupaya menjelaskan harga rumah dengan menggunakan beberapa prediktor secara bersamaan. Dengan menyertakan berbagai variabel, model ini memberikan representasi yang lebih realistis mengenai cara kerja sistem di dunia nyata.

Interpretasi Koefisien

Dalam regresi linear berganda, setiap prediktor memiliki koefisien kemiringan (*slope*) tersendiri. Setiap koefisien menunjukkan perkiraan perubahan pada variabel dependen ketika prediktor tersebut meningkat sebesar satu satuan, dengan asumsi variabel prediktor lainnya tetap konstan. Konsep "menjaga variabel lain tetap konstan" ini sangatlah penting. Hal ini memungkinkan model untuk mengisolasi kontribusi unik dari setiap prediktor. Sebagai contoh, misalkan model regresi mengestimasi:

$$\beta_1 = 165 \, untuk \, variabel \, LuasBangunan.$$

Interpretasinya adalah:

Untuk setiap penambahan satu satuan luas bangunan, prediksi harga rumah akan meningkat sebesar Rp 1.650.000, dengan asumsi jumlah kamar tidur dan kedekatan ke kota tidak berubah. Kemampuan untuk mengendalikan berbagai pengaruh secara bersamaan merupakan salah satu aspek paling kuat dari regresi berganda. Hal ini memungkinkan analis untuk memisahkan dan mengevaluasi dampak individual dari variabel-variabel yang berbeda.

Menjalankan Regresi Linier Berganda di Excel

ToolPak Pemeriksaan Data Excel dapat melakukan pemeriksaan regresi berganda dengan mudah.

Kunjungi:

- Data → Pemeriksaan Data → Pilih Regresi, lalu klik OK.

Tentukan Inputnya

- Input Y Range: Pilih kolom variabel terikat.
- Input X Range: Pilih semua kolom prediktor (misalnya, A1:C201).
- Periksa Label
- Opsi Keluaran: Pilih di mana Anda ingin keluaran regresi muncul.

Diagnostik Opsional

Anda juga dapat memilih:

- Residu
- Petak Sisa
- Plot Kesesuaian Garis
- Plot Probabilitas Normal

Diagnostik ini membantu mengevaluasi apakah model regresi memenuhi asumsinya. Klik OK, dan Excel akan menghasilkan Output Ringkasan Regresi.

	A	B	C	D	E	F	G	H	I
1	HASIL RINGKASAN								
2									
3	*Statistik Regresi*								
4	R Majemuk	0.989							
5	R Kuadrat	0.977							
6	Adjusted R Disesuaikan	0.977							
7	Standard Error	20262.871							
8	Pengamatan	200							
9									
10	ANOVA								
11		df (dk)	SS (JK)	MS (KT)	F	Signifikansi F			
12	Regresi	3	3.45609E+12	1.15203E+12	2805.83292	1.0921E-160			
13	Residu	196	80474452883	410583943.3					
14	Total	199	3.53656E+12						
15									
16		Koefisien	Kesalahan St:	t Hitung	P-value	Batas Bawah	Batas Atas	Batas Bawah	Batas Atas
17	Intersep	48405.5	5092.6	9.5	7.3622E-18	38362.1	58448.9	38362.1	58448.9
18	Luas Bangunan	150.9	4.1	37.2	1.0362E-90	142.9	158.9	142.9	158.9
19	Kamar Tidur	9253.9	2594.4	3.6	0.00045412	4137.3	14370.4	4137.3	14370.4
20	JarakKePusatKota_km	-941.7	167.7	-5.6	6.6768E-08	-1272.4	-610.9	-1272.4	-610.9

Gambar 8.6 Output summary regresi tahunan Excet – regresi muttipte tiwear

Menafsirkan Output

Untuk saat ini, fokuslah pada komponen-komponen utama berikut dari output regresi:

1. **R^2 dan R^2 yang Disesuaikan (*Adjusted* R^2)**
   - R^2 menunjukkan proporsi variasi dalam variabel dependen yang dijelaskan oleh prediktor-prediktor yang ada.
   - *Adjusted* R^2 memperhitungkan jumlah prediktor dan memberikan penalti terhadap variabel yang tidak diperlukan.

   Karena nilai R^2 akan selalu meningkat setiap kali prediktor ditambahkan, Adjusted R^2 umumnya dianggap lebih andal dalam model regresi berganda.
2. **Tabel Koefisien**

   Kolom koefisien menyajikan estimasi parameter regresi. Setiap koefisien menunjukkan seberapa besar perubahan pada hasil prediksi ketika prediktor terkait meningkat sebesar satu unit, dengan asumsi prediktor lainnya tetap konstan. Koefisien-koefisien ini dapat digabungkan untuk membentuk persamaan regresi.
3. **Standar Error (Galat Baku)**

   Standar *error* mengukur ketidakpastian dalam estimasi koefisien. Nilai standar *error* yang lebih kecil menunjukkan estimasi yang lebih presisi.
4. **Diagnostik Residual**

   Plot residual membantu menilai apakah model regresi sudah tepat untuk data yang digunakan. Plot ini memungkinkan analis untuk memeriksa berbagai masalah, seperti:
   - non-linearitas
   - varians yang tidak sama (heteroskedastisitas)
   - pencilan (*outlier*)

- kesalahan spesifikasi model

Kita akan membahas diagnostik ini secara lebih mendetail pada bagian selanjutnya.

Memvisualisasikan Hubungan dalam Regresi Berganda

Meskipun memvisualisasikan hubungan menjadi lebih kompleks dengan adanya banyak prediktor, terdapat beberapa pendekatan yang dapat digunakan:

- Membuat plot sebar (*scatterplot*) antara variabel dependen dengan masing-masing prediktor secara terpisah
- Memeriksa plot residual
- Menggunakan plot diagnostik untuk menilai asumsi-asumsi model

Perangkat visual ini melengkapi output numerik regresi dan membantu membangun intuisi mengenai bagaimana model tersebut berperilaku.

8.6 RINGKASAN REGRESI LINEAR BERGANDA

Regresi linear berganda secara signifikan memperluas kekuatan pemeriksaan regresi dengan memungkinkan kita untuk menelaah bagaimana beberapa prediktor bekerja bersama dalam memengaruhi suatu hasil. Meskipun mekanisme model menjadi lebih kompleks, gagasan dasarnya tetap sama:

- mengestimasi koefisien yang mendeskripsikan hubungan antarvariabel
- menilai seberapa baik model tersebut mencocoki data
- menafsirkan kontribusi dari setiap prediktor

Dengan menyertakan banyak variabel, model regresi memberikan pemahaman yang lebih realistis dan mendalam mengenai pola-pola yang terdapat dalam data dunia nyata. Seiring dengan pendalaman Anda dalam pemodelan regresi, konsep-konsep dasar ini akan terus menjadi panduan dalam melakukan interpretasi dan pemeriksaan.

Sekilas tentang Regresi Polinomial

Terkadang, hubungan antarvariabel tidak bersifat linear. Sebagai contoh, hubungan antara usia dan pendapatan mungkin bermula pada tingkat yang relatif rendah di awal masa dewasa, meningkat selama paruh baya, lalu menurun seiring individu mendekati masa pensiun. Pola melengkung seperti ini tidak dapat ditangkap dengan baik oleh regresi linear sederhana maupun berganda, yang mengasumsikan hubungan berbentuk garis lurus antara prediktor dan variabel hasil. Regresi polinomial memperluas regresi linear dengan memungkinkan model untuk menangkap hubungan yang melengkung. Alih-alih mencocokkan garis lurus, model ini mencocokkan kurva polinomial yang dapat melengkung untuk mengikuti pola dalam data secara lebih akurat. Model regresi polinomial memiliki bentuk umum sebagai berikut:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

Keterangan:

- x^2, x^3, …, x^n adalah suku-suku berorde lebih tinggi yang memungkinkan model untuk menangkap kelengkungan.

- n merepresentasikan derajat polinomial.
  - Derajat 2 menghasilkan kurva kuadratik.
  - Derajat 3 menghasilkan kurva kubik.
- Model ini tetap termasuk dalam keluarga regresi linear karena koefisien (β) diestimasi secara linear, meskipun kurva yang dihasilkan bukan berupa garis lurus.

Contoh

Pertimbangkan prediksi harga rumah berdasarkan ukuran rumah. Untuk rumah berukuran kecil, harga mungkin meningkat pesat seiring bertambahnya ukuran. Untuk rumah berukuran sedang, peningkatannya mungkin lebih bertahap. Untuk rumah yang sangat besar, harga mungkin kembali melonjak tajam karena adanya fitur-fitur mewah. Garis lurus sederhana mungkin mengalami *underfit* (kurang pas) terhadap hubungan ini karena gagal menangkap perubahan-perubahan tersebut. Kurva polinomial dapat melengkung untuk mengikuti pola tersebut dengan lebih akurat.

Poin Penting untuk Diingat

- Regresi polinomial berguna untuk memodelkan pola non-linear dalam data.
- Polinomial berderajat tinggi dapat menyebabkan *overfitting*, di mana model sangat cocok dengan data pelatihan tetapi berkinerja buruk pada data baru.
- Memilih derajat polinomial yang tepat adalah hal penting dan sering kali dipandu oleh teknik evaluasi model, seperti validasi silang (*cross-validation*).

Singkatnya, regresi polinomial menyediakan perluasan fleksibel dari regresi linear, yang memungkinkan analis untuk menangkap tren melengkung ketika model garis lurus tidak lagi memadai.

8.7 ASUMSI REGRESI LINEAR

Agar pemeriksaan regresi menghasilkan hasil yang dapat diandalkan dan dapat diinterpretasikan, beberapa asumsi utama harus terpenuhi secara wajar.

Asumsi Regresi

1. **Linearitas**

   Hubungan antara variabel prediktor X dan variabel hasil Y harus mendekati linear. Jika hubungan yang sebenarnya sangat melengkung, model regresi garis lurus dapat menghasilkan kesimpulan yang menyesatkan. Anda dapat menilainya secara visual menggunakan diagram sebar (*scatterplot*). Jika titik-titik data membentuk pola yang mendekati garis lurus, maka asumsi linearitas kemungkinan besar terpenuhi.

2. **Independensi Observasi**

   Setiap observasi harus bersifat independen satu sama lain. Asumsi ini dapat dilanggar jika:

   - individu yang sama diukur berulang kali;
   - observasi dikumpulkan dari waktu ke waktu dengan ketergantungan temporal;
   - terdapat data yang berkelompok (*clustered*) atau terbagi dalam grup.

   Pelanggaran terhadap independensi dapat mendistorsi inferensi statistik dan mengarah pada kesimpulan yang menyesatkan.

3. **Homoskedastisitas**
   Variabilitas residual harus relatif konstan di seluruh nilai X. Secara praktis, ini berarti sebaran titik-titik secara vertikal di sekitar garis regresi harus tetap cukup konsisten, terlepas dari apakah nilai X kecil atau besar. Jika sebaran tersebut meningkat atau menurun secara sistematis, data menunjukkan heteroskedastisitas, yang dapat membuat estimasi kesalahan baku (*standard error*) menjadi tidak dapat diandalkan.
4. **Normalitas Residual (Untuk Inferensi)**
   Untuk pengujian hipotesis dan interval kepercayaan yang akan kita bahas pada bab-bab selanjutnya, residual harus berdistribusi mendekati normal. Asumsi ini tidak terlalu krusial jika model hanya digunakan untuk prediksi, namun menjadi penting saat melakukan inferensi statistik.
5. **Tidak Adanya Pencilan (*Outlier*) Ekstrem atau Titik yang Berpengaruh Kuat**
   Titik data yang ekstrem dapat memberikan pengaruh yang tidak proporsional terhadap garis regresi dan mendistorsi hubungan yang diestimasi. Oleh karena itu, penting untuk mengidentifikasi dan memeriksa pencilan atau observasi yang memiliki pengaruh sangat kuat sebelum menafsirkan hasil regresi.

Memeriksa Asumsi

Asumsi-asumsi ini dapat dievaluasi menggunakan beberapa pendekatan:

- Pemeriksaan visual melalui diagram pencar (*scatterplot*)
- Plot residual
- Uji statistik diagnostik (akan dibahas kemudian di Bagian III)
- Pengetahuan bidang terkait dan penalaran kontekstual

Jika asumsi-asumsi tersebut dilanggar, analis dapat mempertimbangkan untuk:

- Melakukan transformasi variabel
- Menggunakan model alternatif
- Menghapus observasi bermasalah secara hati-hati (dengan disertai justifikasi yang jelas)

Membuat Plot Diagnostik di Excel

Setelah menjalankan pemeriksaan regresi di Excel, plot diagnostik dapat membantu menentukan apakah asumsi-asumsi model telah terpenuhi dengan cukup baik.

1. **Plot Residual**

Plot residual merupakan salah satu alat diagnostik terpenting dalam pemeriksaan regresi. Plot ini menampilkan residual (selisih antara nilai yang diamati dan nilai prediksi):

$$Residual = y - \hat{y}$$

Residual diplot terhadap salah satu dari berikut ini:

- nilai prediksi, atau
- variabel prediktor

Excel dapat menghasilkan nilai residual secara otomatis jika opsi *Residuals* dipilih dalam kotak dialog regresi.

Langkah-langkah

Buat diagram pencar (*scatterplot*) dengan:

- Nilai prediksi pada sumbu X
- Nilai residual pada sumbu Y

Hal yang Perlu Diperhatikan

Model regresi yang baik biasanya menunjukkan:

- Sebaran titik-titik secara acak di sekitar angka nol. Potensi masalah akan terlihat sebagai:
- Pola melengkung → pelanggaran asumsi linearitas
- Sebaran berbentuk kipas → heteroskedastisitas
- Gugusan atau pola tertentu → kesalahan spesifikasi model

2. Plot Q–Q Normal

Plot Q–Q (*Quantile–Quantile*) membandingkan distribusi residual Anda dengan distribusi normal teoretis. Plot ini menempatkan:

- Kuantil normal yang diharapkan pada satu sumbu
- Nilai residual yang teramati pada sumbu lainnya

Jika residual mengikuti distribusi normal, titik-titik tersebut akan terletak kira-kira di sepanjang garis diagonal lurus. Penyimpangan sistematis dari garis ini dapat mengindikasikan:

- kemencengan (*skewness*)
- ekor yang tebal (*heavy tails*)
- penyimpangan dari kenormalan

8.8 PEDOMAN PRAKTIS ANALISIS REGRESI

Saat Anda mulai menerapkan regresi dalam pemeriksaan dunia nyata, panduan berikut dapat membantu memastikan hasil yang lebih andal dan bermakna.

1. **Mulailah dengan Diagram Pencar (*Scatterplot*)**

   Selalu visualisasikan data Anda sebelum menjalankan model regresi. Diagram pencar membantu Anda untuk:

   - menentukan apakah model linear sudah tepat
   - mengidentifikasi pencilan (*outliers*)
   - mendeteksi pola atau struktur yang tidak biasa

   Visualisasi memberikan intuisi yang berharga sebelum pemodelan formal dimulai.

2. **Periksa Asumsi Model**

   Hasil regresi tidak boleh diinterpretasikan begitu saja tanpa pemeriksaan. Selalu periksa:

   - plot residual
   - pencilan (*outliers*)
   - potensi pelanggaran asumsi linearitas atau varians konstan

   Model yang melanggar asumsinya dapat menghasilkan kesimpulan yang menyesatkan.

3. **Laporkan R^2 Bersama dengan Persamaan Regresi**
   Persamaan regresi menjelaskan bagaimana variabel-variabel saling berhubungan, sementara R^2 menunjukkan seberapa baik model tersebut menjelaskan data. Melaporkan keduanya memberikan gambaran yang lebih lengkap mengenai kinerja model.
4. **Interpretasikan Koefisien dalam Konteksnya**
   Angka saja tidak informatif kecuali dijelaskan dalam konteks yang relevan. Sebagai contoh:
   Alih-alih mengatakan:
   "Koefisien kemiringan (*slope*) adalah 3,05"
   Katakanlah:
   "Untuk setiap tambahan satu jam waktu belajar, nilai ujian meningkat rata-rata sekitar 3,05 poin."
   Konteks mengubah hasil statistik menjadi wawasan yang bermakna.
5. **Akui Adanya Ketidakpastian**
   Model regresi menggambarkan pola dalam data, bukan prediksi yang sempurna. Semua prediksi mengandung tingkat kesalahan tertentu. Nanti di Bab 14, kita akan mempelajari cara mengukur ketidakpastian ini menggunakan selang kepercayaan (*confidence intervals*) dan pengujian hipotesis.
6. **Pertimbangkan Signifikansi Praktis**
   Sebuah model mungkin memiliki performa statistik yang kuat, namun nilai praktisnya terbatas. Sebagai contoh, sebuah model mungkin dapat memprediksi nilai ujian dengan akurasi tinggi, tetapi jika tambahan delapan jam waktu belajar hanya meningkatkan nilai sebesar beberapa poin saja, maka dampak di dunia nyata mungkin sangat kecil. Selalu pertimbangkan apakah hasil tersebut bermakna dalam pengambilan keputusan praktis.

8.9 REGRESI DALAM KONTEKS: PENERAPAN DI DUNIA NYATA

Pemeriksaan regresi merupakan salah satu teknik statistik yang paling banyak digunakan di hampir semua bidang ilmiah maupun profesional. Karena memungkinkan analis untuk mengukur hubungan antarvariabel dan membuat prediksi, teknik ini telah menjadi alat fundamental untuk pengambilan keputusan berbasis data. Berikut adalah beberapa contoh bagaimana regresi diterapkan dalam praktik:

- **Ekonomi**
  Memprediksi pengeluaran konsumen berdasarkan faktor-faktor seperti pendapatan, suku bunga, dan tingkat ketenagakerjaan.
- **Kedokteran**
  Mengestimasi risiko penyakit menggunakan prediktor seperti usia, faktor gaya hidup, riwayat medis, dan penanda genetik.

- **Pendidikan**
  Memodelkan prestasi mahasiswa berdasarkan ukuran kelas, pengalaman guru, sumber daya sekolah, dan faktor sosial-ekonomi.
- **Pemasaran**
  Memperkirakan penjualan produk menggunakan variabel seperti pengeluaran iklan, musiman, penetapan harga, dan kondisi ekonomi.
- **Ilmu Lingkungan**
  Memprediksi tingkat polusi berdasarkan aktivitas industri, kondisi cuaca, kepadatan penduduk, dan kebijakan regulasi.
- **Psikologi**
  Mengkaji bagaimana ciri kepribadian, pengalaman hidup, dan lingkungan sosial memengaruhi hasil seperti kebahagiaan, tingkat stres, atau kesejahteraan.

Dalam setiap kasus tersebut, regresi menyediakan kerangka kerja matematis untuk memeriksa hubungan antarvariabel dan menghasilkan prediksi yang berdasar. Dengan mengukur bagaimana perubahan pada satu variabel berkaitan dengan perubahan pada variabel lainnya, regresi memungkinkan para peneliti dan praktisi untuk lebih memahami sistem yang kompleks serta mendukung pengambilan keputusan berbasis bukti.

8.10 KETERBATASAN REGRESI DAN FAKTOR YANG MEMENGARUHI KEANDALAN MODEL

Meskipun memiliki kemampuan dan fleksibilitas yang tinggi, pemeriksaan regresi memiliki keterbatasan-keterbatasan penting. Memahami keterbatasan ini sangatlah penting untuk menafsirkan hasil regresi secara bertanggung jawab.

1. **Korelasi Tidak Menyiratkan Kausalitas**
   Hubungan dalam regresi tidak secara otomatis menunjukkan hubungan sebab-akibat (kausalitas). Meskipun suatu variabel prediktor dapat memprediksi hasil dengan kuat, hal ini tidak berarti variabel tersebut merupakan penyebab dari hasil tersebut. Hubungan yang teramati bisa saja terjadi karena:
   - kedua variabel dipengaruhi oleh faktor ketiga yang mendasarinya, atau
   - hubungan tersebut muncul karena kebetulan atau adanya variabel pengganggu (*confounding variables*).

   Menetapkan kausalitas biasanya memerlukan rancangan eksperimen yang cermat, studi terkontrol, atau metode inferensi kausal khusus, bukan sekadar pemeriksaan regresi saja.
2. **Kualitas Model Bergantung pada Kualitas Data**
   Model regresi sepenuhnya bergantung pada kualitas data yang digunakan untuk menyusunnya. Jika data mengandung:
   - kesalahan pengukuran
   - sampel yang bias
   - variabel yang hilang atau didefinisikan dengan buruk

maka model regresi yang dihasilkan dapat memberikan prediksi yang menyesatkan atau tidak dapat diandalkan. Dalam sains data, prinsip ini sering diringkas dengan istilah "*garbage in, garbage out*" (sampah yang dimasukkan, sampah pula yang dihasilkan).

3. **Hubungan Nonlinier Mungkin Memerlukan Model Alternatif**
   Regresi linier mengasumsikan bahwa hubungan antara prediktor dan hasil bersifat mendekati linier (garis lurus). Namun, banyak hubungan di dunia nyata yang bersifat melengkung, kompleks, atau nonlinier. Dalam kasus seperti ini, model linier sederhana mungkin tidak cocok dengan data dan menghasilkan prediksi yang tidak akurat. Sebelum menerapkan model regresi, analis sebaiknya selalu memeriksa diagram pencar (*scatterplots*) dan visualisasi data untuk menentukan apakah asumsi linier tersebut masuk akal. Jika terdapat kelengkungan yang nyata, teknik seperti regresi polinomial atau model nonlinier lainnya mungkin lebih tepat untuk digunakan.
4. **Pencilan (*Outliers*) Dapat Sangat Memengaruhi Hasil**
   Observasi ekstrem yang dikenal sebagai pencilan atau *outliers* dapat memberikan pengaruh yang tidak proporsional terhadap garis regresi, terutama pada kumpulan data yang lebih kecil. Satu titik data yang tidak biasa saja dapat mengubah secara signifikan:
   - estimasi kemiringan (*slope*)
   - titik potong (*intercept*)
   - kecocokan model secara keseluruhan

   Oleh karena itu, analis harus selalu memeriksa data mereka untuk mendeteksi kemungkinan adanya pencilan dan menyelidiki apakah hal tersebut mencerminkan:
   - kesalahan dalam memasukkan data
   - masalah dalam pengukuran
   - observasi yang sah namun tidak biasa

   Pencilan tidak boleh dihapus secara otomatis; data tersebut harus diperiksa dengan cermat dan ditangani dengan dasar pemeriksaan yang jelas.

8.11 RINGKASAN ANALISIS REGRESI

Dalam bab ini, kita telah membahas pemeriksaan regresi, salah satu perangkat terpenting untuk memahami hubungan antarvariabel dan melakukan prediksi berdasarkan data. Jika korelasi menunjukkan apakah dua variabel bergerak bersamaan, regresi memungkinkan kita untuk memodelkan hubungan tersebut secara matematis dan menggunakannya untuk memprediksi hasil. Kita memulai dengan konsep regresi linear sederhana, di mana sebuah garis lurus dicocokkan dengan data untuk menggambarkan bagaimana variabel prediktor (X) berhubungan dengan variabel hasil (Y). Persamaan regresi menyediakan cara praktis untuk mengestimasi nilai Y berdasarkan nilai X yang teramati. Dua komponen utama dalam model ini kemiringan (*slope*) dan intersep membantu menjelaskan bagaimana perubahan pada variabel prediktor memengaruhi hasil yang diprediksi. Selanjutnya, kita mempelajari cara mengevaluasi kualitas model regresi. Ukuran seperti R^2 (koefisien determinasi) membantu kita memahami seberapa besar variasi dalam variabel hasil

yang dapat dijelaskan oleh model tersebut. Konsep seperti SS_{tot} (variasi total) dan SS_{res} (variasi residual) menjadi dasar untuk memahami seberapa baik garis regresi mencocokkan data. Kita juga memperkenalkan R^2 yang disesuaikan (*Adjusted R²*), yang memberikan evaluasi model lebih akurat ketika terdapat lebih dari satu variabel prediktor. Bab ini kemudian berlanjut ke pembahasan regresi berganda, di mana lebih dari satu variabel prediktor digunakan untuk menjelaskan variasi dalam variabel hasil. Hal ini memungkinkan analis untuk memodelkan hubungan dunia nyata yang lebih kompleks, di mana suatu hasil jarang sekali dipengaruhi oleh satu faktor saja.

Kita juga secara singkat memperkenalkan regresi polinomial, yang memungkinkan model regresi untuk menangkap hubungan melengkung dalam data ketika garis lurus tidak lagi memadai. Meskipun pendekatan ini dapat meningkatkan fleksibilitas model, penggunaannya harus dilakukan dengan hati-hati untuk menghindari *overfitting*, kondisi di mana model terlalu menyesuaikan diri dengan data pelatihan sehingga kinerjanya buruk saat diterapkan pada data baru. Memahami regresi juga mengharuskan kita untuk memastikan bahwa beberapa asumsi terpenuhi. Asumsi-asumsi tersebut meliputi linearitas, independensi observasi, homoskedastisitas (varians residual yang konstan), distribusi residual yang mendekati normal untuk keperluan inferensi, serta tidak adanya pencilan ekstrem yang dapat mendistorsi garis regresi. Perangkat diagnostik seperti plot residual dan plot Q–Q membantu analis dalam mengevaluasi apakah asumsi-asumsi tersebut telah terpenuhi.

Terakhir, kita telah meninjau penerapan regresi dalam dunia nyata di berbagai bidang, seperti ekonomi, kedokteran, pemasaran, pendidikan, ilmu lingkungan, dan psikologi. Meskipun regresi merupakan alat pemeriksaan yang ampuh, penggunaannya harus dilakukan dengan cermat. Beberapa keterbatasan penting yang perlu diperhatikan antara lain: korelasi tidak menyiratkan kausalitas (sebab-akibat), akurasi model sangat bergantung pada kualitas data, hubungan nonlinear mungkin memerlukan model alternatif, dan keberadaan pencilan dapat sangat memengaruhi hasil pemeriksaan. Secara keseluruhan, pemeriksaan regresi menyediakan kerangka kerja yang kuat untuk menemukan pola, menjelaskan hubungan antarvariabel, serta membuat prediksi. Jika digunakan dengan bijak disertai dengan eksplorasi dan interpretasi data yang cermat, regresi menjadi instrumen esensial bagi para ilmuwan data dan analis yang berupaya menggali wawasan bermakna dari data.

8.12 REGRESI SEBAGAI FONDASI ANALITIK PREDIKTIF DAN INFERENSI STATISTIK

Selamat! Anda kini telah menguasai salah satu perangkat paling esensial dalam perangkat kerja seorang ilmuwan data pemeriksaan regresi. Anda telah mampu mengidentifikasi hubungan antarvariabel, membangun model prediktif, mengevaluasi seberapa baik model tersebut mencocoki data, serta menerapkan pemeriksaan regresi menggunakan Excel. Ini bukan sekadar teknik statistik biasa; regresi merupakan fondasi bagi sebagian besar analitik prediktif dan pembelajaran mesin (*machine learning*) modern. Seiring berjalannya proses belajar Anda, Anda akan menyadari bahwa pengetahuan ini membuat transisi menuju metode pemeriksaan yang lebih lanjut menjadi jauh lebih mudah. Banyak algoritma pembelajaran mesin dibangun di atas gagasan mendasar yang baru saja Anda

pelajari: memanfaatkan pola dalam data masa lalu untuk memprediksi hasil di masa depan. Dengan memahami regresi, Anda telah melangkah maju ke dalam dunia yang menarik tersebut. Namun, perjalanan kita dengan regresi belum berakhir. Di bagian selanjutnya dalam buku ini, kita akan kembali membahas regresi dengan pertanyaan-pertanyaan yang lebih mendalam dan krusial:

- Apakah hubungan yang kita amati signifikan secara statistik, atau mungkinkah hal itu terjadi secara kebetulan?
- Seberapa yakin kita terhadap estimasi kemiringan (*slope*) yang diperoleh?
- Seberapa besar ketidakpastian yang menyertai prediksi kita?
- Bagaimana cara menentukan apakah hubungan regresi tersebut benar-benar ada dalam populasi secara keseluruhan, bukan sekadar dalam sampel yang kita miliki?

Pertanyaan-pertanyaan ini membawa kita melampaui sekadar mendeskripsikan hubungan, menuju pembuatan inferensi statistik. Untuk menjawabnya, kita perlu membangun fondasi yang kuat dalam teori probabilitas, distribusi sampling, dan pengujian hipotesis konsep-konsep yang akan kita bahas lebih lanjut di Bagian 3. Untuk saat ini, luangkan waktu sejenak untuk mengapresiasi sejauh mana Anda telah melangkah. Anda telah menambahkan regresi ke dalam perangkat pemeriksaan Anda, beralih dari sekadar mendeskripsikan data menjadi memodelkan hubungan dan membuat prediksi. Perjalanan ke depan akan memperkenalkan perangkat yang jauh lebih canggih namun dengan fondasi yang telah Anda bangun sejauh ini, Anda sudah sangat siap untuk memanfaatkannya.

BAB 9
PROBABILITAS — BAHASA KETIDAKPASTIAN

9.1 PERAN PROBABILITAS DALAM ANALISIS DATA

Bayangkan Anda adalah seorang prakirawan cuaca yang mencoba memprediksi curah hujan esok hari, seorang analis bisnis yang memperkirakan penjualan kuartal mendatang, atau seorang manajer pengendalian kualitas yang menilai kemungkinan adanya cacat produk dalam proses manufaktur. Dalam setiap situasi tersebut, Anda diminta untuk membuat penilaian mengenai hasil yang belum terjadi. Masa depan penuh dengan ketidakpastian, namun keputusan tetap harus diambil. Dalam dunia sains data dan analitik, ketidakpastian ada di mana-mana. Perilaku pelanggan, pasar keuangan, wabah penyakit, lalu lintas situs web, hingga kerusakan peralatan semuanya melibatkan unsur keacakan. Meskipun kita tidak dapat menghilangkan ketidakpastian, kita dapat belajar untuk mengukurnya, menalarinya, dan mengambil keputusan yang tepat di tengah ketidakpastian tersebut.

Di sinilah probabilitas berperan. Probabilitas menyediakan kerangka kerja matematis untuk mendeskripsikan ketidakpastian. Probabilitas memungkinkan kita untuk mengukur seberapa besar kemungkinan suatu peristiwa terjadi dan menalar hasil acak secara sistematis. Alih-alih mengandalkan intuisi atau tebakan, probabilitas membekali kita dengan perangkat untuk mengevaluasi risiko, menafsirkan data, dan membuat prediksi yang berlandaskan logika. Probabilitas juga merupakan pintu gerbang menuju statistik inferensial. Pada bab-bab sebelumnya, kita berfokus pada statistik deskriptif, metode untuk meringkas dan memvisualisasikan data yang sudah kita miliki. Statistik inferensial melangkah lebih jauh; statistik ini menggunakan probabilitas untuk menarik kesimpulan tentang populasi yang lebih luas berdasarkan sampel yang terbatas. Setiap selang kepercayaan (*confidence interval*), uji hipotesis, dan model prediktif pada dasarnya bersandar pada prinsip-prinsip probabilitas.

Dalam bab ini, Anda akan mempelajari gagasan-gagasan mendasar yang membentuk bahasa ketidakpastian. Kita akan menjelajahi cara kerja probabilitas, mengkaji berbagai jenis peristiwa yang mungkin terjadi, serta mempelajari aturan yang mengatur bagaimana probabilitas digabungkan. Anda juga akan memahami bagaimana probabilitas bersyarat membantu kita memperbarui pemahaman saat informasi baru tersedia, dan bagaimana Teorema Bayes menyediakan kerangka kerja yang ampuh untuk melakukan penalaran di tengah ketidakpastian. Di akhir bab ini, Anda tidak hanya akan memahami prinsip-prinsip inti probabilitas, tetapi juga melihat bagaimana prinsip-prinsip tersebut menjadi landasan bagi teknik-teknik inferensial yang kuat yang akan dibahas dalam bab-bab selanjutnya.

9.2 MEMAHAMI PROBABILITAS: DARI INTUISI MENUJU MATEMATIKA

Apa itu Probabilitas?

Probabilitas adalah ukuran numerik yang menggambarkan seberapa besar kemungkinan suatu peristiwa akan terjadi. Dalam percakapan sehari-hari, kita sering melontarkan pertanyaan seperti "Seberapa besar peluangnya?" atau "Seberapa mungkin hal

itu terjadi?" Probabilitas menyediakan cara matematis yang presisi untuk menjawab pertanyaan-pertanyaan tersebut. Semua nilai probabilitas berada di antara 0 dan 1 (atau setara dengan antara 0% dan 100%):

- Probabilitas = 0 → Peristiwa tersebut mustahil terjadi (tidak mungkin terjadi).
- Probabilitas = 1 → Peristiwa tersebut pasti terjadi.
- Probabilitas antara 0 dan 1 → Peristiwa tersebut mungkin terjadi atau mungkin tidak terjadi.

Perhatikan contoh-contoh berikut:

- Probabilitas mendapatkan angka 7 saat melempar dadu standar bersisi enam adalah 0, karena dadu tersebut hanya memiliki sisi bernomor 1 sampai 6.
- Probabilitas matahari akan terbit esok hari adalah 1, karena hal tersebut hampir pasti terjadi berdasarkan pengetahuan ilmiah dan pengamatan yang berulang.
- Probabilitas munculnya sisi "angka" (*heads*) saat melempar koin yang adil adalah 0,5, yang berarti peluang munculnya sisi "angka" maupun "gambar" (*tails*) adalah sama besar.

Contoh-contoh ini mengilustrasikan bagaimana probabilitas memungkinkan kita untuk menerjemahkan pemahaman intuitif kita mengenai peluang ke dalam bentuk numerik.

Tiga Pendekatan terhadap Probabilitas

Terdapat beberapa cara untuk menentukan probabilitas, bergantung pada informasi yang tersedia. Dalam statistika dan sains data, tiga pendekatan utama sering digunakan.

1. **Probabilitas Klasik (Teoretis)**

   Pendekatan klasik diterapkan ketika semua hasil yang mungkin memiliki peluang yang sama besar untuk terjadi. Dalam situasi seperti itu, probabilitas dapat dihitung menggunakan rumus sederhana:

   $P(E)$ = Jumlah kejadian yang diinginkan / Total jumlah kemungkinan kejadian

   Contoh

   Berapa probabilitas terambilnya kartu As dari satu set standar yang berisi 52 kartu remi?

   - Kejadian yang diinginkan: 4 kartu As
   - Total kemungkinan kejadian: 52 kartu

   $$P(\text{Ace}) = \frac{4}{52} = 0.0769$$

   Ini berarti terdapat peluang sekitar 7,69% untuk mendapatkan kartu As.

   Pendekatan klasik lazim digunakan dalam situasi yang melibatkan permainan peluang, seperti pelemparan dadu, pengambilan kartu, atau pemilihan lotre.

2. Peluang Empiris (Eksperimental)

Dalam banyak situasi nyata, hasil yang mungkin muncul tidak memiliki peluang yang sama besar atau terlalu kompleks untuk diperiksa secara teoretis. Dalam kasus seperti ini, peluang diestimasi menggunakan data hasil pengamatan. Hal ini dikenal sebagai peluang empiris.

Rumusnya adalah:

$P(E)$ = Jumlah kejadian yang muncul / Total jumlah percobaan

Contoh

Misalkan sebuah pabrik memproduksi 10.000 barang, dan pemeriksaan kualitas menunjukkan bahwa 150 barang di antaranya cacat.

$$P(\text{Defective}) = \frac{150}{10000} = 0.015$$

Ini berarti probabilitas dihasilkannya barang yang cacat adalah 1,5%. Probabilitas empiris memegang peranan sentral dalam sains data, di mana probabilitas sering kali diestimasi berdasarkan data historis.

3. Probabilitas Subjektif

Terkadang, data yang dapat diandalkan tidak tersedia, sehingga probabilitas harus diestimasi menggunakan penilaian pakar, pengalaman, atau pendapat yang berdasar. Hal ini dikenal sebagai probabilitas subjektif.

Contohnya meliputi:

- Seorang dokter mengestimasi kemungkinan kesembuhan pasien berdasarkan pengalaman klinis.
- Seorang ekonom memprediksi probabilitas terjadinya resesi.
- Seorang analis olahraga mengestimasi peluang suatu tim untuk memenangkan kejuaraan.

Meskipun probabilitas subjektif tidak diturunkan dari perhitungan formal, jenis probabilitas ini tetap berguna ketika data terbatas atau tidak tersedia.

Terminologi Dasar Probabilitas

Sebelum melangkah lebih jauh, penting untuk memahami beberapa istilah kunci yang digunakan dalam teori probabilitas.

- **Eksperimen**: Suatu tindakan atau proses yang menghasilkan salah satu dari beberapa kemungkinan hasil. Contoh: Melakukan tes diagnostik HIV.
- **Hasil (*Outcome*)**: Satu hasil spesifik dari suatu eksperimen. Contoh: Hasil tes yang menunjukkan positif atau negatif.
- **Ruang Sampel (*Sample Space*)**: Himpunan dari semua kemungkinan hasil suatu eksperimen. Contoh: S = {Positif, Negatif}

- **Kejadian (*Event*)**: Suatu kejadian adalah hasil spesifik atau sekumpulan hasil di dalam ruang sampel. Contoh:
  - Kejadian A: Pasien dinyatakan positif → {Positif}
  - Kejadian B: Pasien dinyatakan negatif → {Negatif}

Memahami konsep-konsep dasar ini akan membantu kita dalam mendeskripsikan dan memeriksa peristiwa yang tidak pasti secara sistematis.

Menerapkan Probabilitas Dasar dalam Excel

Konsep probabilitas menjadi lebih bermakna ketika diterapkan pada data nyata. Mari kita pelajari contoh sederhana menggunakan Excel.

Contoh: Survei Kepuasan Pelanggan

Misalkan sebuah perusahaan melakukan survei terhadap 200 pelanggan mengenai kepuasan mereka terhadap suatu produk. Tanggapan tersebut dirangkum ke dalam kategori-kategori berikut:

Tabel 9.1 Data jumlah survei kepuasan pelanggan

Tingkat Kepuasan	**Hitungan**
Sangat Puas	50
Puas	80
Netral	40
Tidak Puas	20
Sangat Tidak Puas	10

Untuk menghitung probabilitas di Excel:

1. Masukkan tingkat kepuasan di Kolom A.
2. Masukkan jumlah (*frekuensi*) yang sesuai di Kolom B.
3. Hitung total jumlah respons di sel B7 menggunakan rumus: =SUM(B2:B6)
4. Hitung probabilitas untuk kategori 'Sangat Puas' (*Very Satisfied*) di sel C2 menggunakan rumus: =B2/B7
5. Salin rumus tersebut ke baris-baris di bawahnya.
6. Format Kolom C sebagai Persentase agar lebih mudah diinterpretasikan.
1. Pilih sel-sel dalam kolom tersebut, lalu klik simbol persentase pada grup menu angka (number).

Hasilnya:

Tabel 9.2 Proporsi survei kepuasan pelanggan

Tingkat Kepuasan	**Hitungan**	**Probabilitas**
Sangat Puas	50	25%
Puas	80	40%
Netral	40	20%
Tidak Puas	20	10%
Sangat Tidak Puas	10	5%
Total	200	

Ini berarti 25% pelanggan menyatakan sangat puas. Penggunaan Excel dengan cara ini memungkinkan analis untuk mengubah jumlah frekuensi menjadi estimasi probabilitas dengan cepat, yang kemudian dapat digunakan untuk pemeriksaan statistik yang lebih mendalam.

9.3 JENIS-JENIS KEJADIAN DALAM PROBABILITAS

Dalam teori probabilitas, kejadian (*event*) merujuk pada sekumpulan hasil dari suatu eksperimen yang ingin kita pemeriksaan. Kejadian dapat berbentuk beragam: ada yang terdiri dari satu hasil tunggal, sementara yang lain melibatkan beberapa hasil. Beberapa kejadian dapat terjadi secara bersamaan, sedangkan yang lain tidak. Dalam situasi tertentu, terjadinya satu kejadian dapat memengaruhi probabilitas kejadian lainnya. Memahami berbagai jenis kejadian sangatlah penting karena hal ini menentukan aturan probabilitas mana yang harus digunakan saat menghitung kemungkinan dan memeriksa data.

Kejadian Sederhana (*Simple Events*)

Kejadian sederhana terdiri dari tepat satu hasil dalam ruang sampel. Contoh:

- Mendapatkan angka 4 saat melempar dadu bersisi enam
- Seorang pelanggan melakukan pembelian pada kunjungan pertama mereka ke situs web

Karena kejadian sederhana hanya melibatkan satu hasil, probabilitasnya sering kali dapat dihitung secara langsung.

Kejadian Majemuk (*Compound Events*)

Kejadian majemuk terdiri dari dua hasil atau lebih dari ruang sampel. Contoh:

- Mendapatkan angka genap saat melempar dadu → {2, 4, 6}
- Pengunjung situs web melakukan pembelian atau mendaftar untuk menerima buletin (*newsletter*)

Kejadian majemuk memungkinkan kita untuk mempelajari situasi yang lebih luas yang melibatkan berbagai kemungkinan hasil.

Kejadian Saling Lepas (*Mutually Exclusive Events*)

Dua kejadian dikatakan saling lepas jika keduanya tidak dapat terjadi pada saat yang bersamaan. Contoh:

- Sebuah koin tidak mungkin menunjukkan sisi angka dan sisi gambar secara bersamaan.
- Seorang pelanggan tidak mungkin membeli sekaligus tidak membeli barang yang sama dalam satu transaksi.

Ketika dua kejadian bersifat saling lepas, terjadinya satu kejadian secara otomatis mencegah kejadian lainnya untuk terjadi.

Kejadian Bebas (*Independent Events*)

Dua kejadian dikatakan bebas (independen) jika terjadinya salah satu kejadian tidak memengaruhi probabilitas kejadian lainnya. Contoh:

- Melempar koin dan melempar dadu — hasil dari lemparan koin tidak memengaruhi hasil dari lemparan dadu.

- Melempar koin sebanyak dua kali — hasil lemparan pertama tidak memengaruhi hasil lemparan kedua.
- Pengambilan sampel acak dengan pengembalian (*sampling with replacement*), di mana setiap observasi diambil secara independen dari populasi yang sama.

Independensi merupakan konsep krusial dalam probabilitas dan statistik karena banyak model statistik mengasumsikan bahwa observasi bersifat independen (bebas).

Peristiwa Dependen (Saling Bergantung)

Peristiwa dikatakan dependen jika hasil dari suatu peristiwa mengubah probabilitas peristiwa lainnya. Contoh:

- Mengambil dua kartu dari setumpuk kartu tanpa pengembalian. Kartu pertama yang diambil memengaruhi probabilitas untuk pengambilan kedua.
- Perilaku pembelian pelanggan, di mana pembelian suatu produk dapat meningkatkan kemungkinan pembelian produk lain yang berkaitan.

Mengenali dependensi sangatlah penting karena perhitungan probabilitas harus memperhitungkan hubungan antarperistiwa tersebut.

Peristiwa Komplementer (Saling Melengkapi)

Dua peristiwa dikatakan komplementer jika keduanya mencakup seluruh kemungkinan hasil, dan terjadinya salah satu peristiwa secara otomatis berarti peristiwa lainnya tidak terjadi. Contoh:

- Melempar koin:
  - Peristiwa A: Muncul sisi Angka (*Heads*)
  - Peristiwa A^c: Tidak muncul sisi Angka (yaitu, muncul sisi Gambar/Tails)

Jumlah probabilitas dari peristiwa-peristiwa komplementer selalu sama dengan 1.

Contoh: Peristiwa Independen vs. Dependen

Dalam sains data dan analitik bisnis, peristiwa jarang diperiksa secara terpisah. Memahami apakah peristiwa dapat terjadi secara bersamaan atau apakah probabilitas suatu peristiwa bergantung pada peristiwa lain sangatlah penting untuk menentukan rumus matematika yang tepat. Sebagai contoh, dalam pemeriksaan pemasaran, kita mungkin tertarik pada urutan peristiwa, seperti perjalanan pelanggan (*customer journey*) melalui corong pemasaran (*marketing funnel*). Bayangkan seorang analis pemasaran yang mempelajari bagaimana pengguna berinteraksi dengan kampanye surel (*email campaign*). Misalkan diamati probabilitas berikut:

- Probabilitas seorang pengguna membuka surel
- Probabilitas seorang pengguna mengeklik tautan, dengan asumsi surel tersebut telah dibuka

Sebelum kita dapat menghitung peluang seorang pengguna melakukan keduanya (membuka dan mengeklik), kita harus bertanya:

Apakah satu peristiwa memengaruhi peristiwa lainnya?

Jika keduanya bersifat independen (bebas): Ini berarti mengeklik tautan tidak ada hubungannya dengan membuka surel. Kita cukup mengalikan kedua peluang tersebut:

$$P(Membuka\ dan\ Mengeklik) = P(Membuka) \times P(Mengeklik)$$

Namun, keduanya bersifat dependen (bergantung): Pada kenyataannya, pengguna harus membuka surel terlebih dahulu sebelum dapat mengeklik tautan di dalamnya. Tindakan "mengeklik" bergantung pada tindakan "membuka". Ketergantungan ini mengubah perhitungan matematisnya:

$$P(Membuka\ dan\ Mengeklik) = P(Membuka) \times P(Mengeklik \mid Membuka)$$

Menentukan jenis peristiwa bukan sekadar teori; hal ini sangat penting untuk pemeriksaan yang akurat. Kesalahan dalam memahami hubungan antarperistiwa dapat menyebabkan penerapan aturan probabilitas yang keliru, sehingga menghasilkan nilai probabilitas akhir yang tidak tepat. Pelajari bagaimana perbedaan ini memengaruhi perhitungan pada bagian-bagian selanjutnya.

9.4 ATURAN PROBABILITAS

Untuk bekerja secara efektif dengan probabilitas, kita perlu memahami beberapa aturan dasar. Aturan-aturan ini menyediakan kerangka kerja matematis yang konsisten untuk menggabungkan berbagai peristiwa dan menghitung kemungkinan terjadinya hasil yang lebih kompleks. Dengan menguasai aturan-aturan ini, seorang analis dapat melampaui pertanyaan probabilitas sederhana dan mulai memecahkan masalah yang sering ditemui dalam bidang sains data, pemeriksaan risiko, serta pengambilan keputusan di tengah ketidakpastian.

1. **Aturan Rentang Probabilitas**

   Setiap nilai probabilitas harus berada dalam rentang 0 hingga 1 (inklusif).

   $$0 \leq P(A) \leq 1$$

   Aturan ini mencerminkan makna dasar dari probabilitas:

   - $P(A) = 0 \rightarrow$ Peristiwa A mustahil terjadi.
   - $P(A) = 1 \rightarrow$ Peristiwa A pasti terjadi.
   - $0 < P(A) < 1 \rightarrow$ Peristiwa A mungkin terjadi atau mungkin tidak terjadi.

   Contoh

   Bayangkan melempar dadu standar bersisi enam. Peluang munculnya angka kurang dari 7 adalah:

   $$P(angka < 7) = 1$$

   karena setiap kemungkinan hasil (angka 1 sampai 6) memenuhi kondisi tersebut.

2. Aturan Komplemen

Komplemen dari suatu kejadian merepresentasikan situasi di mana kejadian tersebut tidak terjadi. Probabilitas dari komplemen tersebut adalah:

$$P(A^c) = 1 - P(A)$$

Di mana:

- A = kejadian tersebut
- A^c = komplemen dari kejadian A

Contoh

Jika probabilitas hujan hari ini adalah:

$$P(Hujan) = 0{,}30$$

maka probabilitas tidak turun hujan adalah:

$$P(Tidak\ Hujan) = 1 - 0{,}30 = 0{,}70$$

Aturan ini sangat berguna karena terkadang lebih mudah untuk menghitung probabilitas bahwa suatu kejadian tidak terjadi.

3. Aturan Penjumlahan

Aturan penjumlahan memungkinkan kita untuk menghitung probabilitas bahwa setidaknya satu dari dua kejadian akan terjadi.

Kasus 1: Kejadian yang Saling Lepas (*Mutually Exclusive*)

Jika dua kejadian tidak dapat terjadi pada saat yang bersamaan, maka kedua kejadian tersebut dikatakan saling lepas (*mutually exclusive*). Probabilitasnya adalah sebagai berikut:

$$P(A \text{ or } B) = P(A) + P(B)$$

Contoh

Berapakah peluang munculnya angka 2 atau 5 saat melempar sebuah dadu yang adil?

$$P(2) = \frac{1}{6}, \quad P(5) = \frac{1}{6}$$

$$P(2 \text{ atau } 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Kasus 2: Kejadian yang Tidak Saling Lepas (*Non-Mutually Exclusive Events*)

Jika dua kejadian dapat terjadi secara bersamaan, kita harus mengurangi bagian yang tumpang tindih agar tidak terhitung dua kali.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Jika:

$$P(A) = 0.55$$
$$P(B) = 0.45$$
$$P(A \cap B) = 0.20$$

Maka

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \approx 0.80$$

Contoh

Misalkan kita melempar sebuah dadu dan mempertimbangkan dua kejadian:

- Kejadian A: muncul angka genap → {2, 4, 6}
- Kejadian B: muncul angka lebih besar dari 3 → {4, 5, 6}

Beberapa hasil (4 dan 6) termasuk dalam kedua kejadian tersebut $P(A \cap B)$.

$$P(A) = \frac{3}{6}, \quad P(B) = \frac{3}{6}, \quad P(A \cap B) = \frac{2}{6}$$

$$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

4. Aturan Perkalian

Aturan perkalian digunakan untuk menghitung probabilitas bahwa dua peristiwa terjadi secara bersamaan.

Kasus 1: Peristiwa Independen (Bebas)

Jika peristiwa-peristiwa tersebut tidak saling memengaruhi, kita mengalikan probabilitas masing-masing peristiwa tersebut.

$$P(A \cap B) = P(A) \times P(B)$$

Jika:

$$P(A) = 0.60$$
$$P(B) = 0.45$$

Kemudian

$$P(A \cap B) = 0.60 \times 0.45 = 0.27$$

Contoh

Misalkan kita melempar koin dan mengundi dadu.

- $P(\text{Heads}) = 0.5$
- $P(\text{Rolling a } 4) = \frac{1}{6}$

$$P(\text{Heads and } 4) = 0.5 \times \frac{1}{6} = \frac{1}{12}$$

Kasus 2: Kejadian Bergantung

Ketika kejadian-kejadian saling memengaruhi, kita harus menggunakan peluang bersyarat.

$$P(A \cap B) = P(A) \times P(B \mid A)$$

Jika:

$$P(A) = 0.60$$
$$P(B \mid A) = 0.35$$

Maka:

$$P(A \cap B) = P(A) \cdot P(B \mid A) \approx 0.21$$

Contoh

Misalkan kita mengambil dua kartu dari setumpuk kartu tanpa pengembalian. Jika kartu pertama adalah As, peluang untuk mendapatkan kartu As lainnya akan berubah karena satu kartu As telah dikeluarkan dari tumpukan tersebut. Ketergantungan ini harus diperhitungkan dalam perhitungan peluang. Pelajari lebih lanjut mengenai peluang bersyarat pada bagian berikutnya.

5. Hukum Peluang Total

Ketika beberapa kejadian yang saling lepas (*mutually exclusive*) mencakup seluruh kemungkinan hasil, probabilitas dari keseluruhan ruang sampel sama dengan jumlah dari probabilitas masing-masing kejadian tersebut.

$$P(\boldsymbol{A_1}) + P(\boldsymbol{A_2}) + \cdots + P(\boldsymbol{A_n}) = 1$$

Contoh

Saat melempar dadu bersisi enam:

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$$

karena salah satu dari hasil tersebut pasti akan terjadi.

Mengapa Aturan-Aturan Ini Penting

Aturan-aturan probabilitas menyediakan kerangka logis untuk menalar ketidakpastian. Aturan ini memastikan bahwa perhitungan probabilitas tetap konsisten dan secara matematis dapat dipertanggungjawabkan. Dalam sains data, aturan-aturan ini digunakan secara terus-menerus. Aturan ini membantu analis untuk:

- menggabungkan probabilitas dari berbagai kejadian
- mengevaluasi risiko dan hasil
- memodelkan sistem yang tidak pasti
- membangun model prediktif

Setiap metode statistik yang diperkenalkan selanjutnya dalam buku ini mulai dari distribusi sampling hingga pengujian hipotesis pada dasarnya bergantung pada prinsip-prinsip probabilitas mendasar ini. Dengan menguasai aturan-aturan ini, Anda mulai mengembangkan kefasihan dalam apa yang kita sebut sebagai bahasa ketidakpastian.

9.5 PROBABILITAS BERSYARAT

Memahami Konsep "Dengan Syarat"

Probabilitas bersyarat menjelaskan peluang terjadinya suatu peristiwa dengan asumsi bahwa peristiwa lain telah terjadi. Dengan kata lain, konsep ini memungkinkan kita untuk memperbarui estimasi probabilitas saat informasi baru tersedia. Alih-alih bertanya, "Berapa peluang terjadinya Peristiwa B?", probabilitas bersyarat mengajukan pertanyaan: "Berapa peluang terjadinya Peristiwa B, dengan syarat Peristiwa A telah terjadi?"
Hal ini dituliskan sebagai:

$$P(B \mid A)$$

yang dibaca sebagai "probabilitas B dengan syarat A."

Interpretasi dalam Dunia Nyata

Probabilitas bersyarat sering muncul dalam pengambilan keputusan dan pemeriksaan data di dunia nyata:

- P(Hujan | Berawan) – Berapa peluang terjadinya hujan dengan syarat langit dalam keadaan berawan?
- P(Pembelian | Mengunjungi Situs Web) – Berapa peluang seorang pengunjung melakukan pembelian dengan syarat ia telah mengunjungi situs web tersebut?
- P(Spam | Berisi kata "Gratis") – Berapa peluang sebuah surel (email) merupakan spam dengan syarat surel tersebut mengandung kata "gratis"?

Dalam setiap kasus, kita menggunakan informasi yang sudah diketahui untuk menyempurnakan prediksi kita.

Rumus Probabilitas Bersyarat

Definisi formal dari probabilitas bersyarat adalah:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Rumus ini menyatakan bahwa probabilitas B dengan syarat A sama dengan probabilitas terjadinya A dan B secara bersamaan, dibagi dengan probabilitas terjadinya A.

Penjelasan Intuitif

Bayangkan probabilitas bersyarat sebagai tindakan mempersempit fokus ke dalam lingkup hasil yang lebih kecil. Alih-alih melihat semua kemungkinan hasil, kita kini hanya berfokus pada kasus-kasus di mana Kejadian A terjadi. Di dalam kelompok yang lebih kecil tersebut, kita bertanya:

Berapa proporsi kasus yang juga mencakup Kejadian B?

Contoh: Segmentasi Pelanggan

Bayangkan sebuah situs web *e-commerce* dengan 1.000 pengunjung:

- 200 adalah pelanggan yang kembali (Kejadian A)
- 150 pengunjung melakukan pembelian (Kejadian B)
- 75 adalah pelanggan yang kembali dan melakukan pembelian (A ∩ B)

Kita ingin menghitung:

$$P(Pembelian \mid Pelanggan\ yang\ Kembali)$$

Dengan menggunakan rumus probabilitas bersyarat:

$$P(\text{Purchase} \mid \text{Returning}) = \frac{P(\text{Purchase} \cap \text{Returning})}{P(\text{Returning})}$$

$$= \frac{75/1000}{200/1000}$$
$$= 0.375$$

Interpretasi:

Sekitar 37,5% pelanggan yang kembali melakukan pembelian; angka ini jauh lebih tinggi dibandingkan tingkat pembelian secara keseluruhan yang sebesar 15% (150 dari 1.000 pengunjung). Wawasan ini dapat membantu bisnis untuk memfokuskan upaya pemasaran pada pelanggan yang kembali.

Latihan Praktik Excel

Mari kita hitung probabilitas bersyarat menggunakan Excel.

Gunakan dataset `*email_recipients*` (1.000 penerima email) dan lacak dua kejadian:

- Apakah email dibuka
- Apakah tautan di dalam email diklik

Struktur Dataset

Kolom	Deskripsi
A	ID Pelanggan

B Email Dibuka (*TRUE*/*FALSE*)
C Tautan Diklik (*TRUE*/*FALSE*)
Kita ingin menghitung:

$$P(Klik \mid Buka)$$

Rumus Excel

$$= COUNTIFS(B2{:}B1001, TRUE, C2{:}C1001, TRUE) \,/\, COUNTIF(B2{:}B1001, TRUE)$$

Penjelasan

- `COUNTIFS` menghitung jumlah penerima yang membuka email dan mengeklik tautan
- `COUNTIF` menghitung jumlah penerima yang membuka email
- Hasil pembagian keduanya memberikan probabilitas bersyarat terjadinya klik, dengan asumsi email telah dibuka

Perhitungan semacam ini lazim digunakan dalam analitik pemasaran, pemeriksaan perilaku pengguna, dan optimalisasi tingkat konversi (*conversion rate*).

Menguji Independensi / Dependensi

Probabilitas bersyarat juga membantu menentukan apakah dua kejadian bersifat independen (bebas) atau dependen (bergantung). Dua kejadian A dan B dikatakan independen jika mengetahui bahwa A telah terjadi tidak mengubah probabilitas terjadinya B.
Ini berarti:

$$P(B \mid A) = P(B)$$

Interpretasi

- Kejadian Independen:
  Terjadinya satu kejadian tidak memengaruhi probabilitas kejadian lainnya.
- Kejadian Dependen:
  Terjadinya satu kejadian mengubah probabilitas kejadian lainnya.

Contoh:

Skenario	**Hubungan**
Melempar dua koin	Independen
Mengambil kartu tanpa pengembalian	Dependen
Membuka surel dan mengeklik tautan	Dependen

Menguji Independensi di Excel

Untuk menguji apakah dua peristiwa bersifat independen:

1. Hitung $P(B)$ (probabilitas keseluruhan dari B).
2. Hitung $P(B \mid A)$ (probabilitas bersyarat).

Jika kedua probabilitas tersebut sama atau sangat mendekati, maka peristiwa-peristiwa tersebut kemungkinan besar independen. Jika terdapat perbedaan yang signifikan, maka peristiwa-peristiwa tersebut bersifat dependen (saling bergantung). Memahami perbedaan ini sangat penting dalam sains data, karena banyak model termasuk inferensi Bayesian, algoritma klasifikasi, dan pemeriksaan kausal bergantung pada penalaran mengenai probabilitas bersyarat dan independensi.

9.6 TEOREMA BAYES: PEMBARUAN KEYAKINAN DAN ANALISIS PROBABILITAS BERSYARAT

Memperbarui Keyakinan dengan Informasi Baru

Teorema Bayes merupakan salah satu konsep paling ampuh dalam statistika dan sains data modern. Teorema ini menyediakan cara sistematis untuk memperbarui probabilitas ketika bukti baru tersedia. Secara sederhana, Teorema Bayes membantu menjawab pertanyaan seperti:

> *Jika kita mengamati bukti tertentu, bagaimana kita harus merevisi keyakinan kita mengenai apa yang menyebabkannya?*

Konsep ini bersifat fundamental dalam berbagai penerapan di dunia nyata, antara lain:

- **Deteksi spam** – Memperbarui probabilitas bahwa suatu surel (email) adalah spam berdasarkan isinya.
- **Diagnosis medis** – Mengestimasi probabilitas penyakit berdasarkan hasil tes.
- **Pembelajaran mesin (*machine learning*)** – Banyak algoritma menerapkan penalaran Bayesian untuk membuat prediksi.
- **Pengujian A/B (A/B *testing*)** – Memperbarui tingkat keyakinan terhadap hasil eksperimen seiring dengan masuknya data baru.

Oleh karena itu, penalaran Bayesian bukan sekadar teori—konsep ini digunakan secara luas dalam pengambilan keputusan berbasis data di era modern.

Rumus Teorema Bayes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Istilah	Makna
P(A|B)	Posterior – hal yang ingin kita cari
P(B|A)	Likelihood (Kemungkinan) – Probabilitas terjadinya B jika A benar.
P(A)	Probabilitas prior – keyakinan awal kita mengenai A
P(B)	Bukti (atau probabilitas marginal) – total probabilitas teramatinya B

Pada intinya, teorema ini menggabungkan pengetahuan awal dengan bukti baru untuk menghasilkan keyakinan yang telah diperbarui.

Contoh yang Relevan: Paradoks dalam Tes Medis

Mari kita lihat sebuah skenario yang sering membingungkan bahkan bagi para profesional berpengalaman sekalipun.

Situasi:

Bayangkan sebuah tes diagnostik untuk penyakit langka yang menyerang 1% populasi. Tes tersebut memiliki karakteristik sebagai berikut:

- **Tingkat positif sejati (*true positive rate*): 95%**
  (Jika seseorang mengidap penyakit tersebut, tes akan mendeteksinya dalam 95% kasus.)
- **Tingkat positif palsu (*false positive rate*): 5%**
  (Orang sehat akan mendapatkan hasil tes positif dalam 5% kasus.)

Pertanyaan

Jika seseorang mendapatkan hasil tes positif, berapakah probabilitas ia benar-benar mengidap penyakit tersebut?

Intuisi Umum (Biasanya Keliru)

Banyak orang berasumsi: "Tes ini memiliki akurasi 95%, jadi pasti ada peluang 95% bahwa orang tersebut mengidap penyakit itu." Namun, anggapan ini mengabaikan fakta seberapa langka penyakit tersebut.

Proses Perhitungan:

Langkah 1: Menentukan Kejadian

Misalkan:

- D = Seseorang mengidap penyakit tersebut
- T^+ = Hasil tes positif

Informasi yang diketahui:

- $P(D)$ = 0,01
- $P(T^+ \mid D)$ = 0,95
- $P(T^+ \mid \text{bukan } D)$ = 0,05

Langkah 2: Menghitung Probabilitas Hasil Tes Positif

Hasil tes positif dapat terjadi melalui dua cara:

1. Seseorang mengidap penyakit tersebut dan mendapatkan hasil tes positif
2. Seseorang tidak mengidap penyakit tersebut tetapi mendapatkan hasil positif palsu

$$P(T+) = P(T+ \mid D)P(D) + P(T+ \mid \text{bukan } D)P(\text{bukan } D)$$
$$P(T+) = (0{,}95 \times 0{,}01) + (0{,}05 \times 0{,}99)$$
$$P(T+) = 0{,}0095 + 0{,}0495 = 0{,}059$$

Jadi, sekitar 5,9% dari seluruh tes akan menunjukkan hasil positif.

Langkah 3: Menerapkan Teorema Bayes

$$P(D \mid T^+) = \frac{P(T^+ \mid D)P(D)}{P(T^+)}$$
$$P(D \mid T^+) = \frac{0.95 \times 0.01}{0.059}$$
$$P(D \mid T^+) \approx 0.161$$

Hasil yang Mengejutkan

Bahkan setelah hasil tes positif, probabilitas bahwa orang tersebut benar-benar mengidap penyakit tersebut hanya sekitar 16,1%. Hal ini terjadi karena penyakit tersebut sangat langka, artinya banyak hasil positif sebenarnya adalah positif palsu. Fenomena ini menggambarkan pentingnya tingkat dasar (probabilitas awal) dalam penalaran statistik.

Intuisi di Balik Teorema Bayes

Teorema Bayes memungkinkan kita untuk membalikkan probabilitas bersyarat. Alih-alih bertanya:

> *"Jika seseorang mengidap penyakit tersebut, seberapa besar kemungkinan hasil tes positif?"*

Teorema Bayes membantu menjawab pertanyaan yang lebih bermanfaat:

> *"Jika kita mengamati hasil tes positif, seberapa besar kemungkinan orang tersebut mengidap penyakit tersebut?"*

Pembalikan ini sangat penting dalam banyak aplikasi ilmu data.

Aplikasi Dunia Nyata

Penalaran Bayes banyak digunakan dalam analitik modern:

Tabel 9.3 Aplikasi Teorema Bayes

Penerapan	**Pertanyaan Bayesian**
Diagnosis medis	Berdasarkan gejala yang ada, penyakit apa yang paling mungkin terjadi?
Penyaringan spam	Berdasarkan kata-kata tertentu, seberapa besar kemungkinan suatu surel merupakan spam?
Pengendalian kualitas	Berdasarkan cacat yang ditemukan, mesin mana yang kemungkinan besar menghasilkannya?
Riset pasar	Berdasarkan perilaku pembelian, segmen pelanggan mana yang paling mungkin?

Dalam semua kasus ini, kita berangkat dari bukti yang teramati, lalu menelusuri ke belakang untuk menyimpulkan penyebab yang paling mungkin.

9.7 RINGKASAN PROBABILITAS

Dalam bab ini, kita telah membahas probabilitas, yaitu bahasa matematika yang digunakan untuk mendeskripsikan ketidakpastian. Dalam sains data, probabilitas menjadi landasan untuk menalar keacakan, mengevaluasi risiko, dan membuat prediksi yang tepat berdasarkan data. Kita memulainya dengan mengeksplorasi eksperimen, hasil (*outcomes*), dan ruang sampel, yang mendefinisikan kemungkinan hasil dari suatu proses acak. Sebuah kejadian (*event*) merepresentasikan satu atau lebih hasil yang menjadi perhatian di dalam ruang sampel tersebut. Memahami bagaimana suatu kejadian didefinisikan memungkinkan kita untuk menerjemahkan situasi dunia nyata seperti perilaku pelanggan, kegagalan sistem, atau respons pemasaran ke dalam pertanyaan probabilitas yang dapat diperiksa secara matematis.

Selanjutnya, kita meninjau berbagai jenis kejadian, termasuk kejadian sederhana, kejadian majemuk, kejadian yang saling lepas (*mutually exclusive*), kejadian independen, dan kejadian dependen. Memahami perbedaan ini sangat penting karena jenis kejadian akan menentukan aturan probabilitas mana yang harus diterapkan saat melakukan perhitungan. Kemudian, kita membahas aturan-aturan dasar probabilitas yang menyediakan kerangka kerja terstruktur untuk menangani hasil yang tidak pasti. Aturan-aturan ini mencakup aturan rentang probabilitas, aturan komplemen, aturan penjumlahan untuk menggabungkan kejadian, serta aturan perkalian untuk probabilitas gabungan. Aturan-aturan tersebut memastikan bahwa perhitungan probabilitas tetap konsisten secara logis dan valid secara matematis. Konsep kunci dalam pemeriksaan data modern adalah probabilitas bersyarat (*conditional probability*), yang mengukur peluang terjadinya suatu kejadian dengan asumsi bahwa kejadian lain telah terjadi. Probabilitas bersyarat mencerminkan bagaimana informasi baru mengubah pemahaman kita tentang ketidakpastian. Konsep ini sangat penting dalam berbagai penerapan dunia nyata, seperti memprediksi perilaku pelanggan, mengevaluasi kinerja pemasaran, dan memeriksa keterlibatan pengguna.

Berdasarkan konsep probabilitas bersyarat, kita memperkenalkan Teorema Bayes, sebuah metode yang ampuh untuk memperbarui probabilitas ketika terdapat bukti baru. Penalaran Bayesian memainkan peran krusial dalam bidang-bidang seperti diagnosis medis, deteksi spam, pembelajaran mesin (*machine learning*), dan sains pengambilan keputusan. Dengan menggabungkan pengetahuan awal (*prior knowledge*) dengan data yang teramati, Teorema Bayes memungkinkan kita untuk menyempurnakan prediksi dan menarik kesimpulan yang lebih tepat. Sepanjang bab ini, kita menekankan pada contoh-contoh praktis yang diambil dari konteks sains data, termasuk analitik situs web, kampanye pemasaran, dan pemeriksaan keterlibatan melalui surel (email).

Contoh-contoh ini mengilustrasikan bagaimana konsep-konsep probabilitas dapat diterapkan secara langsung dalam tugas-tugas pemeriksaan yang nyata. Probabilitas menjadi jembatan antara statistik deskriptif dan statistik inferensial. Jika statistik deskriptif merangkum pola dalam data yang diamati, maka probabilitas menyediakan kerangka teoretis yang diperlukan untuk memahami bagaimana data berperilaku dalam kondisi ketidakpastian. Pada bab berikutnya, kita akan mengembangkan gagasan-gagasan ini dengan mengeksplorasi distribusi probabilitas, yang menggambarkan bagaimana variabel acak berperilaku di berbagai

kemungkinan hasil. Distribusi probabilitas memungkinkan kita untuk memodelkan fenomena dunia nyata seperti pembelian oleh pelanggan, kesalahan pengukuran, dan variasi alami serta menjadi landasan bagi banyak metode statistik yang digunakan dalam sains data.

BAB 10
DISTRIBUSI PROBABILITAS

10.1 PENGANTAR DISTRIBUSI PROBABILITAS

Pada Bab 9, kita telah mempelajari cara menghitung probabilitas dari peristiwa-peristiwa tertentu. Contohnya, munculnya sisi angka pada lemparan koin, ditemukannya produk yang cacat, atau pelanggan lama yang melakukan pembelian kembali. Perhitungan semacam ini membantu kita menjawab pertanyaan mengenai hasil (*outcome*) individual. Namun, dalam banyak situasi nyata, kita tertarik pada sesuatu yang lebih luas. Alih-alih bertanya:

"Berapa probabilitas terjualnya tepat 5 Unit hari ini?"

kita mungkin ingin memahami keseluruhan pola dari berbagai kemungkinan hasil. Sebagai contoh, berapa probabilitas terjualnya 0, 1, 2, 3, 4, 5, atau lebih unit pada hari tertentu? Untuk menjawab pertanyaan seperti ini, kita menggunakan distribusi probabilitas. Distribusi probabilitas adalah deskripsi matematis mengenai bagaimana probabilitas ditetapkan pada nilai-nilai yang mungkin dari suatu variabel acak. Alih-alih berfokus pada satu hasil saja, distribusi probabilitas memberikan gambaran lengkap mengenai seberapa besar kemungkinan terjadinya berbagai hasil yang berbeda. Anda dapat membayangkan distribusi probabilitas sebagai peta ketidakpastian. Alih-alih hanya mengetahui satu nilai probabilitas, kita dapat melihat keseluruhan lanskap dari hasil-hasil yang mungkin terjadi beserta tingkat kemungkinan untuk setiap hasil tersebut. Distribusi probabilitas memegang peranan sentral dalam statistika dan sains data. Distribusi ini memungkinkan kita untuk:

- Mendeskripsikan pola dalam fenomena acak
- Menghitung probabilitas untuk rentang hasil tertentu
- Membuat prediksi mengenai observasi di masa depan
- Memahami perilaku statistik sampel

Yang terpenting, distribusi probabilitas menjadi jembatan antara teori probabilitas dan statistika inferensial. Jika teori probabilitas membantu kita bernalar mengenai ketidakpastian, maka distribusi probabilitas memungkinkan kita untuk memodelkan perilaku data acak, yang merupakan hal esensial bagi inferensi statistik. Dalam bab ini, kita akan mengeksplorasi distribusi-distribusi probabilitas terpenting yang digunakan dalam statistika dan sains data. Kita juga akan mempelajari cara menggunakan distribusi-distribusi tersebut di Excel serta melihat penerapannya dalam berbagai masalah nyata di bidang bisnis, sains, dan pemeriksaan. Kita akan memulainya dengan memperkenalkan konsep variabel acak, yang merupakan landasan dari distribusi probabilitas.

10.2 DASAR-DASAR VARIABEL ACAK

Memahami Variabel

Sebelum membahas distribusi probabilitas, ada baiknya kita meninjau kembali konsep mengenai variabel. Dalam statistika dan sains data, variabel adalah suatu karakteristik atau sifat yang dapat memiliki nilai yang berbeda-beda. Variabel merepresentasikan aspek-aspek di dunia yang dapat diukur atau diamati, yang kemudian kita pemeriksaan menggunakan data. Sebagai contoh:

- Dalam studi kesehatan, tekanan darah merupakan variabel yang berbeda antara satu pasien dengan pasien lainnya.
- Dalam analitik ritel, pendapatan penjualan harian adalah variabel yang berubah dari hari ke hari.
- Dalam studi pendidikan, nilai ujian merupakan variabel yang bervariasi di antara para mahasiswa.

Variabel secara umum diklasifikasikan ke dalam dua kategori besar.

1. **Variabel Kualitatif (Kategorikal)**

   Variabel kualitatif merepresentasikan kategori atau label kelompok, bukan pengukuran numerik. Dua jenis yang umum adalah:

   Variabel Nominal

   Variabel ini merepresentasikan kategori tanpa urutan atau peringkat alami. Contohnya meliputi:

   - Jenis buah (apel, jeruk, pisang)
   - Golongan darah (A, B, AB, O)
   - Negara tempat tinggal

   Variabel Ordinal

   Variabel ini merepresentasikan kategori yang memiliki urutan bermakna, namun perbedaan antar kategori tidak selalu sama. Contohnya meliputi:

   - Ukuran kaus (Kecil, Sedang, Besar)
   - Tingkat pendidikan (SMA, Sarjana, Magister, PhD)
   - Peringkat kepuasan (Buruk, Cukup, Baik, Sangat Baik)

2. **Variabel Kuantitatif (Numerik)**

   Variabel kuantitatif merepresentasikan pengukuran atau hitungan numerik. Variabel ini terbagi menjadi dua jenis utama.

   Variabel Diskrit

   Variabel diskrit memiliki nilai numerik yang dapat dihitung, biasanya berupa bilangan bulat. Contohnya meliputi:

   - Jumlah pelanggan yang masuk ke toko
   - Jumlah produk cacat dalam satu kelompok produksi
   - Jumlah surel yang diterima dalam satu jam

   Variabel Kontinu

   Variabel kontinu merepresentasikan kuantitas terukur yang dapat memiliki nilai yang tak terhingga banyaknya dalam suatu rentang tertentu.

Contohnya meliputi:

- Berat badan pasien
- Waktu yang diperlukan untuk menyelesaikan suatu tugas
- Suhu di dalam ruangan

Sejauh ini, kita telah membahas variabel terutama dalam konteks pengumpulan dan deskripsi data. Namun, saat mempelajari probabilitas, kita akan mengenal konsep yang berkaitan tetapi lebih spesifik, yaitu variabel acak (*random variable*).

Memahami Variabel Acak

Variabel acak adalah nilai numerik yang merepresentasikan hasil dari suatu eksperimen acak. Lebih tepatnya, variabel acak menetapkan suatu angka untuk setiap kemungkinan hasil dari eksperimen yang hasilnya tidak pasti.

- Dalam eksperimen acak, hasil akhirnya tidak dapat diketahui sebelumnya.
- Variabel acak menerjemahkan hasil yang tidak pasti tersebut ke dalam angka-angka yang dapat diperiksa secara matematis.

Dengan cara ini, variabel acak menjadi jembatan antara probabilitas dan statistika. Variabel acak memungkinkan kita menggunakan perangkat matematika untuk memodelkan, memeriksa, dan memahami proses-proses acak.

Contoh Variabel Acak

Variabel Acak Diskret

Misalkan sebuah rumah sakit mencatat jumlah pasien yang dinyatakan positif COVID-19 setiap harinya. Misalkan:

X = jumlah pasien yang dinyatakan positif COVID-19 di Rumah Sakit A pada hari tertentu. Kemungkinan nilai X bisa berupa:

0, 1, 2, 3, …

Karena hasil-hasil ini dapat dihitung (dicacah), maka X merupakan variabel acak diskret.

Variabel Acak Kontinu

Pertimbangkan sebuah kampanye donor darah di universitas yang bertajuk “Donor Darah untuk Menyelamatkan Nyawa.” Sebelum mendonorkan darah, mahamahasiswa harus memenuhi kriteria kelayakan tertentu, termasuk persyaratan berat badan minimum.

Misalkan:

Y = berat badan mahamahasiswa yang menjadi sukarelawan donor darah. Kemungkinan nilainya bisa berupa:

52,5 kg, 63,8 kg, 70,12 kg, dan seterusnya.

Karena berat badan dapat berupa nilai riil apa pun dalam suatu rentang, maka Y merupakan variabel acak kontinu. Di antara 50 kg dan 51 kg saja, terdapat kemungkinan nilai yang tak terhingga banyaknya.

Mengapa Variabel Acak Itu Penting

Variabel acak memungkinkan kita untuk melampaui pernyataan yang samar, seperti: "Penjualan terkadang naik dan terkadang turun." Sebaliknya, variabel acak memungkinkan kita untuk memeriksa ketidakpastian dengan cara yang presisi dan matematis. Dengan variabel acak, kita dapat:

- Menetapkan probabilitas untuk setiap kemungkinan hasil
- Merangkum ketidakpastian menggunakan distribusi probabilitas
- Menghitung besaran-besaran seperti nilai harapan (*expected value*) dan variabilitas
- Memprediksi kemungkinan terjadinya peristiwa ekstrem atau tidak biasa

Singkatnya, variabel acak mengubah ketidakpastian menjadi besaran terukur yang dapat diperiksa dan dimodelkan. Kemampuan ini sangat penting dalam bidang-bidang seperti sains data, analitik bisnis, studi kesehatan, keuangan, dan teknik, di mana keputusan sering kali harus diambil dalam kondisi ketidakpastian. Pada bagian selanjutnya, kita akan membahas distribusi probabilitas, yang menjelaskan bagaimana nilai-nilai suatu variabel acak tersebar dan seberapa besar kemungkinan terjadinya setiap hasil yang mungkin.

10.3 KONSEP DASAR DISTRIBUSI PROBABILITAS

Pada bagian sebelumnya, kita telah memperkenalkan konsep variabel acak representasi numerik dari hasil suatu proses acak. Sebagai contoh, saat sebuah dadu standar bersisi enam dilempar, hasilnya tidak dapat dipastikan, namun kita dapat menetapkan nilai numerik (1, 2, 3, 4, 5, atau 6) untuk setiap kemungkinan hasil tersebut. Dengan mengubah hasil menjadi angka, kita dapat menerapkan perangkat matematika untuk memeriksa keacakan. Akan tetapi, sekadar mengetahui nilai-nilai yang mungkin dari suatu variabel acak saja tidaklah cukup. Hal ini ibarat memiliki daftar tujuan perjalanan tanpa mengetahui seberapa besar peluang Anda untuk sampai ke setiap tujuan tersebut. Untuk benar-benar memahami suatu proses acak, kita juga harus mengetahui seberapa besar kemungkinan terjadinya setiap hasil. Di sinilah distribusi probabilitas menjadi sangat penting. Distribusi probabilitas memberikan deskripsi lengkap mengenai suatu variabel acak dengan menetapkan:

- Semua nilai yang mungkin dapat diambil oleh variabel acak tersebut
- Probabilitas yang dikaitkan dengan setiap nilai tersebut

Dengan kata lain, distribusi probabilitas berfungsi sebagai cetak biru ketidakpastian, yang menggambarkan bagaimana probabilitas tersebar di seluruh kemungkinan hasil. Bayangkan distribusi probabilitas sebagai peta cuaca untuk fenomena acak. Sebuah peta cuaca tidak hanya mencantumkan kondisi cuaca yang mungkin terjadi (cerah, berawan, hujan), tetapi juga menunjukkan probabilitas terjadinya setiap kondisi tersebut. Demikian pula, distribusi probabilitas memperlihatkan hasil apa saja yang mungkin terjadi beserta tingkat kemungkinan terjadinya masing-masing hasil tersebut. Dengan mengubah ketidakpastian menjadi deskripsi matematis yang terstruktur, distribusi probabilitas memungkinkan kita untuk memeriksa pola dalam fenomena acak dan membuat prediksi yang berdasar.

Komponen Penyusun Distribusi Probabilitas

Setiap distribusi probabilitas dibangun dari dua komponen mendasar.

1. **Ruang Sampel (S): Semesta Kemungkinan**

   Ruang sampel adalah himpunan lengkap dari semua hasil yang mungkin terjadi dalam suatu eksperimen acak. Sebagai contoh, saat melempar dadu bersisi enam yang adil, ruang sampelnya adalah:

   $$S = \{1, 2, 3, 4, 5, 6\}$$

   Setiap nilai mewakili kemungkinan hasil dari eksperimen tersebut. Ruang sampel menentukan apa saja yang mungkin terjadi, namun belum menunjukkan seberapa besar peluang munculnya masing-masing hasil tersebut.

2. **Probabilitas: Mengukur Kemungkinan**

   Setelah hasil-hasil yang mungkin diketahui, langkah selanjutnya adalah menetapkan probabilitas untuk setiap hasil tersebut. Probabilitas mengukur seberapa besar kemungkinan suatu peristiwa akan terjadi. Untuk sebuah dadu yang adil, setiap sisi memiliki peluang yang sama untuk muncul. Karena terdapat enam kemungkinan hasil, maka probabilitas untuk masing-masing hasil adalah:

   $$P(X = x) = 1 / 6$$

   dengan keterangan:

   - X melambangkan variabel acak (hasil lemparan dadu)
   - x melambangkan nilai spesifik yang dapat diambil oleh X

   Distribusi probabilitas ini dapat diringkas sebagai berikut:

Hasil (X)	**Probabilitas P(X)**
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Aturan Dasar Probabilitas

Untuk setiap distribusi probabilitas: Jumlah probabilitas dari semua hasil yang mungkin harus sama dengan 1. Aturan ini mencerminkan suatu kebenaran sederhana: sesuatu di dalam ruang sampel pasti akan terjadi. Jika jumlah probabilitas tidak sama dengan 1, maka berarti:

- beberapa hasil tidak tercantum, atau

- penetapan probabilitas tidak tepat

Dua Dunia Keacakan: Distribusi Diskrit Vs. Kontinu

Distribusi probabilitas terbagi menjadi dua kategori utama, bergantung pada jenis variabel acak yang terlibat.

1. Distribusi Probabilitas Diskrit

Distribusi diskrit menggambarkan variabel acak yang memiliki nilai-nilai yang dapat dihitung.

Karakteristik

- ❖ Hasil bersifat terpisah dan dapat dihitung
- ❖ Biasanya berupa bilangan bulat

Contoh

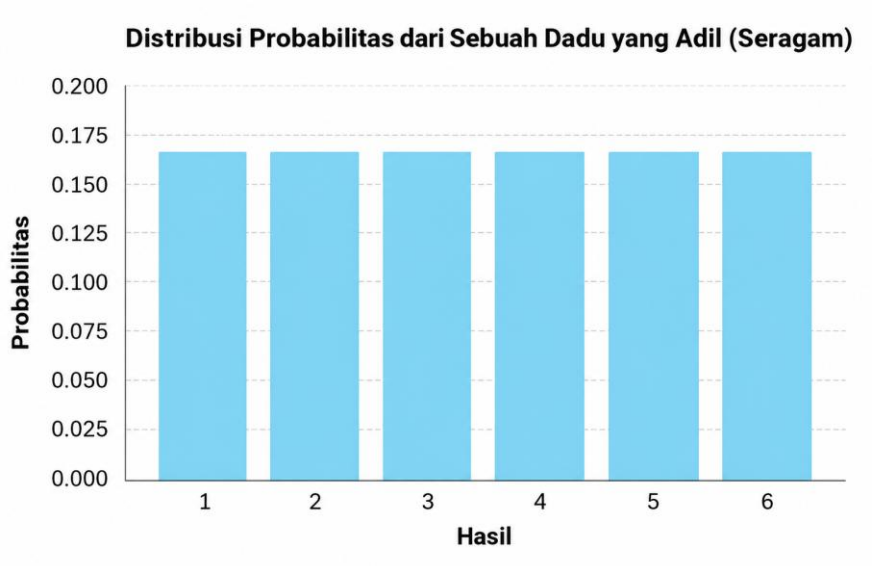
- ❖ Hasil lemparan dadu (Distribusi seragam/*uniform*)
- ❖ Jumlah sisi angka (*heads*) dalam 10 kali lemparan koin (Distribusi binomial)
- ❖ Jumlah pelanggan yang datang ke toko dalam satu jam (sering dimodelkan dengan distribusi Poisson)

Visualisasi

Distribusi diskrit biasanya direpresentasikan menggunakan diagram batang, di mana setiap batang menunjukkan probabilitas dari hasil tertentu.

Dari Kewajaran yang Seragam hingga Pola Keberhasilan: Dasar-Dasar Keacakan

Untuk dadu yang adil, setiap hasil memiliki probabilitas yang sama (1/6), yang menghasilkan bentuk "datar" di mana semua batang memiliki tinggi yang sama. Ini disebut Distribusi Seragam (*Uniform Distribution*).



Gambar 10.1 Distribusi seragam (*uniform distribution*)

Namun, dalam sains data, kita sering berurusan dengan proses di mana hasil-hasil yang mungkin muncul tidak memiliki peluang yang sama. Banyak fenomena di dunia nyata mengikuti pola biner—Sukses atau Gagal, Ya atau Tidak, 1 atau 0. Untuk memahami pola-pola

ini, kita mulai dengan eksperimen acak yang paling sederhana: satu percobaan dengan hanya dua kemungkinan hasil.

1. **Distribusi Bernoulli: Percobaan Tunggal**

   Distribusi Bernoulli memodelkan satu percobaan yang dapat menghasilkan tepat dua kemungkinan hasil. Contoh yang familier adalah pelemparan koin.

   - Hasil: Hanya ada dua kemungkinan hasil—Angka (Sukses) atau Gambar (Gagal).
   - Peluang: Jika koin tersebut adil (*fair*), peluang (p) untuk mendapatkan hasil sukses adalah 0,5.

   Struktur sederhana ini mendefinisikan percobaan Bernoulli, dan model peluang yang mengaturnya disebut distribusi Bernoulli. Distribusi ini menjadi landasan untuk menganalisis data biner dalam sains data, seperti:

   - Apakah pelanggan membeli produk tersebut (Ya/Tidak)?
   - Apakah pengguna mengeklik tautan tersebut (Klik/Tidak Klik)?
   - Apakah transaksi mengalami kegagalan (Gagal/Sukses)?

   Karena begitu banyak proses digital dan bisnis yang menghasilkan keluaran biner, distribusi Bernoulli menjadi salah satu elemen dasar yang paling fundamental dalam analitik terapan.

2. **Distribusi Binomial: Gambaran Besar**

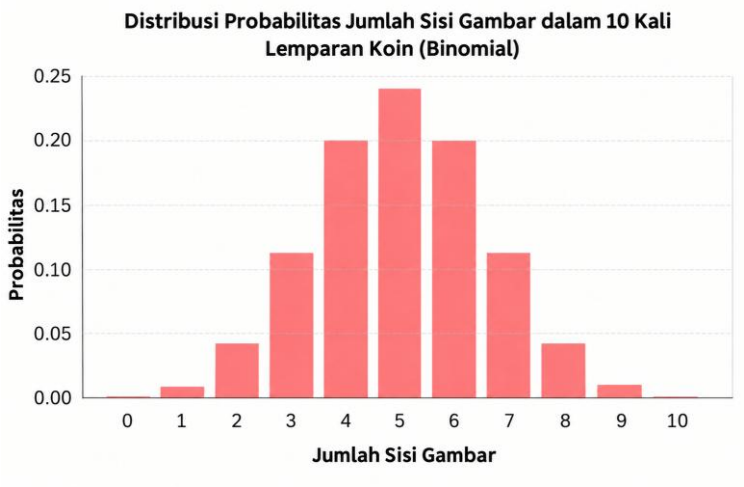
   Apa yang terjadi jika kita mengulangi percobaan Bernoulli tersebut berkali-kali? Misalkan kita melempar koin adil yang sama sebanyak 10 kali. Sekarang, kita tidak lagi hanya melihat satu hasil saja, melainkan jumlah total keberhasilan yang diperoleh. Misalkan variabel acak kita adalah:

   $$Y = Jumlah\ sisi\ angka\ (heads)\ yang\ muncul\ dalam\ 10\ kali\ lemparan.$$

   Kemungkinan hasil yang muncul kini berkisar dari 0 hingga 10 sisi angka. Dalam skenario ini, beberapa hasil jauh lebih mungkin terjadi dibandingkan yang lain:

   - 0 atau 10 sisi angka: Hasil ini mungkin terjadi, tetapi sangat jarang.
   - 5 sisi angka: Ini adalah hasil yang paling sering muncul karena merupakan hasil yang "diharapkan" (*expected result*) untuk sebuah koin yang adil.

   Jika kita memetakan probabilitas untuk setiap kemungkinan hasil (dari 0 hingga 10), kita akan melihat pola berbentuk lonceng yang berpusat di angka 5. Pola ini disebut Distribusi Binomial. Distribusi binomial pada dasarnya merupakan "jumlah" dari beberapa percobaan Bernoulli yang saling bebas.



Gambar 10.2 Distribusi Binomial

2. Distribusi Probabilitas Kontinu

Distribusi kontinu menggambarkan variabel acak yang dapat mengambil nilai apa pun dalam suatu rentang tertentu.

Karakteristik

- Hasil pengamatan berada pada skala kontinu
- Terdapat nilai yang tak terhingga banyaknya di antara dua titik mana pun

Contoh

- Tinggi atau berat badan manusia
- Pengukuran suhu
- Waktu yang diperlukan untuk menyelesaikan suatu tugas

Visualisasi

Distribusi kontinu direpresentasikan oleh kurva mulus yang disebut fungsi kepadatan probabilitas (*probability density functio*n/PDF). Untuk variabel kontinu, probabilitas untuk mendapatkan satu nilai spesifik yang tepat adalah nol. Sebagai gantinya, probabilitas dihitung berdasarkan rentang nilai, yang setara dengan luas area di bawah kurva.

Primadona Utama: Distribusi Normal (Kurva Lonceng)

Di antara berbagai distribusi kontinu, yang paling terkenal adalah Distribusi Normal, yang sering disebut sebagai kurva lonceng. Pola simetris berbentuk lonceng ini muncul dalam banyak fenomena alam maupun sosial. Contohnya meliputi:

- tinggi badan manusia
- skor tes standar
- kesalahan pengukuran dalam eksperimen ilmiah
- fluktuasi dalam berbagai pengukuran ekonomi dan biologis

Karena kemunculannya yang luas serta sifat-sifat matematisnya yang kuat, distribusi normal memegang peranan sentral dalam inferensi statistik, teori pengambilan sampel, dan pengujian hipotesis. Bagian selanjutnya akan membahas distribusi normal secara mendalam.

Bahasa Fungsi: PMF dan PDF

Meskipun tabel dan grafik membantu memvisualisasikan distribusi, matematika menyediakan cara yang tepat untuk mendeskripsikannya menggunakan fungsi.

Fungsi Massa Peluang (*Probability Mass Function* – PMF)

Untuk distribusi diskret, kita menggunakan Fungsi Massa Peluang (PMF). PMF menetapkan nilai peluang untuk setiap kemungkinan nilai dari variabel acak diskret. Sifat-sifat utama PMF:

1. $0 \leq P(X = x) \leq 1$ untuk semua nilai x
2. Jumlah seluruh peluang dari semua nilai yang mungkin adalah 1

Untuk contoh dadu yang adil:

$$P(X = x) = 1/6 \text{ untuk } x \in \{1, 2, 3, 4, 5, 6\}$$

PMF mendistribusikan peluang sebagai massa individual yang ditetapkan pada setiap nilai.

Fungsi Kepadatan Peluang (*Probability Density Function* – PDF)

Untuk distribusi kontinu, kita menggunakan Fungsi Kepadatan Peluang (PDF). Berbeda dengan PMF, PDF tidak memberikan nilai peluang untuk suatu nilai yang tepat (spesifik). Sebaliknya, PDF mendeskripsikan bagaimana peluang didistribusikan di sepanjang rentang nilai tertentu. Sifat-sifat penting PDF:

- Fungsi ini tidak pernah bernilai negatif
- Total luas di bawah kurva sama dengan 1

Peluang dihitung sebagai luas area di bawah kurva di antara dua nilai.

Wawasan Utama

Pada tahap ini, gagasan yang paling penting bukanlah rumus-rumus itu sendiri, melainkan konsep di baliknya:

- Setiap distribusi peluang merepresentasikan pola kemungkinan dari berbagai hasil yang mungkin terjadi.
- PMF dan PDF hanyalah alat matematika yang digunakan untuk mendeskripsikan pola-pola tersebut secara tepat.

Perjalanan Anda ke dalam Dunia Distribusi

Pada bagian-bagian selanjutnya, kita akan menjelajahi beberapa distribusi peluang yang paling penting dalam statistika dan sains data. Untuk setiap distribusi, kita akan mengkaji:

1. Struktur matematika yang mendefinisikannya
2. Proses di dunia nyata yang dimodelkannya
3. Bagaimana distribusi tersebut dapat diterapkan dalam analisis data menggunakan Excel

Anda akan menjumpai distribusi-distribusi yang sering muncul dalam data nyata:

- **Distribusi Binomial**

  Memodelkan proses yang melibatkan hasil sukses atau gagal, seperti dalam pengujian A/B (A/B testing) atau penghitungan jumlah cacat (defek).

- **Distribusi Poisson**
  Menggambarkan peristiwa langka yang terjadi sepanjang waktu atau ruang, seperti kedatangan pelanggan, kegagalan sistem, atau permintaan pada situs web.
- **Distribusi Normal**
  Kurva berbentuk lonceng yang terkenal, yang mendasari banyak fenomena alam maupun sosial.

Setiap distribusi memberikan sudut pandang berbeda untuk memahami keacakan. Menguasai distribusi-distribusi ini akan memungkinkan Anda mengenali pola yang tersembunyi dalam data serta membangun model statistik yang lebih tangguh.
Dan ingat:

> *Setiap kumpulan data memiliki distribusi probabilitas yang mendasarinya. Belajar mengenali distribusi tersebut merupakan salah satu keterampilan paling ampuh dalam sains data.*

10.4 DISTRIBUSI DISKRIT

Distribusi Binomial

Apa itu Distribusi Binomial?

Distribusi Binomial merupakan salah satu distribusi probabilitas diskrit yang paling penting dalam statistika. Distribusi ini memodelkan jumlah keberhasilan dalam sejumlah percobaan berulang yang tetap, di mana setiap percobaan hanya menghasilkan dua kemungkinan hasil:

- Keberhasilan (Sukses)
- Kegagalan

Proses semacam ini sering dijumpai dalam sains data dan analisis dunia nyata. Contohnya meliputi:

- Pengguna mengeklik atau tidak mengeklik (pengujian A/B)
- Produk mengalami cacat atau tidak (pengendalian kualitas)
- Pasien memberikan respons atau tidak terhadap suatu pengobatan (studi medis)

Distribusi binomial menjawab pertanyaan mendasar:
Berapakah probabilitas untuk memperoleh tepat k keberhasilan dalam n percobaan yang saling bebas?

Kapan Harus Menggunakan Distribusi Binomial

Model binomial dapat diterapkan jika kondisi-kondisi berikut terpenuhi:

1. Terdapat jumlah percobaan yang tetap (n)
2. Setiap percobaan bersifat saling bebas (hasil dari satu percobaan tidak memengaruhi percobaan lainnya)
3. Setiap percobaan memiliki dua kemungkinan hasil (keberhasilan/kegagalan)
4. Probabilitas keberhasilan (p) adalah sama untuk setiap percobaan

Jika salah satu dari asumsi ini tidak terpenuhi, maka model binomial mungkin tidak tepat untuk digunakan.

Rumus Binomial (Fungsi Massa Probabilitas)

Probabilitas untuk memperoleh tepat k keberhasilan dalam n percobaan ditentukan oleh fungsi massa probabilitas binomial:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Di mana:

- X = variabel acak (jumlah keberhasilan)
- k = jumlah keberhasilan yang diminati
- n = jumlah total percobaan
- p = peluang keberhasilan pada setiap percobaan
- $\binom{n}{k} = \left(\frac{n!}{k!(n-k)!}\right)$ =jumlah cara untuk memilih k keberhasilan dari n percobaan.

Rumus ini menggabungkan tiga gagasan utama:

- Peluang keberhasilan → pk
- Peluang kegagalan → $(1-p)n-k$
- Jumlah kemungkinan susunan keberhasilan dan kegagalan

Contoh: Pelemparan Koin

Misalkan kita melempar koin yang adil sebanyak 3 kali:

- n = 3
- p = 0,5
- Tentukan peluang munculnya tepat 2 sisi angka (k = 2)

$$P(X = 2) = \binom{3}{2} (0.5)^2 (0.5)^1 = 3 \times 0.25 \times 0.5 = 0.375$$

Interpretasi:

Terdapat peluang sebesar 37,5% untuk mendapatkan tepat 2 sisi angka dalam 3 kali lemparan.

Memahami Faktorial

Dalam matematika, $n!$ adalah singkatan dari n-faktorial, yaitu jumlah cara untuk menyusun n item yang berbeda.

Ini mewakili hasil perkalian semua bilangan bulat positif dari n hingga 1.

Ini adalah konsep fundamental dalam probabilitas dan statistik karena mewakili jumlah cara berbeda Anda dapat mengatur sekumpulan objek.

Untuk setiap bilangan bulat positif n:

$$n! = n\ x\ (n-1)\ x\ (n-2)\ x \ldots x\ 1$$

Contoh:

$-\ 3!\ =\ 3\ x\ 2\ x\ 1\ =\ 6$
$-\ 5!\ =\ 5\ x\ 4\ x\ 3\ x\ 2\ x\ 1\ =\ 120$
$-\ 1!\ =\ 1$

Implementasi Excel: BINOM.DIST()

Dalam praktiknya, praktisi data jarang menghitung probabilitas binomial secara manual. Perangkat lunaklah yang menangani perhitungan tersebut. Excel menyediakan fungsi bawaan:

$$BINOM.DIST(number_s, trials, probability_s, cumulative)$$

Parameter:

- number_s → jumlah keberhasilan (k)
- trials → total percobaan (n)
- probability_s → probabilitas keberhasilan (p)
- cumulative →
  - FALSE → probabilitas eksak $P(X\ =\ k)$
  - TRUE → probabilitas kumulatif $P(X\ \leq\ k)$

Contoh di Excel

$$= BINOM.DIST(2, 3, 0.5, FALSE)$$

Hasil:

$$0.375$$

Probabilitas Kumulatif vs. Probabilitas Eksak

Memahami perbedaan ini sangatlah penting:

- Probabilitas eksak
  $P(X\ =\ k)$ → probabilitas tepat sebanyak k keberhasilan
- Probabilitas kumulatif
  $P(X\ \leq\ k)$ → probabilitas paling banyak k keberhasilan

Contoh:

$$= BINOM.DIST(2, 3, 0.5, TRUE)$$

Ini menghitung:

$$P(X\ \leq\ 2)\ =\ P(0)\ +\ P(1)\ +\ P(2)$$

Karakteristik Utama Distribusi Binomial

Rata-rata (Nilai Harapan)

$$\mu = n \times p$$

Varians

$$\sigma^2 = n \times p \times (1 - p)$$

Simpangan Baku

$$\sigma = \sqrt{n \times p \times (1 - p)}$$

Interpretasi

- Rata-rata (*mean*) merepresentasikan jumlah keberhasilan yang diharapkan
- Varians mengukur variabilitas dalam hasil
- Simpangan baku (standar deviasi) mencerminkan sebaran tipikal di sekitar rata-rata

Bentuk Distribusi

- Simetris jika $p = 0{,}5$
- Menceng ke kanan (*right-skewed*) jika $p < 0{,}5$
- Menceng ke kiri (*left-skewed*) jika $p > 0{,}5$

Contoh

Jika:

- n = 10
- p = 0,3

Maka:

- ✓ Rata-rata $= 10 \times 0{,}3 = 3$
- ✓ Varians $= 10 \times 0{,}3 \times 0{,}7 = 2{,}1$

Interpretasi:

Kita mengharapkan rata-rata sekitar 3 keberhasilan, dengan variabilitas di sekitar nilai tersebut.

Kaitan dengan Distribusi Lain

Distribusi Bernoulli: Kasus khusus di mana $n = 1$. Distribusi Normal (Pendekatan): Untuk nilai n yang besar, distribusi binomial dapat didekati dengan distribusi normal. Kaitan ini menjadi penting dalam bab-bab selanjutnya mengenai statistik inferensial.

Penerapan di Dunia Nyata

Distribusi binomial digunakan secara luas dalam praktik:

- Pengendalian Kualitas → jumlah barang cacat
- Pembelajaran Medis → tingkat keberhasilan pengobatan
- Analitik Pemasaran → tingkat konversi (*conversion rates*)
- Analitik Olahraga → hasil menang/kalah

- Analisis Survei → respons ya/tidak

Contoh Praktis

Contoh 1: Pengendalian Kualitas

Sebuah pabrik memproduksi bola lampu dengan tingkat kerusakan 5%. Jika 20 bola lampu dipilih secara acak, berapakah probabilitas tepat dua di antaranya rusak?

$$n = 20, p = 0{,}05, k = 2$$

$$= BINOM.DIST(2, 20, 0{,}05, FALSE)$$

Hasil:

$$0{,}1887\ (18{,}87\%)$$

Contoh 2: Pengobatan Medis

Sebuah pengobatan memiliki tingkat keberhasilan 70%. Jika 15 pasien menerima pengobatan tersebut, berapakah probabilitas setidaknya 12 pasien akan sembuh?

$$n = 15$$

$$p = 0{,}70$$

Kita ingin mencari

$$P(X \geq 12)$$

yang setara dengan

$$1 - P(X \leq 11)$$

$$= 1 - BINOM.DIST(11, 15, 0{,}70, TRUE)$$

Hasil:

$$0{,}2969\ (29{,}69\%)$$

Distribusi Poisson

Apa itu Distribusi Poisson?

Distribusi Poisson adalah distribusi probabilitas diskret yang memodelkan jumlah kejadian yang berlangsung dalam interval waktu, ruang, atau area tertentu, dengan asumsi tingkat kejadian rata-rata yang telah diketahui. Distribusi ini sangat berguna ketika kita menghitung seberapa sering suatu peristiwa terjadi dalam domain kontinu, seperti:

- waktu (jumlah panggilan per menit)
- ruang (jumlah kerusakan per meter persegi)
- volume (jumlah partikel per satuan volume)

Distribusi Poisson berlaku ketika peristiwa-peristiwa terjadi:

- Secara independen satu sama lain
- Dengan tingkat rata-rata yang konstan, yang dilambangkan dengan λ (lambda).

Secara sederhana, distribusi ini menjawab pertanyaan:

Dengan tingkat kejadian rata-rata tertentu, berapakah probabilitas terjadinya tepat k peristiwa dalam interval yang telah ditetapkan?

Kapan Menggunakan Distribusi Poisson

Model Poisson tepat digunakan jika kondisi-kondisi berikut terpenuhi:

1. Peristiwa dihitung dalam interval tertentu (waktu, ruang, atau wilayah)
2. Peristiwa terjadi secara independen
3. Tingkat rata-rata (λ) bersifat konstan
4. Peristiwa terjadi satu per satu (tidak ada kejadian yang berlangsung secara bersamaan)
5. Peristiwa tergolong jarang terjadi dalam interval yang kecil

Kondisi-kondisi ini menjadikan distribusi Poisson sangat cocok untuk memodelkan peristiwa acak yang jarang terjadi.

Rumus Poisson (Fungsi Massa Probabilitas)

Probabilitas terjadinya tepat k peristiwa, dengan tingkat rata-rata λ, dirumuskan sebagai berikut:

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

di mana:

- X = variabel acak (jumlah kejadian)
- k = jumlah kejadian (0, 1, 2, …)
- λ (lambda) = rata-rata jumlah kejadian per interval
- $e \approx 2{,}71828$ (bilangan Euler)

Memahami λ (Lambda)

Parameter λ memegang peranan sentral:

- Parameter ini merepresentasikan jumlah kejadian yang diharapkan dalam suatu interval
- Parameter ini menentukan pusat sekaligus sebaran dari distribusi tersebut

Sebagai contoh:

- Jika λ = 5 pelanggan/jam → rata-rata, terdapat 5 pelanggan yang datang setiap jam
- Jika λ = 0,5 cacat/halaman → cacat merupakan kejadian yang jarang terjadi

Contoh: Kedatangan Pelanggan

Sebuah toko menerima rata-rata 5 pelanggan per jam (λ = 5). Berapakah probabilitas tepat 3 pelanggan datang dalam satu jam ke depan?

$$P(X = 3) = \frac{e^{-5} \cdot 5^3}{3!} \approx 0.1404$$

Interpretasi:
Terdapat probabilitas sekitar 14,04% untuk tepat 3 kedatangan.
Implementasi Excel: POISSON.DIST()
Excel menyediakan fungsi bawaan untuk probabilitas Poisson:

$$POISSON.DIST(x, mean, cumulative)$$

Parameter:

- x → jumlah kejadian (k)
- mean → λ (rata-rata tingkat kejadian)
- cumulative →
  - FALSE → probabilitas tepat $P(X = x)$
  - TRUE → probabilitas kumulatif $P(X \leq x)$

Contoh di Excel

$$= POISSON.DIST(3, 5, FALSE)$$

Hasil:

$$0{,}1404$$

Probabilitas Kumulatif vs. Probabilitas Tepat

- Probabilitas tepat $P(X = k)$: probabilitas terjadinya tepat k kejadian
- Probabilitas kumulatif $P(X \leq k)$: probabilitas terjadinya paling banyak k kejadian

Contoh

$$= POISSON.DIST(3, 5, TRUE)$$

Ini menghitung:

$$P(X \leq 3) = P(0) + P(1) + P(2) + P(3)$$

Sifat-Sifat Utama Distribusi Poisson
Untuk distribusi Poisson: Rata-rata (Mean)

Mean $\mu = \lambda$
Variance $\sigma^2 = \lambda$

Standard Deviation $\sigma = \sqrt{\lambda}$

Wawasan Penting

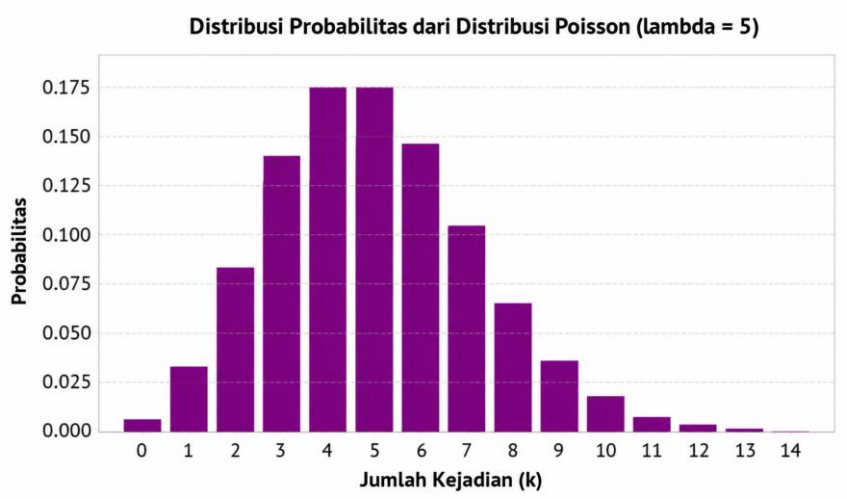
Untuk distribusi Poisson:

$$Rata - rata = Varians = \lambda$$

Sifat unik ini membedakannya dari banyak distribusi lainnya.

Bentuk Distribusi

- Menceng ke kanan (*right-skewed*) ketika λ bernilai kecil
- Menjadi lebih simetris dan berbentuk lonceng seiring dengan meningkatnya nilai λ
- Mendekati distribusi Normal untuk nilai λ yang besar



Gambar 10.3 Distribusi Poisson (λ = 5)

Hubungan dengan Distribusi Lain

- **Aproksimasi Binomial**

  Distribusi Poisson dapat digunakan untuk mengaproksimasi distribusi binomial ketika:
  - n bernilai besar
  - p bernilai kecil
  - $\lambda = np$
- **Aproksimasi Normal**

  Untuk nilai λ yang besar, distribusi Poisson mendekati distribusi normal.

Hubungan-hubungan ini penting dalam pemodelan dan inferensi statistik.

Penerapan di Dunia Nyata

Distribusi Poisson banyak digunakan untuk memodelkan jumlah kejadian:

- Perbankan → jumlah nasabah yang datang per jam
- Pusat Panggilan (*Call Center*) → jumlah panggilan per menit
- Manufaktur → jumlah cacat per satuan luas
- Asuransi → jumlah klaim per bulan

- Penyuntingan → jumlah kesalahan ketik per halaman
- Rekayasa Lalu Lintas → jumlah kecelakaan di persimpangan jalan
- Genetika → jumlah mutasi dalam urutan DNA
- Seismologi → jumlah gempa bumi per tahun

Contoh Praktis

Contoh 1: Pusat Panggilan (*Call Center*)

Sebuah pusat panggilan menerima rata-rata 3 panggilan per menit. Berapakah probabilitas menerima tepat 5 panggilan dalam satu menit tertentu?

$$\lambda = 3$$
$$k = 5$$
$$= POISSON.DIST(5, 3, FALSE)$$

Hasil: $0{,}1008\ (10{,}08\%)$

Contoh 2: Lalu Lintas Situs Web

Sebuah situs web menerima rata-rata 8 pengunjung per jam. Berapakah probabilitas menerima paling banyak 5 pengunjung dalam satu jam ke depan?

$$\lambda = 8$$
$$Kita\ mencari\ P(X \leq 5)$$
$$= POISSON.DIST(5, 8, TRUE)$$

Hasil: $0{,}1912\ (19{,}12\%)$

Contoh 3: Cacat dalam Manufaktur

Sebuah produsen tekstil menemukan rata-rata 2 cacat per 100 m persegi kain. Berapakah probabilitas ditemukannya lebih dari 3 cacat dalam sampel seluas 100 m persegi?

$$\lambda = 2$$
$$Kita\ ingin\ mencari\ P(X > 3) = 1 - P(X \leq 3)$$
$$= 1 - POISSON.DIST(3, 2, TRUE)$$

Hasil: $0{,}1429\ (14{,}29\%)$

10.5 DISTRIBUSI PROBABILITAS KONTINU

Distribusi Normal (*Gaussian*)

Sejauh ini dalam bab ini, kita telah membahas distribusi Binomial dan Poisson—keduanya merupakan contoh distribusi probabilitas diskret, di mana hasil pengamatannya dapat dihitung (contoh: jumlah keberhasilan, jumlah kedatangan). Sekarang, kita akan beralih ke distribusi yang paling penting dalam statistika dan sains data: Distribusi Normal. Berbeda dengan distribusi diskret, distribusi normal berlaku untuk data kontinu yaitu pengukuran yang dapat bernilai apa pun dalam rentang tertentu, seperti tinggi badan, berat badan, waktu, dan suhu.

Memahami Kurva Lonceng (*Bell Curve*)

Distribusi Normal, yang sering disebut sebagai kurva lonceng, dicirikan oleh bentuknya yang simetris menyerupai lonceng. Karakteristik utamanya meliputi:

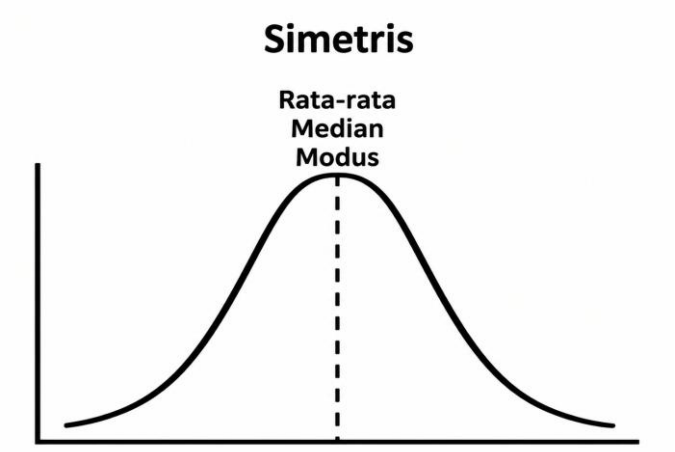
- Memiliki satu puncak (unimodal) di bagian tengah
- Simetri sempurna di sekitar titik tengah
- Nilai-nilai yang mengecil secara bertahap menuju kedua ujung ekstrem

Dalam banyak kumpulan data dunia nyata, nilai-nilai cenderung mengelompok di sekitar nilai rata-rata, dengan jumlah observasi yang semakin sedikit seiring menjauhnya nilai tersebut dari pusat. Pola alami ini menghasilkan bentuk yang menyerupai lonceng.

Karakteristik Utama Distribusi Normal

1. **Simetri**: Sisi kiri dan sisi kanan distribusi merupakan cerminan satu sama lain.
2. **Tendensi Sentral**: Nilai rata-rata (*mean*), median, dan modus adalah sama dan terletak di bagian tengah.
3. **Ekor Asimtotik**: Kurva mendekati sumbu horizontal, namun tidak pernah menyentuhnya. Kurva ini memanjang tanpa batas ke kedua arah.

Kurva Distribusi Normal:



Gambar 10.4 Kurva distribusi normal

Notasi dan Parameter

Variabel yang terdistribusi normal ditulis sebagai:

$$X \sim N(\mu, \sigma^2)$$

Di mana:

- X = variabel acak
- N = distribusi normal
- μ = rata-rata (pusat distribusi)
- σ^2 = varians
- σ = simpangan baku (standar deviasi)

Contoh

Jika tinggi badan mahasiswa (H) merupakan variabel yang mengikuti distribusi normal dengan:

- Rata-rata = 170 cm
- Simpangan baku = 8 cm

Maka kita tulis:

$$H \sim N(170,8^2)$$

Artinya: H berdistribusi normal dengan rata-rata 170 dan simpangan baku 8.

Peran μ dan σ dalam Distribusi Normal

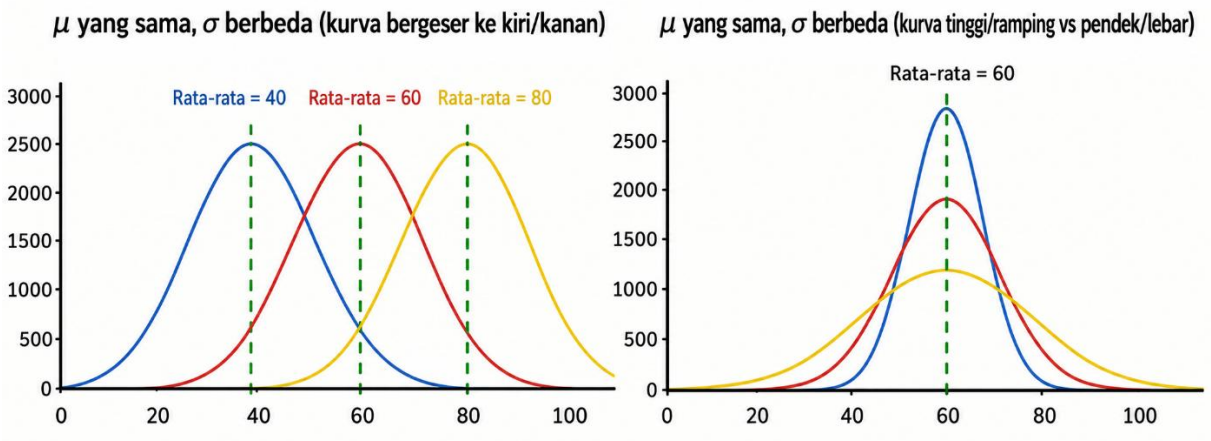
Distribusi normal sepenuhnya ditentukan oleh dua parameter: rata-rata (μ) dan simpangan baku (σ). Keduanya memiliki peran yang berbeda dan tidak tumpang tindih dalam membentuk kurva tersebut.

1. **Rata-rata (μ): Lokasi**
   - Menentukan titik pusat distribusi.
   - Menggeser seluruh kurva secara horizontal tanpa mengubah bentuknya.
   - Rata-rata yang lebih kecil menempatkan kurva lebih ke kiri; rata-rata yang lebih besar menggesernya ke kanan.

     Gagasan utama: Rata-rata hanya memengaruhi lokasi distribusi, bukan sebaran atau bentuknya.
2. **Simpangan Baku (σ): Sebaran**
   - Mengukur tingkat variabilitas dalam data.
   - Simpangan baku (σ) yang kecil menghasilkan kurva yang sempit dan tinggi, yang menunjukkan variabilitas rendah.
   - Simpangan baku (σ) yang besar menghasilkan kurva yang lebih lebar dan datar, yang menunjukkan variabilitas tinggi.

     Gagasan utama: Simpangan baku mengendalikan sebaran distribusi tanpa mengubah lokasi pusatnya.

Perbedaan Rata-rata vs Perbedaan Simpangan Baku (Std)



Gambar 10.5 Distribusi normal - perbedaan rata-rata vs perbedaan simpangan baku

Mengapa Hal Ini Penting

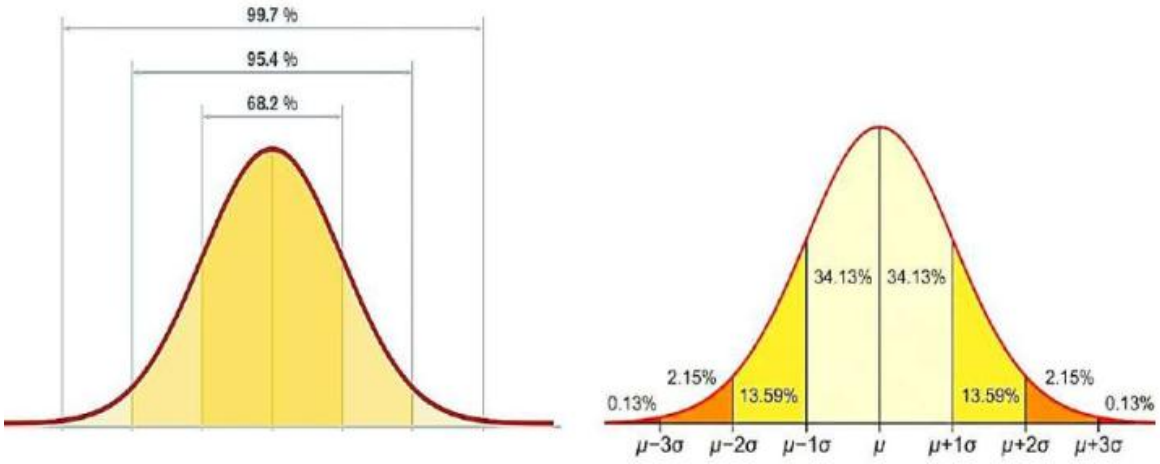
Memahami μ (rata-rata) dan σ (simpangan baku) memungkinkan Anda untuk:

- Membandingkan berbagai kumpulan data.
- Menafsirkan variabilitas dan konsistensi.
- Menafsirkan probabilitas dengan tepat (contohnya, mengapa nilai 80 bisa dianggap "biasa" di satu kelas, namun "luar biasa" di kelas lain).
- Memahami kegunaan standardisasi (skor-Z), yang menghilangkan perbedaan tersebut dan memungkinkan perbandingan yang adil antar kumpulan data yang berbeda.

Aturan Empiris (*Empirical Rule*)

Salah satu karakteristik utama dari distribusi normal adalah Aturan Empiris. Aturan Empiris memberikan cara intuitif untuk memahami probabilitas dalam distribusi normal apa pun. Terlepas dari berapa pun nilai rata-rata atau simpangan bakunya, semua distribusi normal menunjukkan proporsi perkiraan sebagai berikut:

- 68% dari nilai-nilai yang ada berada dalam rentang 1 simpangan baku dari rata-rata.
- 95% berada dalam rentang 2 simpangan baku.
- 99,7% berada dalam rentang 3 simpangan baku.



Gambar 10.6 Aturan Empiris – probabilitas dalam distribusi normal

Aturan ini menyediakan cara cepat untuk memahami sebaran data dan mendeteksi observasi yang tidak lazim.

Mengapa Distribusi Normal Begitu Penting?

1. **Teorema Limit Pusat (*Central Limit Theorem*/CLT)**: Distribusi sampling dari nilai rata-rata akan mendekati distribusi normal seiring dengan bertambahnya ukuran sampel—meskipun data aslinya tidak berdistribusi normal.
2. **Kejadian Alami**: Banyak fenomena di dunia nyata mengikuti pola normal, seperti tinggi dan berat badan, skor tes, serta kesalahan pengukuran.
3. **Kemudahan Matematis**: Distribusi normal memiliki sifat-sifat yang terdefinisi dengan baik, sehingga perhitungan probabilitas menjadi lebih mudah dilakukan. Distribusi ini memiliki probabilitas yang dapat diprediksi (aturan empiris).

4. **Landasan untuk Inferensi Statistik**: Banyak metode statistik (seperti selang kepercayaan dan pengujian hipotesis) bergantung pada asumsi normalitas.

Fungsi Kepadatan Probabilitas (*Probability Density Function*/PDF)

Untuk distribusi kontinu, kita tidak menetapkan probabilitas pada satu nilai yang tepat. Sebaliknya, probabilitas didefinisikan dalam suatu rentang (*interval*). Fungsi Kepadatan Probabilitas (PDF) menggambarkan bagaimana probabilitas tersebar di seluruh nilai yang ada. Fungsi ini dinyatakan dengan rumus:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Di mana:

- e adalah basis logaritma natural (sekitar 2,718).
- π adalah konstanta matematika pi (sekitar 3,14159).

Konsep Utama

- PDF (Fungsi Kepadatan Peluang) memberikan tingkat kemungkinan relatif dari nilai-nilai tertentu.
- Tinggi kurva bukanlah sebuah peluang.
- Peluang diperoleh dari luas area di bawah kurva.
- Total luas area di bawah kurva = 1 (100%).

Perbedaan ini sangat penting saat bekerja dengan distribusi kontinu.

Implementasi dalam Excel

Dalam praktiknya, kita menggunakan perangkat lunak untuk menghitung peluang. Excel menyediakan fungsi bawaan yang sangat berguna.

1. NORM.DIST()

Fungsi ini menghitung peluang untuk distribusi normal.

$$NORM.DIST(x, mean, standard_dev, cumulative)$$

Parameter

- x → nilai yang ditinjau
- mean (μ)
- standard_dev (σ)
- cumulative
  - TRUE → peluang kumulatif $P(X \leq x)$
  - FALSE → kepadatan peluang (tinggi kurva)

Contoh

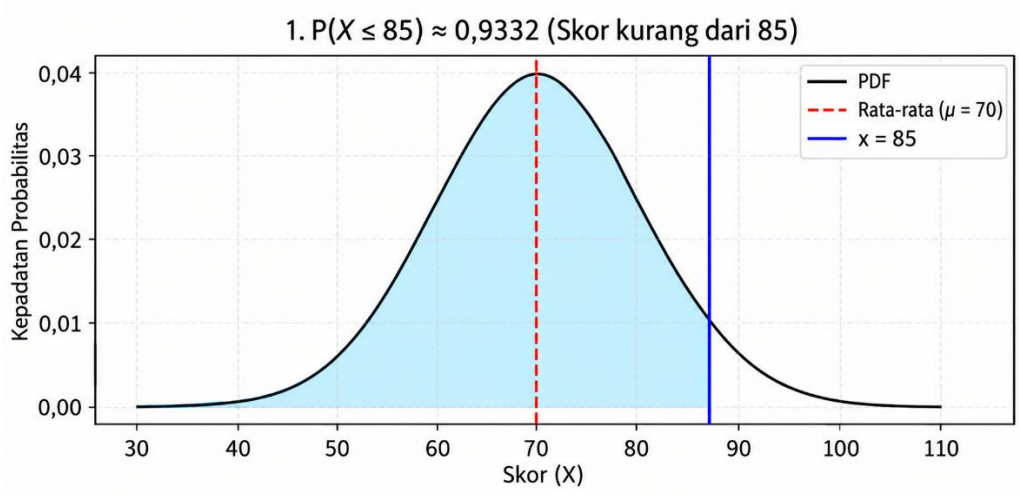
Misalkan nilai ujian berdistribusi normal dengan rata-rata (*mean*) = 70 dan standar deviasi = 10.

a) Berapa peluang seorang mahasiswa memperoleh nilai kurang dari 85?

Kita mencari $P(X \leq 85)$

$$= NORM.DIST(85, 70, 10, TRUE)$$
$$= 0{,}9332$$

Sekitar 93,32% mahasiswa memperoleh nilai di bawah 85. Hal ini ditunjukkan oleh area yang diarsir di bawah kurva kepadatan peluang berikut.



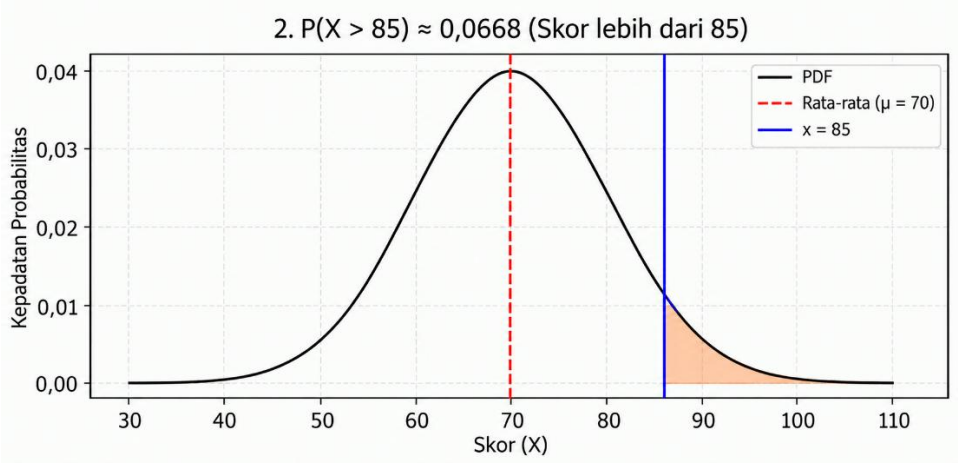
Gambar 10.7 Luas di bawah kurva kerja - probabilitas untuk benar $P(X \leq x)$

b) Temukan probabilitas bahwa seorang mahasiswa mendapat nilai lebih dari 85

Kita menginginkan $P(X > 85)$

$$= 1 - P(X \leq 85)$$
$$= 1 - NORM.DIST(85, 70, 10, TRUE)$$
$$= 0{,}0668\ (6{,}68\%)$$

Sekitar 6,68% mahasiswa mendapat nilai di atas 85.



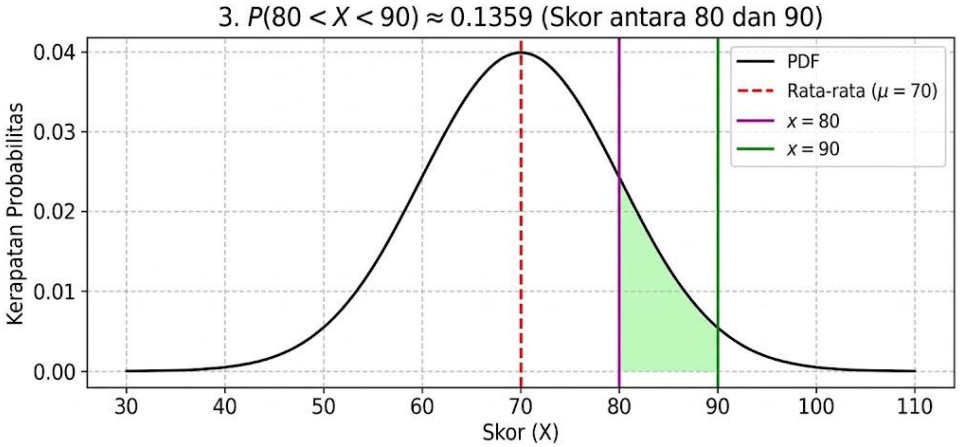
Gambar 10.8 Luas di bawah kurva kerja - probabilitas untuk benar $P(X > x)$

c) Temukan probabilitas bahwa seorang mahasiswa mendapat nilai antara 80 dan 90?

Kita menginginkan $P(80 < X < 90)$

$$= P(X \leq 90) - P(X \leq 80)$$
$$= NORM.DIST(90, 70, 10, TRUE) - NORM.DIST(80, 70, 10, TRUE)$$
$$= 0{,}1359\ (13{,}59\%)$$

Sekitar 13,59% mahasiswa memperoleh nilai antara 80 dan 90.



Gambar 10.9 Area di bawah kurva normal - probabilitas antara $P(x < X < x)$

2. $NORM.INV()$

Jika $NORM.DIST()$ dimulai dengan suatu nilai (x) dan menghasilkan probabilitas, maka fungsi $NORM.INV()$ melakukan sebaliknya. Fungsi ini menjawab pertanyaan: "Jika diketahui suatu probabilitas, nilai x berapakah yang bersesuaian dengannya?"
Sintaksnya:

$$= NORM.INV(probability, mean, standard_dev)$$

Contoh

Melanjutkan contoh nilai ujian (rata-rata/mean = 70, standar deviasi = 10).
Nilai berapakah yang bersesuaian dengan persentil ke-90 (10% teratas)?

$$= NORM.INV(0.90, 70, 10)$$
$$= 82{,}81$$

90% mahasiswa memperoleh nilai di bawah 82,8.

Standardisasi dan Distribusi Normal Standar

Sampai titik ini, kita telah mempelajari distribusi Normal—sebuah distribusi probabilitas kontinu yang memodelkan banyak fenomena alam maupun kejadian di dunia nyata. Kita juga telah melihat bagaimana fungsi-fungsi Excel seperti $NORM.DIST()$ dan

$NORM.INV()$ memungkinkan kita untuk menghitung probabilitas dan nilai-nilai dari distribusi ini. Namun, masih ada tantangan praktis:

Tidak semua distribusi normal dapat dibandingkan secara langsung.

Beberapa distribusi memiliki pusat (rata-rata/*mean*, μ) yang berbeda. Distribusi lainnya mungkin lebih lebar atau lebih sempit, bergantung pada standar deviasinya (σ).
Mengapa Hal Ini Penting? Perhatikan contoh berikut:

- Distribusi skor IQ dengan rata-rata 100 dan standar deviasi 15.
- Distribusi nilai ujian dengan rata-rata 65 dan standar deviasi 16.

Keduanya mengikuti distribusi normal, tetapi berada pada skala yang sangat berbeda. Membandingkan nilai mentah di antara keduanya akan memberikan hasil yang keliru.
Di sinilah standardisasi menjadi sangat penting.

Apa Itu Standardisasi?

Standardisasi adalah proses mengubah nilai-nilai dari suatu distribusi normal ke dalam skala umum, dengan ketentuan:

- Rata-rata (*mean*) = 0
- Simpangan baku (standar deviasi) = 1

Transformasi ini menghasilkan skor-z (*z-score*), yang didefinisikan sebagai:

$$z = \frac{x - \mu}{\sigma}$$

Keterangan:

- x = nilai asli
- μ = rata-rata
- σ = simpangan baku
- z = nilai terstandarisasi (skor z)

Skor z menunjukkan berapa banyak simpangan baku suatu nilai berada di atas atau di bawah rata-rata.

Contoh Ilustrasi

Pertimbangkan kumpulan data tinggi badan orang dewasa (dalam cm):

- Tinggi rata-rata (μ) = 170 cm
- Simpangan baku (σ) = 10 cm

Untuk menstandarisasi tinggi badan x, kurangi rata-rata dan bagi dengan simpangan baku. Skor z yang dihasilkan menyatakan tinggi badan dalam satuan simpangan baku dari rata-rata.

Tabel 10.1 Menghitung Skor Z

Tinggi Badan (cm)	Skor-Z
150	-2,0
160	-1,0

170	0,0
180	1,0
190	2,0

Wawasan Utama

Sebelum standardisasi:

- Rata-rata (Mean) = 170,
- Simpangan baku (SD) = 10
- $X \sim N(170, 10^2)$

Setelah standardisasi:

- ✓ Rata-rata (Mean) = 0,
- ✓ Simpangan baku (SD) = 1
- ✓ $Z \sim N(0,1)$

Dalam contoh tersebut:

- ❖ Tinggi 190 berada 2 skor-z di atas rata-rata.
- ❖ z = 2 berarti: nilai asli = rata-rata asli + 2*SD
- ❖ 170 + (2 x 10) = 190.

Standardisasi mengubah pengukuran dari satuan aslinya (cm, kg, °C, USD, dll.) menjadi satuan simpangan baku. Skor-z sebesar 2 selalu berarti "dua simpangan baku di atas rata-rata," terlepas dari skala aslinya.

Mengapa Standardisasi Penting

1. Dapat Dibandingkan di Berbagai Skala

Standardisasi memungkinkan perbandingan yang bermakna antara variabel-variabel yang diukur dalam satuan yang sangat berbeda atau berasal dari distribusi yang berbeda.

Contoh:

- ❖ Ukuran rumah mungkin berada +1 simpangan baku di atas rata-ratanya.
- ❖ Harga rumah mungkin juga berada +1 simpangan baku di atas rata-ratanya.

Meskipun satuannya berbeda (m^2 vs. Rp), kedua nilai tersebut menempati posisi relatif yang sama dalam distribusinya masing-masing, keduanya sama-sama berada "di atas rata-rata."

2. Menyederhanakan Perhitungan Probabilitas

Standardisasi memungkinkan semua distribusi normal dipetakan ke dalam satu distribusi acuan tunggal: distribusi normal standar. Alih-alih bekerja dengan kombinasi rata-rata dan simpangan baku yang tak terhingga banyaknya, kita mengubah nilai-nilai tersebut menjadi skor-z dan menggunakan satu distribusi probabilitas universal. Hal ini sangat menyederhanakan perhitungan dan interpretasi probabilitas.

3. Landasan untuk Inferensi Statistik

Standardisasi menjadi dasar bagi:

- Pengujian hipotesis
- Selang kepercayaan (*confidence intervals*)
- Nilai-p (*p-values*)

Metode-metode ini sangat bergantung pada skor-z.

Distribusi Normal Standar

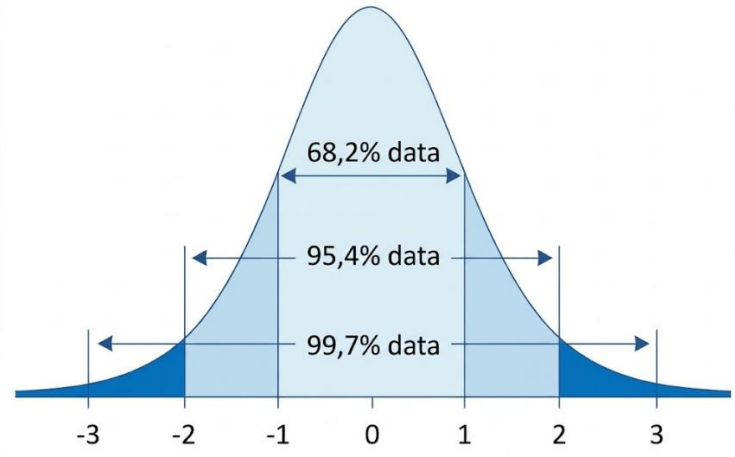
Distribusi normal standar adalah kasus khusus dari distribusi normal dengan:

- Rata-rata (*Mean*) = 0
- Simpangan baku (*Standard deviation*) = 1

Kita menotasikannya sebagai:

$$Z \sim N(0,1)$$

Di mana Z adalah variabel acak normal standar.



Gambar 10.10 Distribusi kerja standar

Probabilitas ditulis sebagai:

$$P(Z \leq z)$$

Setelah nilai distandarisasi, kita dapat menggunakan:

- Fungsi Excel
- Perangkat lunak statistik
- Tabel Z

untuk menghitung probabilitas secara efisien.

Menghitung Probabilitas Menggunakan Skor Z di Excel

1. Fungsi NORM.S.DIST()

Pada bagian sebelumnya tentang distribusi normal, kita menggunakan fungsi Excel

$$NORM.DIST(x, mean, standard_dev, cumulative)$$

untuk menghitung probabilitas untuk variabel yang terdistribusi normal. Untuk distribusi normal standar, logikanya sama, tetapi dengan satu perbedaan utama: data telah distandarisasi, sehingga rata-ratanya adalah 0 dan deviasi standarnya adalah 1. Karena itu, Excel menyediakan fungsi khusus untuk bekerja langsung dengan skor z:

$$NORM.S.DIST()$$

Tabel 10.2 Fungsi NORM.DIST()

Tugas	**Fungsi Excel**	**Interpretasi**
Probabilitas bahwa ($\boldsymbol{Z \leq z}$)	$= NORM.S.DIST(z, TRUE)$	Mengembalikan probabilitas kumulatif di sebelah kiri z
Tinggi kurva di z	$= NORM.S.DIST(z, FALSE)$	Mengembalikan kerapatan probabilitas (bukan probabilitas)

Fungsi ini bekerja seperti $NORM.DIST()$, namun tidak memerlukan nilai rata-rata (mean) atau simpangan baku (*standard deviation*) karena dalam distribusi normal standar, nilai-nilai tersebut ditetapkan masing-masing sebesar 0 dan 1.

Perhitungan Probabilitas Umum

- Ekor kiri (*Left-tail*): $P(Z \leq z)$
- Ekor kanan (*Right-tail*): $P(Z > z) = 1 - P(Z \leq z)$
- Di antara dua nilai (*Between*): $P(z1 < Z < z2) = P(Z \leq z2) - P(Z \leq z1)$

Contoh Soal

Misalkan nilai ujian berdistribusi normal dengan rata-rata 65 dan simpangan baku 15.

Contoh 1: Probabilitas Ekor Kiri (*Left-Tail*)

Tentukan probabilitas memperoleh nilai kurang dari 80. Nilai ujian:

- Rata-rata (*Mean*) = 65
- Simpangan baku (SD) = 15
- Dicari: $P(X \leq 80)$

Metode 1 (Nilai mentah/*Raw values*):

$= NORM.DIST(80, 65, 15, TRUE)$

$= 0{,}8413$

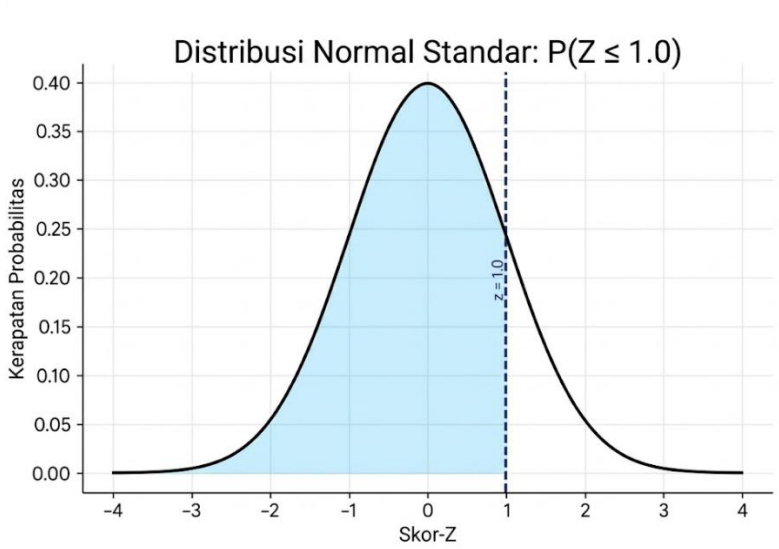
≈84% mahasiswa memperoleh nilai di bawah 80

Metode 2 (Skor-Z/Z-score):

- $Z = \frac{80-65}{15} = 1$
- $= NORM.S.DIST(1, TRUE)$
- = 0,8413

✓ Hasil sama — pendekatan berbeda.

Area yang diarsir di bawah kurva distribusi normal standar menunjukkan probabilitas saat z kurang dari atau sama dengan 1, yaitu $P(Z \leq 1)$.



Gambar 10.11 Luas di bawah kurva normal standar – probabilitas di sebelah kiri $P(Z \leq z)$

Tinjau kembali aturan empiris (*empirical rule*) dan jumlahkan probabilitas-probabilitasnya untuk melakukan verifikasi.

Contoh 2: Probabilitas di Antara Dua Nilai

Misalkan nilai ujian berdistribusi normal dengan rata-rata 65 dan simpangan baku 10. Tentukan probabilitas memperoleh nilai di antara 65 dan 85:

- Rata-rata (*Mean*) = 65
- Simpangan Baku (SD) = 10
- Cari: $P(65 < X < 85)$

Metode 1 (Nilai Mentah/Raw Values):

$= NORM.DIST(85, 65, 10, TRUE) - NORM.DIST(65, 65, 10, TRUE)$

$= 0{,}4772$

≈47,72% nilai berada di antara 65 dan 85.

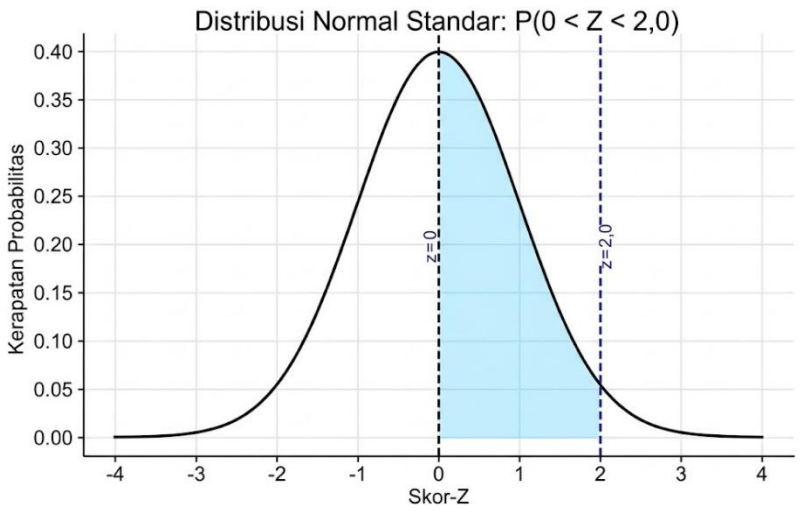
Metode 2 (Skor-Z):

$z_1 = \frac{65-65}{10} = 0, \quad z_2 = \frac{85-65}{10} = 2$

$= NORM.S.DIST(2, TRUE) - NORM.S.DIST(0, TRUE)$

$= 0{,}4772$

Interpretasi: ≈ 47,72% skor berada dalam rentang ini.



Gambar 10.12 Luas di bawah kurva normal standar - probabilitas antara $P(z < Z < z)$

Tinjau kembali aturan empiris (*empirical rule*) dan jumlahkan probabilitas-probabilitasnya untuk melakukan verifikasi.

Contoh 3: Probabilitas Ekor Kanan (*Right-Tail Probability*)

Melanjutkan contoh nilai ujian yang berdistribusi normal dengan rata-rata 65 dan simpangan baku 10. Tentukan probabilitas memperoleh nilai lebih dari 90.

- Rata-rata (Mean) = 65
- Simpangan Baku (SD) = 10
- Dicari: $P(X > 90) = 1 - P(X \leq 90)$

Metode 1 (Nilai Mentah/Raw Values):

$= 1 - NORM.DIST(90, 65, 10, TRUE)$

$= 0{,}00621$

≈0,62% memperoleh nilai di atas 90.

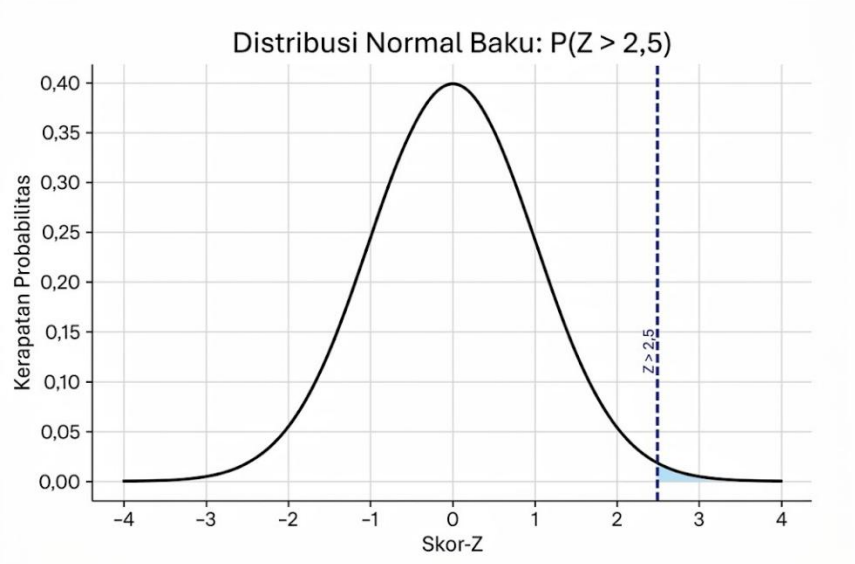
Metode 2 (Skor-Z):

$z = \frac{90-65}{10} = 2.5,$

$= 1 - NORM.S.DIST(2.5, TRUE)$

$= 0{,}00621$

Interpretasi: ≈ 0,62% skor berada di atas 90.



Gambar 10.13 Luas di bawah kurva normal standar - probabilitas di sisi kanan $P(Z > z)$

2. Fungsi $NORM.S.INV()$

Sama seperti $NORM.INV()$ yang menghasilkan nilai x (nilai data mentah) yang bersesuaian dengan probabilitas kumulatif tertentu, fungsi $NORM.S.INV()$ menghasilkan skor-z (*z-score*) yang bersesuaian dengan probabilitas ekor kiri (*left-tail probability*) tertentu.

$$= NORM.S.INV(probabilitas)$$

Misalkan nilai ujian berdistribusi normal dengan rata-rata 70 dan standar deviasi 10. Tentukan batas (persentil): "Batas nilai untuk 10% teratas".

Cari nilai x sedemikian sehingga $P(X \leq x) = 0{,}90$.

Metode 1 (Nilai Mentah):

$$= NORM.INV(0{,}90;\ 70;\ 10)$$
$$\approx 82{,}8$$

Metode 2 (Skor-z):

Pertama, cari kuantil Z:

$$= NORM.S.INV(0{,}90)$$
$$\approx 1{,}2816$$

Kemudian konversikan kembali:

$$x = \mu + z\sigma = 70 + 1{,}2816 \times 10 \approx 82{,}8$$

Hasil ini berarti seorang mahasiswa harus memperoleh nilai sekitar 82,8 atau lebih tinggi untuk masuk ke dalam kelompok 10% teratas dari seluruh peserta ujian.

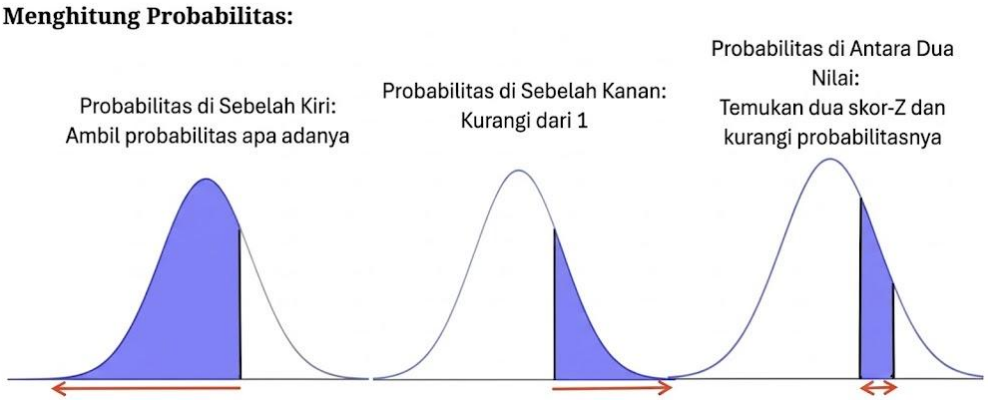
$NORM.S.INV()$ sangat berguna ketika Anda memerlukan:

- nilai kritis untuk pengujian hipotesis

- batas-batas selang kepercayaan (*confidence interval*)
- batas persentil pada data yang telah distandarisasi

Poin utama

- ❖ Standarisasi mengubah distribusi normal apa pun ke dalam skala universal
- ❖ Skor-z mengukur jarak dari rata-rata dalam satuan standar deviasi
- ❖ Distribusi normal standar menyederhanakan perhitungan probabilitas
- ❖ Excel menyediakan perangkat yang efisien ($NORM.S.DIST$, $NORM.S.INV$)
- ❖ Probabilitas bersesuaian dengan luas area di bawah kurva:
  - Ekor kiri (*Left-tail*) → dari $-\infty$ hingga z
  - Ekor kanan (*Right-tail*) → dari z hingga $+\infty$
  - Di antara (*Between*) → area di antara dua nilai z



Gambar 10.14 Panduan untuk menghitung probabilitas

10.6 MENGHUBUNGKAN PROBABILITAS DENGAN DATA

Bagaimana Probabilitas Mendorong Inferensi, Prediksi, dan Pengambilan Keputusan

Hingga saat ini, kita telah mempelajari probabilitas sebagai kerangka kerja matematis untuk memahami ketidakpastian. Namun dalam sains data, probabilitas bukan sekadar teori melainkan mesin praktis yang menjadi landasan bagi cara kita belajar dari data, membuat prediksi, dan menentukan arah pengambilan keputusan. Bagian ini merangkum semuanya dengan menunjukkan bagaimana probabilitas mendasari tugas-tugas utama dalam sains data.

1. Inferensi Statistik: Mempelajari Hal yang Belum Diketahui

Dalam situasi nyata, kita jarang memiliki akses ke seluruh populasi. Sebaliknya, kita bekerja dengan sampel dan menggunakan probabilitas untuk menarik kesimpulan yang berdasar mengenai populasi yang lebih luas. Dengan probabilitas, kita dapat:

- ➢ **Menarik kesimpulan dari data sampel**: Mengestimasi parameter populasi seperti rata-rata, proporsi, dan hubungan antarvariabel.
- ➢ **Mengukur ketidakpastian**: Menyatakan tingkat keyakinan kita terhadap estimasi yang dibuat dengan menggunakan nilai probabilitas.
- ➢ **Membuat klaim berbasis bukti**: Menentukan apakah pola yang teramati bermakna atau sekadar terjadi karena kebetulan.

Pada intinya, probabilitas memungkinkan kita beralih dari data yang kita miliki menuju wawasan yang dapat dipercaya.

2. Prediksi: Mengestimasi Apa yang Akan Terjadi Selanjutnya

Prediksi merupakan inti dari sains data. Baik untuk meramalkan penjualan, mendeteksi penipuan, maupun merekomendasikan produk, probabilitas menyediakan cara terstruktur untuk memodelkan ketidakpastian mengenai peristiwa di masa depan. Dengan menggunakan probabilitas, kita dapat:

- **Mengestimasi kemungkinan hasil di masa depan**: Contoh: "Terdapat peluang 70% bahwa permintaan akan meningkat bulan depan."
- **Membangun model probabilistik**: Banyak model (seperti regresi logistik) menghasilkan nilai probabilitas, bukan sekadar jawaban ya/tidak yang sederhana.
- **Mengukur ketidakpastian prediksi**: Alih-alih memberikan satu angka tunggal, kita dapat menyajikan rentang kemungkinan hasil beserta probabilitas yang menyertainya.

Prediksi yang baik tidak hanya akurat, tetapi juga jujur mengenai ketidakpastian yang ada.

3. Pengambilan Keputusan: Bertindak di Tengah Ketidakpastian

Pada akhirnya, sains data hadir untuk mendukung pengambilan keputusan yang lebih baik. Probabilitas membantu mengubah wawasan menjadi tindakan dengan memperhitungkan risiko dan ketidakpastian secara eksplisit. Dengan probabilitas, kita dapat:

- **Membandingkan berbagai tindakan alternatif**: Mengevaluasi opsi mana yang memiliki manfaat harapan (*expected benefit*) tertinggi.
- **Mengoptimalkan hasil**: Mengambil keputusan yang memaksimalkan keberhasilan atau meminimalkan biaya/risiko.
- **Mengelola risiko secara efektif**: Memahami dan bersiap menghadapi peristiwa yang mungkin jarang terjadi namun memiliki dampak besar. Dalam praktiknya, probabilitas mengubah data menjadi strategi yang dapat ditindaklanjuti.

Menyatukan Semuanya

Dalam inferensi, prediksi, dan pengambilan keputusan, probabilitas memainkan peran pemersatu:

- Memungkinkan kita untuk mengekstrak wawasan bermakna dari data (inferensi).
- Memungkinkan kita untuk mengantisipasi hasil di masa depan (prediksi).
- Memungkinkan kita untuk memilih dengan bijak (pengambilan keputusan). Secara keseluruhan, hal-hal ini membentuk fondasi sains data modern.

10.7 DISTRIBUSI PROBABILITAS DAN NORMAL, SERTA PENERAPAN ANALISIS DATA

Dalam bab ini, kita telah mempelajari bagaimana distribusi probabilitas membantu kita memodelkan, memahami, dan menafsirkan keacakan dalam data dunia nyata. Alih-alih memandang ketidakpastian sebagai sesuatu yang tidak teratur, kita belajar bahwa keacakan sering kali mengikuti pola yang dapat diprediksi, yang dapat dijelaskan secara matematis dan dianalisis secara praktis. Kita memulainya dengan memperkenalkan konsep distribusi probabilitas, sebuah kerangka kerja yang menetapkan nilai probabilitas untuk setiap kemungkinan hasil dari suatu variabel acak. Hal ini memungkinkan kita untuk beralih dari

probabilitas yang terpisah-pisah menuju gambaran menyeluruh tentang bagaimana nilai-nilai tersebut terdistribusi.

Selanjutnya, kita membedakan dua jenis distribusi utama:

- Distribusi diskret, yang menggambarkan hasil yang dapat dihitung (dicacah).
- Distribusi kontinu, yang menggambarkan pengukuran di sepanjang suatu rentang.

Dalam kategori distribusi diskret, kita membahas:

- Distribusi Binomial, yang digunakan untuk memodelkan jumlah keberhasilan dalam sejumlah percobaan independen yang telah ditentukan.
- Distribusi Poisson, yang digunakan untuk memodelkan jumlah kejadian yang berlangsung dalam suatu interval waktu atau ruang tertentu.

Distribusi-distribusi ini menyediakan perangkat praktis untuk menganalisis berbagai skenario dunia nyata, seperti kedatangan pelanggan, jumlah cacat produk, dan hasil yang bersifat biner (dua kemungkinan). Kemudian, kita beralih ke distribusi kontinu dengan berfokus pada model yang paling penting dan paling luas digunakan dalam statistika:

- Distribusi Normal, yang dicirikan oleh kurva simetris berbentuk lonceng.

Kita mengeksplorasi karakteristik utamanya, yaitu:

- Simetri di sekitar nilai rata-rata (*mean*).
- Kesamaan antara nilai rata-rata (*mean*), median, dan modus.
- Peran nilai rata-rata (μ) dan simpangan baku (σ) dalam menentukan bentuk serta sebaran kurva.
- Aturan Empiris (68–95–99,7), yang memberikan tolok ukur probabilitas secara intuitif.

Kita juga memperkenalkan konsep Fungsi Kepadatan Probabilitas (*Probability Density Function*/PDF), dengan menekankan bahwa probabilitas untuk variabel kontinu direpresentasikan sebagai luas area di bawah kurva, bukan sebagai titik-titik individual. Agar distribusi Normal dapat diterapkan secara praktis, kita menggunakan fungsi-fungsi dalam Excel, seperti:

- NORM.DIST() untuk menghitung probabilitas.
- NORM.INV () untuk mencari nilai yang bersesuaian dengan probabilitas tertentu.

Salah satu poin penting dalam bab ini adalah standardisasi, di mana kita mengubah nilai-nilai data menjadi skor-z (*z-score*). Hal ini memungkinkan kita untuk:

- Membandingkan nilai di berbagai distribusi yang berbeda
- Menyederhanakan perhitungan probabilitas
- Bekerja dengan Distribusi Normal Baku (*Standard Normal Distribution*) universal ($Z \sim N(0,1)$)

Kita telah mempelajari cara menghitung probabilitas menggunakan:

- NORM.S.DIST() untuk probabilitas yang telah dibakukan (distandarisasi)
- NORM.S.INV () untuk nilai kritis dan persentil

Terakhir, kita menghubungkan distribusi probabilitas dengan aspek inti dalam sains data:

- Inferensi Statistik: menarik kesimpulan mengenai populasi berdasarkan sampel
- Prediksi: mengestimasi hasil di masa depan dengan tingkat ketidakpastian yang terukur

- Pengambilan Keputusan: menggunakan probabilitas sebagai panduan dalam bertindak dan mengelola risiko

Menguasai distribusi probabilitas merupakan langkah krusial untuk mulai berpikir layaknya seorang praktisi sains data. Hal ini membekali Anda dengan kemampuan untuk tidak sekadar mendeskripsikan data, tetapi juga menafsirkan ketidakpastian, membuat prediksi yang berdasar, serta mendukung pengambilan keputusan yang tepat. Pada bab berikutnya, kita akan mengembangkan pemahaman dari landasan ini dengan beralih ke inferensi statistik di mana kita akan merumuskan cara menarik kesimpulan dari data menggunakan prinsip-prinsip yang telah Anda pelajari di sini.

BAB 11
TEORI PENGAMBILAN SAMPEL DAN ESTIMASI

11.1 TEORI PENGAMBILAN SAMPEL DAN ESTIMASI DALAM INFERENSI STATISTIK

Dalam dunia nyata, praktisi data (*data scientist*) jarang sekali memiliki kemewahan berupa informasi yang lengkap. Sebaliknya, mereka diminta untuk mengambil keputusan, menarik kesimpulan, dan memberikan wawasan berdasarkan data yang terbatas. Sebuah rumah sakit mungkin mempelajari beberapa ratus pasien untuk memahami penyakit yang menjangkiti jutaan orang. Sebuah perusahaan mungkin menganalisis sampel transaksi pelanggan untuk menentukan strategi bagi seluruh organisasi. Pemerintah mungkin menyurvei sebagian penduduknya sebagai dasar untuk merumuskan kebijakan nasional. Hal ini memunculkan pertanyaan mendasar: Bagaimana kita bisa dengan yakin mempelajari keseluruhan populasi hanya dari sebuah sampel? Bab ini akan menjawab pertanyaan tersebut.

Teori Pengambilan Sampel dan Estimasi merupakan tulang punggung statistik inferensial, cabang statistika yang memungkinkan kita untuk melangkah lebih jauh dari sekadar data yang diamati, guna menarik kesimpulan yang berdasar mengenai dunia yang lebih luas. Jika bab-bab sebelumnya berfokus pada pendeskripsian dan pemahaman data, kini kita melangkah ke tahap krusial berikutnya: menggunakan data untuk melakukan generalisasi, membuat prediksi, dan mengambil keputusan di tengah ketidakpastian. Inti dari bab ini mencakup dua gagasan utama:

- Pengambilan Sampel (*Sampling*) — proses memilih sebagian data (*subset*) dari populasi yang lebih besar.
- Estimasi — proses menggunakan sampel tersebut untuk menyimpulkan parameter populasi yang tidak diketahui.

Gagasan-gagasan ini sangat ampuh, namun juga memerlukan kehati-hatian. Pengambilan sampel yang buruk dapat menghasilkan kesimpulan yang bias. Salah menafsirkan estimasi dapat menyebabkan rasa percaya diri yang berlebihan dan pengambilan keputusan yang keliru. Sebagai seorang praktisi data, tanggung jawab Anda bukan sekadar menganalisis data, melainkan memastikan bahwa kesimpulan yang Anda hasilkan valid, dapat diandalkan, dan dapat dipertanggungjawabkan. Dalam bab ini, Anda akan mengembangkan pemahaman mendalam mengenai:

- Perbedaan antara populasi dan sampel, serta mengapa hal ini penting
- Berbagai metode pengambilan sampel dan pengaruhnya terhadap hasil
- Bias pengambilan sampel dan cara mengevaluasi seberapa representatif suatu sampel
- Ukuran sampel yang tepat berdasarkan kebutuhan akurasi dan kendala praktis
- Konsep distribusi sampling—mekanisme di balik inferensi statistik
- Teorema Limit Pusat (*Central Limit Theorem*) dan perannya dalam memungkinkan inferensi statistik
- Distribusi yang digunakan dalam inferensi, serta kapan harus menggunakan statistik z atau statistik t

- ❖ Cara menyusun dan menafsirkan estimasi titik (*point estimates*) dan selang kepercayaan (*confidence intervals*)

Lebih penting lagi, Anda akan mulai berpikir layaknya seorang praktisi statistik sejati—seseorang yang tidak begitu saja menerima data apa adanya, melainkan mengajukan pertanyaan seperti:

- Apakah sampel ini representatif?
- Seberapa besar ketidakpastian dalam estimasi ini?
- Seberapa yakin kita terhadap kesimpulan ini?

Pergeseran pola pikir ini sangatlah penting. Sebagaimana ditekankan di sepanjang buku ini, statistik bukan sekadar kumpulan rumus; statistik adalah cara bernalar mengenai ketidakpastian dan bukti.

Di akhir bab ini, Anda akan memperoleh kemampuan baru yang ampuh: kemampuan untuk menarik kesimpulan bermakna dari data yang terbatas, sembari memahami dan mengomunikasikan ketidakpastian yang menyertainya. Inilah keterampilan yang mengubah data mentah menjadi wawasan tepercaya dan merupakan salah satu kemampuan yang menjadi ciri khas seorang ilmuwan data (*data scientist*). Mari kita mulai.

11.2 POPULASI, SAMPEL, DAN BIAS DALAM INFERENSI STATISTIK

Memahami Perbedaannya

Dalam statistika, salah satu konsep paling mendasar adalah perbedaan antara populasi dan sampel. Populasi merujuk pada keseluruhan kelompok individu, objek, atau pengamatan yang ingin kita pelajari. Ini bisa mencakup seluruh pemilih terdaftar di suatu negara, semua pohon di dalam hutan, atau setiap produk yang dihasilkan oleh sebuah perusahaan dalam satu tahun. Pada intinya, populasi mewakili keseluruhan kumpulan data yang relevan dengan pertanyaan tertentu.

Dalam kondisi ideal, kita akan mengumpulkan data dari setiap anggota populasi. Proses ini dikenal sebagai sensus atau pencacahan lengkap. Meskipun sensus memberikan informasi yang paling akurat dan lengkap, cara ini sering kali tidak praktis karena keterbatasan biaya, waktu, dan kelayakan terutama ketika berhadapan dengan populasi yang sangat besar atau tak terhingga. Sebaliknya, sampel adalah bagian dari populasi yang dipilih untuk dianalisis. Alih-alih mempelajari seluruh populasi, kita menganalisis sampel yang dipilih secara cermat dan menggunakannya untuk menarik kesimpulan mengenai populasi secara keseluruhan. Pendekatan inilah yang menjadi inti dari statistika inferensial.

Parameter vs. Statistik

Untuk membedakan secara jelas antara populasi dan sampel, para ahli statistik menggunakan terminologi khusus: Parameter adalah karakteristik numerik dari suatu populasi. Karena populasi sering kali terlalu besar untuk diamati secara menyeluruh, parameter biasanya tidak diketahui nilainya dan harus diestimasi. Parameter populasi yang umum meliputi:

- ➢ Rata-rata populasi: μ
- ➢ Simpangan baku populasi: σ

- Proporsi populasi: p

Statistik adalah karakteristik numerik yang dihitung dari suatu sampel. Statistik merupakan nilai yang diketahui karena dihitung dari data yang diamati, dan nilai ini digunakan untuk mengestimasi parameter populasi. Statistik sampel yang umum meliputi:

- Rata-rata sampel: $\bar{x}$
- Simpangan baku sampel: s
- Proporsi sampel: $\hat{p}$

Tujuan utama dari statistika inferensial sederhana namun sangat bermanfaat: menggunakan statistik sampel untuk mengestimasi parameter populasi yang tidak diketahui.

Contoh: Misalkan rata-rata tinggi badan pria dewasa yang sebenarnya di suatu negara (rata-rata populasi μ) tidak diketahui. Kita mengambil sampel sebanyak 2000 pria dan menghitung rata-rata tinggi badannya, yaitu $\bar{x}$ = 174,5 cm. Statistik sampel ini berfungsi sebagai estimasi kita terhadap parameter populasi.

Mengapa Kita Menggunakan Sampel?

Pengambilan sampel bukan sekadar untuk kemudahan sering kali, hal ini merupakan suatu keharusan. Keputusan untuk menggunakan sampel alih-alih melakukan sensus menyeluruh didasari oleh beberapa faktor utama:

- **Biaya**: Mengumpulkan data dari seluruh populasi bisa sangat mahal. Pengambilan sampel secara signifikan mengurangi biaya tersebut.
- **Waktu**: Sensus bisa memakan waktu begitu lama sehingga hasilnya menjadi tidak relevan lagi saat selesai. Pengambilan sampel memungkinkan pengambilan keputusan yang lebih cepat.
- **Keterlaksanaan**: Beberapa populasi tidak mungkin diamati secara keseluruhan. Sebagai contoh, menguji masa pakai setiap bola lampu akan menghabiskan seluruh persediaan yang ada.
- **Akurasi**: Sampel yang dirancang dengan baik terkadang dapat menghasilkan hasil yang lebih dapat diandalkan dibandingkan sensus yang dilaksanakan dengan buruk, terutama jika pengendalian kualitasnya lebih ketat.

Penerapan di Dunia Nyata

Pengambilan sampel digunakan di hampir semua bidang:

- **Jajak Pendapat Politik**: Survei terhadap sebagian pemilih untuk memprediksi hasil pemilihan umum.
- **Riset Pasar**: Mempelajari preferensi konsumen dengan menggunakan sampel pelanggan.
- **Pengendalian Kualitas**: Memeriksa sampel produk untuk menilai kualitas manufaktur.
- **Uji Coba Medis**: Menguji pengobatan pada sampel pasien sebelum digunakan secara lebih luas.

Contoh-contoh ini mengilustrasikan gagasan penting: Kita tidak memerlukan semua data untuk membuat keputusan yang baik, kita hanya memerlukan data yang tepat.

Konsep Keterwakilan (*Representativeness*)

Efektivitas pengambilan sampel bergantung pada satu prinsip krusial: keterwakilan. Sampel yang representatif mencerminkan karakteristik utama populasi secara akurat. Sampel ini pada dasarnya merupakan versi miniatur dari populasi tersebut, yang mempertahankan struktur dan keragamannya. Sebagai contoh:

- Jika suatu populasi terdiri dari 60% perempuan dan 40% laki-laki, maka sampel harus mencerminkan distribusi yang serupa.
- Jika populasi mencakup semua kelompok usia, maka sampel tidak boleh didominasi oleh satu kelompok usia saja.

Ketika sampel bersifat representatif, kita dapat menggeneralisasi temuan dengan penuh keyakinan. Sebaliknya, jika tidak representatif, kesimpulan yang kita tarik bisa jadi menyesatkan.

Kesalahan Pengambilan Sampel (*Sampling Error*) vs. Bias Pengambilan Sampel (*Sampling Bias*)

Bahkan sampel yang dirancang dengan baik pun tidak sempurna. Dua konsep penting membantu kita memahami keterbatasannya:

Kesalahan Pengambilan Sampel (*Sampling Error*)

Kesalahan pengambilan sampel adalah perbedaan alami antara statistik sampel dan parameter populasi yang sebenarnya. Hal ini terjadi semata-mata karena kita hanya mengamati sebagian dari populasi. Sifatnya:

- ❖ Tidak dapat dihindari
- ❖ Bersifat acak
- ❖ Dapat diukur

Jika kita mengambil beberapa sampel, hasilnya akan bervariasi namun hasil tersebut cenderung mengelompok di sekitar nilai populasi yang sebenarnya.

Bias Pengambilan Sampel (Sampling Bias)

Bias pengambilan sampel terjadi ketika metode pengambilan sampel kita secara sistematis melebih-lebihkan (*over-represent*) atau kurang mewakili (*under-represent*) kelompok tertentu dalam populasi. Berbeda dengan kesalahan pengambilan sampel, bias merupakan cacat dalam metodologi yang menyebabkan estimasi menjadi salah secara konsisten. Jenis-jenis bias pengambilan sampel yang umum meliputi:

- **Bias Seleksi (*Selection Bias*)**: Terjadi ketika metode pengambilan sampel mengecualikan atau kurang mewakili kelompok tertentu.
  - Contoh: Melakukan survei melalui telepon mengenai penggunaan internet, namun hanya menelepon ke nomor telepon rumah (hal ini mengecualikan banyak anak muda yang hanya menggunakan ponsel).
- **Bias Respons (*Response Bias*)**: Terjadi ketika kelompok orang tertentu lebih cenderung berpartisipasi dibandingkan kelompok lainnya.
  - Contoh: Survei kepuasan pelanggan di mana hanya pelanggan yang sangat puas atau sangat tidak puas yang bersedia memberikan tanggapan.

- **Bias Kelangsungan Hidup (*Survivorship Bias*)**: Terjadi ketika kita hanya mengamati kasus-kasus yang berhasil dan mengabaikan kegagalan.
  - Contoh: Mempelajari bisnis yang sukses untuk memahami faktor keberhasilan perusahaan, tanpa mempertimbangkan bisnis yang gagal.
- **Bias Sukarelawan (*Volunteer Bias*)**: Terjadi ketika partisipan memilih sendiri untuk terlibat dalam pembelajaran.
  - Contoh: Studi kesehatan di mana hanya orang-orang yang peduli terhadap kesehatan yang secara sukarela berpartisipasi.

Bias menyebabkan estimasi yang secara konsisten tidak akurat, sehingga jauh lebih bermasalah dibandingkan kesalahan acak (*random error*).

Strategi untuk Meminimalkan Bias

Meminimalkan bias sangat penting untuk menghasilkan temuan pembelajaran yang valid, reliabel, dan dapat digeneralisasi. Strategi-strategi berikut merupakan praktik terbaik yang diakui secara luas untuk mengurangi kesalahan sistematis di sepanjang proses pembelajaran:

1. **Perancangan Studi yang Cermat**: Susun rencana pembelajaran yang jelas dan ketat, yang mencakup populasi target, kerangka sampel, dan prosedur pengumpulan data sebelum proses pengumpulan data dimulai.
2. **Penggunaan Pengambilan Sampel Probabilitas (*Probability Sampling*)**: Jika memungkinkan, gunakan metode pengambilan sampel berbasis probabilitas seperti pengambilan sampel acak sederhana, berstrata, sistematis, atau klister untuk mendorong pemilihan secara acak dan meningkatkan keterwakilan sampel.
3. **Instrumen Pengukuran yang Jelas dan Tidak Bias**: Buat butir pertanyaan survei, panduan wawancara, dan alat ukur yang bersifat netral, tepat, serta bebas dari bahasa yang mengarahkan (*leading*) atau ambigu.
4. **Pelatihan Komprehensif**: Berikan pelatihan menyeluruh kepada semua pengumpul data untuk memastikan penerapan prosedur yang konsisten serta mengurangi bias yang mungkin timbul dari pihak pewawancara.
5. **Uji Coba (*Pilot Testing*)**: Lakukan studi percontohan untuk mengidentifikasi potensi masalah pada prosedur pengambilan sampel, perumusan pertanyaan, atau protokol pengumpulan data, lalu perbaiki rancangan studi berdasarkan temuan tersebut.
6. **Validasi dan Pembersihan Data**: Terapkan pemeriksaan sistematis untuk mendeteksi kesalahan, ketidakkonsistenan, nilai yang hilang (*missing values*), dan pencilan (*outliers*) guna memastikan kualitas data sebelum dianalisis.
7. **Transparansi dalam Pelaporan**: Dokumentasikan dan laporkan secara jelas metode pengambilan sampel, tingkat respons (*response rates*), serta sumber bias yang diketahui atau mungkin ada, sehingga pembaca dapat menilai kredibilitas temuan tersebut secara kritis. Dengan menangani sumber-sumber bias ini secara proaktif, para peneliti memperkuat ketelitian studi mereka serta meningkatkan kredibilitas inferensi statistik yang ditarik dari data mereka.

Latihan Excel: Mendemonstrasikan Kesalahan Pengambilan Sampel (*Sampling Error*)

Mari kita gunakan Excel untuk menyimulasikan kesalahan pengambilan sampel dan melihat bagaimana hal tersebut terjadi dalam praktik. Gunakan kumpulan data `sampling_error` yang tersedia dalam buku kerja Excel bab ini untuk mengikuti dan menyelesaikan latihan tersebut.

1. Hitung rata-rata populasi yang sebenarnya
   - Masukkan rumus di sel C1: `$= AVERAGE(A1:A1000)$`
2. Ambil beberapa sampel
   - Di kolom D hingga M, buat 10 sampel berbeda dengan ukuran masing-masing 50
   - Di sel D1, masukkan: `$=INDEX(A:A, RANDBETWEEN(1,1000))$`
   - Salin rumus ini ke bawah hingga sel D50 dan ke samping hingga kolom M50
   - Hitung rata-rata sampel untuk setiap sampel di baris 52: `$=AVERAGE(D1:D50)$`
3. Amati kesalahan pengambilan sampel
   - Bandingkan setiap rata-rata sampel dengan rata-rata populasi
   - Perhatikan bagaimana nilai-nilai tersebut bervariasi di sekitar nilai yang sebenarnya
   - Tekan F9 untuk menghitung ulang dan lihat bagaimana rata-rata sampel berubah

Hal ini menunjukkan bahwa kesalahan pengambilan sampel adalah sesuatu yang wajar dan bersifat acak terkadang rata-rata sampel kita lebih tinggi daripada rata-rata populasi, terkadang lebih rendah, namun secara rata-rata, nilai-nilai tersebut berpusat di sekitar nilai yang sebenarnya.

11.3 METODE PENGAMBILAN SAMPEL DAN IMPLIKASINYA

Kualitas kesimpulan statistik sangat bergantung pada bagaimana sampel dipilih. Analisis secanggih apa pun tidak dapat memperbaiki sampel yang dipilih dengan buruk. Metode pengambilan sampel yang dirancang dengan baik akan menghasilkan sampel yang representatif, sehingga kita dapat menggeneralisasi temuan dengan penuh keyakinan. Sebaliknya, metode yang dirancang dengan buruk akan menimbulkan bias, yang berujung pada kesimpulan yang menyesatkan atau tidak valid. Metode pengambilan sampel secara garis besar dibagi menjadi dua kategori:

- Pengambilan Sampel Probabilitas (*Probability Sampling*) — setiap anggota populasi memiliki peluang yang diketahui (lebih dari nol) untuk terpilih
- Pengambilan Sampel Non-Probabilitas (*Non-Probability Sampling*) — pemilihan didasarkan pada kemudahan atau penilaian subjektif, dengan peluang terpilih yang tidak diketahui

Metode Pengambilan Sampel Probabilitas

Pengambilan sampel probabilitas memastikan bahwa setiap unit dalam populasi memiliki peluang yang diketahui untuk terpilih. Karakteristik ini memungkinkan kita untuk:

- Mengukur kesalahan pengambilan sampel (*sampling error*)
- Menerapkan statistik inferensial
- Menarik kesimpulan yang dapat diandalkan mengenai populasi

Oleh karena itu, metode pengambilan sampel probabilitas dianggap sebagai standar emas (*gold standard*) dalam analisis statistik.

1. Pengambilan Sampel Acak Sederhana (*Simple Random Sampling*/SRS)

Pengambilan Sampel Acak Sederhana adalah metode pengambilan sampel yang paling mendasar. Setiap anggota populasi memiliki peluang yang sama dan independen untuk terpilih.

Cara Kerja

1. Buat daftar lengkap populasi (kerangka sampel)
2. Berikan pengenal (*identifier*) unik kepada setiap anggota
3. Gunakan proses acak untuk memilih sampel

Kelebihan

- ✓ Menghilangkan bias pemilihan (*selection bias*)
- ✓ Mudah dipahami dan diterapkan
- ✓ Menjadi landasan teoretis bagi inferensi statistik

Keterbatasan

- o Memerlukan daftar populasi yang lengkap
- o Dapat menyebabkan kurang terwakilinya subkelompok yang kecil namun penting
- o Bisa memakan biaya besar untuk populasi yang berjumlah banyak atau tersebar luas

Latihan Excel: Pengambilan Sampel Acak Sederhana

Skenario: Anda memiliki daftar 200 karyawan dan ingin menyurvei 30 di antaranya. Gunakan dataset `simple_random_sampling` yang tersedia dalam buku kerja Excel bab ini untuk mengikuti langkah-langkah dan menyelesaikan latihan tersebut.

Langkah-langkah:

1. Kolom A berisi ID karyawan (Karyawan 1, Karyawan 2, dst.)
2. Pada sel B2, buat angka acak menggunakan rumus: `$= RAND()$`
3. Salin rumus ini ke bawah untuk seluruh 200 baris
4. Urutkan (*sort*) dataset berdasarkan kolom angka acak tersebut
5. Pilih 30 baris pertama sebagai sampel Anda

Wawasan Utama: Setiap kali Anda melakukan kalkulasi ulang, sampel acak yang berbeda akan dihasilkan menunjukkan keacakan yang sesungguhnya.

2. Pengambilan Sampel Bertingkat

Pengambilan sampel bertingkat membagi populasi menjadi subkelompok (strata) berdasarkan karakteristik penting, kemudian mengambil sampel acak dari setiap stratum. Metode ini meningkatkan keterwakilan dengan memastikan bahwa subkelompok (strata) penting terwakili dalam sampel.

Cara Kerjanya

1. Bagi populasi menjadi strata homogen
2. Lakukan pengambilan sampel acak di dalam setiap strata
3. Gabungkan semua unit yang dipilih menjadi satu sampel

Jenis Pengambilan Sampel Bertingkat

- ❖ Proporsional: Ukuran sampel mencerminkan proporsi populasi

- ❖ Alokasi Sama: Jumlah yang sama dipilih dari setiap strata
- ❖ Alokasi Optimal: Berdasarkan variabilitas dan biaya

Keuntungan

- ✓ Memastikan representasi semua kelompok kunci
- ✓ Seringkali menghasilkan perkiraan yang lebih tepat
- ✓ Memungkinkan analisis subkelompok

Keterbatasan

- o Membutuhkan pengetahuan sebelumnya tentang struktur populasi
- o Lebih kompleks untuk diimplementasikan
- o Membutuhkan penanganan terpisah untuk setiap kelompok

Latihan Excel: Pengambilan Sampel Bertingkat

Skenario: Survei kepuasan kerja di berbagai departemen di sebuah perusahaan. Pilih sampel sebanyak 60 menggunakan alokasi proporsional. Gunakan dataset stratified_sampling untuk mengikuti dan menyelesaikan latihan ini.

Langkah 1: Perusahaan memiliki 300 karyawan di 3 departemen:

- Penjualan: 150 karyawan
- Teknik: 100 karyawan
- SDM: 50 karyawan

Langkah 2: Buat sampel bertingkat proporsional (ukuran sampel total = 60)

1. Sampel Penjualan: (150/300) × 60 = 30 karyawan
2. Sampel Teknik: (100/300) × 60 = 20 karyawan
3. Sampel SDM: (50/300) × 60 = 10 karyawan

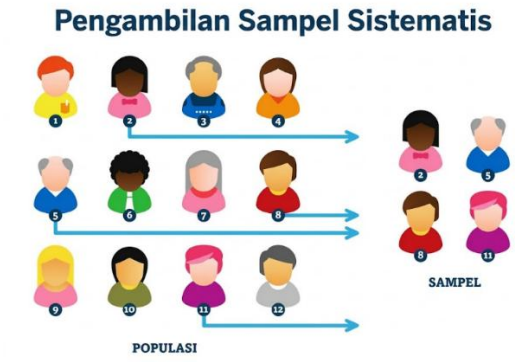
Langkah 3: Implementasikan di Excel

1. Saring data Anda berdasarkan departemen
2. Gunakan pengambilan sampel acak di setiap departemen
3. Gabungkan hasilnya

Wawasan Utama: Stratifikasi memastikan representasi yang seimbang, mengurangi risiko kehilangan subkelompok penting.

3. Pengambilan Sampel Sistematis

Pengambilan sampel sistematis memilih setiap anggota kth dari populasi setelah titik awal acak. Misalnya, mengambil setiap orang ketiga dari suatu barisan.



Gambar 11.1 Pengambilan Sampel Sistematis

Cara Kerjanya

1. Hitung interval pengambilan sampel:
   $k = \frac{N}{n}$ (N = ukuran populasi, n = ukuran sampel yang diinginkan).
2. Pilih titik awal secara acak antara 1 dan k.
3. Pilih setiap elemen ke-k setelahnya.

Kelebihan

- ✓ Sederhana dan efisien.
- ✓ Memastikan cakupan populasi yang merata.
- ✓ Mudah diterapkan pada kumpulan data yang sudah terurut.

Kekurangan

- o Dapat menimbulkan bias jika terdapat pola tersembunyi.
- o Kurang acak dibandingkan dengan pengambilan sampel acak sederhana (*simple random sampling*).

Latihan Excel: Pengambilan Sampel Sistematis (*Systematic Sampling*)

Contoh: Gunakan kumpulan data *systematic_sampling* untuk mengikuti dan menyelesaikan latihan ini. Kolom A berisi ID Karyawan dan Kolom B berisi gaji.

Hitung Interval Pengambilan Sampel (k):

- ❖ Populasi: 200 karyawan (N=200)
- ❖ Ukuran sampel yang diinginkan: 50 (n=50)
- ❖ Interval pengambilan sampel: k = 200/50 = 4

Langkah 1: Tambahkan Kolom Indeks

- Tambahkan judul kolom baru di sel C1 dengan nama "Indeks".
- Ketik 1 di sel C2 dan 2 di sel C3.
- Sorot kedua sel tersebut, lalu klik ganda pada *fill handle* (pegangan pengisi) untuk memberi nomor pada baris dari 1 hingga 200.

Langkah 2: Tentukan Parameter

- Cari ruang kosong di lembar kerja Anda (misalnya, kolom E dan F) untuk menetapkan aturan.
- Sel E1 (Titik Awal): Ketik =$RANDBETWEEN(1,4)$ Misalkan hasilnya adalah 2 untuk contoh ini.
- Sel F1 (Interval k): Ketik 4.
- Kunci Titik Awal:

  Salin sel E1, lalu tempel sebagai nilai (*Paste Values* — Ctrl + Alt + V > *Values*) kembali ke sel E1.

  Langkah ini mencegah titik awal berubah setiap kali Anda menekan tombol *Enter*.

Langkah 3: Buat Rumus "Pemilih" (*Selector*)

Kita memerlukan rumus yang menandai orang ke-2, lalu orang ke-6 (2+4), kemudian orang ke-10 (6+4), dan seterusnya. Kita menggunakan logika berikut:

Jika (Indeks dikurangi Titik Awal) habis dibagi 4, maka pilih data tersebut.

1. Buat judul kolom baru di sel D1 dengan nama "Keep?".

2. Pada baris data pertama (D2), masukkan rumus: $= IF(MOD(C2 - E1, F1) = 0$, "YES", "NO")
3. Tarik atau klik ganda untuk menerapkan rumus ini ke seluruh 200 baris.

Langkah 4: Memfilter dan Mengekstrak

- Klik pada baris judul.
- Buka menu Data > Filter.
- Klik tanda panah tarik-turun (dropdown) pada kolom "*Keep*?".
- Hapus centang pada "*NO*" dan pilih "*YES*" saja.

Pengambilan sampel sistematis (*systematic sampling*) memberikan struktur ke dalam proses yang acak—metode ini efisien, namun harus digunakan dengan hati-hati.

4. Pengambilan Sampel Klaster (*Cluster Sampling*)

Pengambilan sampel klaster digunakan ketika populasi berukuran besar dan tersebar secara geografis. Alih-alih mengambil sampel individu, kita mengambil sampel kelompok (klaster).

Cara Kerja

1. Bagi populasi ke dalam klaster-klaster (contoh: wilayah, sekolah, kota).
2. Pilih sebagian klaster secara acak.
3. Teliti semua anggota di dalam klaster yang terpilih (atau ambil sampel dari dalam klaster tersebut).

Kapan Digunakan

- Populasi tersebar secara geografis.
- Tidak tersedia daftar populasi yang lengkap.
- Terdapat kendala biaya yang signifikan.

Kelebihan

- ✓ Hemat biaya.
- ✓ Praktis secara logistik.
- ✓ Mengurangi upaya pengumpulan data.

Kekurangan

- Kurang presisi dibandingkan metode lain.
- Risiko bias lebih tinggi jika klaster-klaster memiliki perbedaan yang signifikan.
- Memerlukan analisis yang lebih kompleks.

Wawasan Bagian: Memilih Metode yang Tepat

Setiap metode pengambilan sampel melibatkan pertimbangan antara:

- ➢ Akurasi
- ➢ Biaya
- ➢ Kompleksitas
- ➢ Kelayakan (*feasibility*)

Tidak ada pendekatan yang "satu ukuran untuk semua" (*one-size-fits-all*). Seorang ilmuwan data yang terampil akan memilih metode yang paling sesuai dengan tujuan pembelajaran, kendala data, dan kondisi di lapangan.

Metode Pengambilan Sampel Non-Probabilitas

Berbeda dengan pengambilan sampel probabilitas, metode pengambilan sampel non-probabilitas tidak bergantung pada pemilihan acak. Akibatnya, tidak setiap anggota populasi memiliki peluang yang diketahui atau setara untuk disertakan dalam sampel. Hal ini memiliki implikasi penting:

- Kesalahan pengambilan sampel (*sampling error*) tidak dapat dihitung.
- Inferensi statistik tidak dapat diterapkan secara formal.
- Temuan tidak dapat digeneralisasikan secara meyakinkan ke populasi yang lebih luas.

Meskipun memiliki keterbatasan tersebut, pengambilan sampel non-probabilitas memainkan peran penting dalam praktik; khususnya dalam pembelajaran eksploratif, studi kualitatif, dan situasi di mana pengambilan sampel probabilitas tidak dapat dilakukan. Kuncinya bukan menghindari metode-metode ini, melainkan menggunakannya dengan tepat dan menafsirkan hasilnya secara hati-hati.

1. Pengambilan Sampel Berdasarkan Kemudahan (*Convenience Sampling*)

Konsep

Pengambilan sampel berdasarkan kemudahan melibatkan pemilihan individu yang mudah diakses oleh peneliti.

Cara Kerja

Peserta direkrut dari sumber yang tersedia dengan mudah, seperti:

- Mahamahasiswa di dalam kelas
- Pengunjung di pusat perbelanjaan (mal)
- Pengguna di platform daring

Kelebihan

- ✓ Cepat dan murah
- ✓ Mudah diterapkan
- ✓ Berguna untuk studi percontohan (*pilot study*) dan eksplorasi awal

Keterbatasan

- o Sangat rentan terhadap bias pemilihan (*selection bias*)
- o Sering kali menghasilkan sampel yang tidak representatif
- o Hasil tidak dapat digeneralisasikan

Contoh: Melakukan survei hanya kepada mahamahasiswa universitas mengenai preferensi politik nasional kemungkinan besar akan menghasilkan kesimpulan yang bias.

2. Pengambilan Sampel Bertujuan (*Purposive Sampling/Judgmental Sampling*)

Konsep

Pengambilan sampel bertujuan melibatkan pemilihan peserta secara sengaja berdasarkan karakteristik, keahlian, atau relevansi tertentu terhadap pertanyaan pembelajaran.

Cara Kerja

Peneliti mengidentifikasi dan memilih secara khusus individu atau kelompok yang memiliki:

- Pengetahuan tertentu,

- Pengalaman, atau
- Atribut yang relevan dengan studi tersebut.

Kelebihan

- ✓ Ideal untuk pembelajaran kualitatif mendalam
- ✓ Memungkinkan fokus pada kasus yang kaya informasi
- ✓ Berguna untuk studi kasus dan analisis pakar

Keterbatasan

- o Risiko tinggi terhadap bias peneliti
- o Proses pemilihan yang bersifat subjektif
- o Generalisasi yang terbatas

Contoh: Contohnya termasuk mewawancarai para pakar di industri tertentu atau memilih kasus yang sangat kaya akan informasi. Sebagai contoh: Mewawancarai ilmuwan data (data scientist) yang berpengalaman untuk memahami tren industri dapat memberikan wawasan mendalam namun tidak mewakili seluruh profesional di bidang tersebut.

3. Pengambilan Sampel Kuota (*Quota Sampling*)

Konsep

Pengambilan sampel kuota memastikan bahwa subkelompok tertentu terwakili dalam sampel, namun tanpa pemilihan secara acak di dalam subkelompok tersebut.

Cara Kerja

1. Identifikasi karakteristik utama (misalnya, usia, jenis kelamin, wilayah)
2. Tentukan proporsi dalam populasi
3. Tetapkan kuota (jumlah peserta yang akan diambil sebagai sampel) untuk setiap subkelompok
4. Kumpulkan data menggunakan metode non-acak hingga kuota terpenuhi

Kelebihan

- ✓ Memastikan keterwakilan subkelompok-subkelompok penting
- ✓ Lebih terstruktur dibandingkan pengambilan sampel kemudahan (*convenience sampling*) sederhana
- ✓ Lebih cepat dan lebih murah dibandingkan pengambilan sampel acak berstrata (*stratified random sampling*)

Keterbatasan

- o Masih rentan terhadap bias pemilihan di dalam kelompok
- o Tidak menjamin keterwakilan yang benar-benar akurat
- o Tidak cocok untuk inferensi statistik

Pengambilan sampel kuota mungkin tampak seperti pengambilan sampel berstrata tetapi karena tidak melibatkan unsur acak, metode ini tidak memberikan keandalan statistik yang setara.

Penggunaan yang Tepat dan Keterbatasan Utama Pengambilan Sampel Non-Probabilitas

Kapan Pengambilan Sampel Non-Probabilitas Tepat Digunakan?

Pengambilan sampel non-probabilitas paling berguna ketika:

- Mengeksplorasi masalah baru atau menyusun hipotesis

- Melakukan pembelajaran kualitatif yang berfokus pada kedalaman, bukan generalisasi
- Bekerja dengan keterbatasan waktu, biaya, atau akses terhadap data
- Mempelajari populasi yang langka, bersifat khusus, atau sulit dijangkau

Dalam konteks ini, tujuannya bukanlah untuk melakukan generalisasi, melainkan untuk memperoleh wawasan, arah, atau pemahaman.

Keterbatasan Penting yang Perlu Diperhatikan

Meskipun bermanfaat, metode ini memiliki keterbatasan yang signifikan:

- Tidak Dapat Digeneralisasi: Hasil hanya berlaku untuk individu yang dijadikan sampel
- Risiko Bias yang Tinggi: Proses pemilihan dapat mendistorsi temuan secara sistematis
- Tidak Ada Ukuran Ketidakpastian: Tanpa unsur acak, kita tidak dapat menghitung margin kesalahan (*margin of error*) atau interval kepercayaan (*confidence interval*)

Perspektif Seorang Ilmuwan Data (*Data Scientist*)

Pengambilan sampel *non-probabilitas* bukanlah sesuatu yang "salah" tetapi harus digunakan dengan tujuan yang jelas dan interpretasi yang disiplin.

- Gunakan untuk mengeksplorasi, bukan untuk mengonfirmasi
- Gunakan untuk menghasilkan gagasan, bukan untuk membuat klaim pada tingkat populasi
- Gunakan saat diperlukan namun akui keterbatasannya secara transparan

Kapan pun tujuannya adalah inferensi statistik, pengambilan sampel probabilitas harus selalu menjadi pendekatan yang diutamakan.

11.4 PERTIMBANGAN MENGENAI UKURAN SAMPEL

Salah satu pertanyaan paling penting dalam setiap studi adalah:

"Seberapa besar ukuran sampel yang seharusnya?"

Tidak ada jawaban tunggal yang berlaku universal. Ukuran sampel yang diperlukan bergantung pada tingkat akurasi yang Anda butuhkan, karakteristik populasi, serta sumber daya yang tersedia. Dalam praktiknya, penentuan ukuran sampel merupakan keseimbangan antara ketepatan statistik dan kendala di dunia nyata.

Faktor Utama yang Memengaruhi Ukuran Sampel

1. **Presisi yang Diinginkan (Margin Kesalahan/*Margin of Error*)**

   Presisi mengacu pada seberapa dekat estimasi sampel Anda dengan nilai populasi yang sebenarnya.

   - Margin kesalahan yang lebih kecil memerlukan ukuran sampel yang lebih besar
   - Margin kesalahan yang lebih besar memungkinkan ukuran sampel yang lebih kecil

   Wawasan: Jika Anda menginginkan hasil yang sangat akurat, Anda harus mengumpulkan lebih banyak data.

2. **Variabilitas Populasi**

   Variabilitas menggambarkan seberapa besar perbedaan nilai data di dalam suatu populasi.

- Variabilitas rendah → sampel yang dibutuhkan lebih kecil
- Variabilitas tinggi → sampel yang dibutuhkan lebih besar

Wawasan: Semakin beragam data yang ada, semakin banyak observasi yang diperlukan untuk menangkapnya secara akurat.

3. **Tingkat Kepercayaan (*Confidence Level*)**

   Tingkat kepercayaan mencerminkan seberapa yakin Anda terhadap estimasi yang dibuat.

   - Pilihan umum: 90%, 95%, 99%
   - Tingkat kepercayaan lebih tinggi → ukuran sampel lebih besar

   Wawasan: Kepastian yang lebih tinggi menuntut jumlah data yang lebih banyak.

4. **Ukuran Populasi**

   Untuk populasi yang besar, total ukuran populasi hanya memberikan dampak minimal terhadap ukuran sampel yang diperlukan.

   Wawasan: Baik populasi Anda berjumlah 100.000 maupun 10 juta, ukuran sampel yang dibutuhkan bisa jadi hampir sama.

Pertimbangan Praktis (*Trade-Off*)

Dalam praktik sains data di dunia nyata, penentuan ukuran sampel melibatkan keseimbangan antara tiga faktor yang saling bersaing:

1. **Biaya:** Sampel yang lebih besar membutuhkan biaya lebih tinggi
    1. Biaya survei
    2. Waktu pengumpulan data
    3. Kompleksitas pemeriksaan
2. **Waktu:** Sampel yang lebih besar memerlukan waktu lebih lama untuk dikumpulkan
    1. Durasi kerja lapangan
    2. Waktu pemrosesan data
    3. Penundaan dalam pengambilan keputusan
3. **Akurasi:** Sampel yang lebih besar memberikan estimasi yang lebih akurat
    1. Margin kesalahan (*margin of error*) yang lebih kecil
    2. Interval kepercayaan (*confidence interval*) yang lebih presisi
    3. Representasi subkelompok yang lebih baik

Prinsip Utama: Selalu ada kompromi (*trade-off*) mencapai akurasi sempurna jarang sekali bisa dilakukan secara praktis.

Menentukan Ukuran Sampel

Untuk beralih dari pemahaman konseptual ke perhitungan yang presisi, kita menggunakan rumus statistik yang sudah baku. Rumus yang tepat bergantung sepenuhnya pada apa yang ingin Anda estimasi apakah itu rata-rata populasi (*mean*) atau persentase (proporsi). Sebelum kita membahas rumus ukuran sampel secara mendetail, penting untuk memahami dua konsep inti berikut:

- **Margin kesalahan (*margin of error*)**: Besarnya kemungkinan perbedaan antara estimasi sampel Anda dengan nilai populasi yang sebenarnya.

- **Interval kepercayaan (*confidence interval*)**: Rentang nilai yang kemungkinan besar mencakup nilai populasi yang sebenarnya.

Kita akan mengeksplorasi kedua konsep ini secara mendalam pada bagian selanjutnya, dengan menyusunnya langkah demi langkah agar logikanya tetap jelas dan mudah dipahami.

1. Ukuran Sampel untuk Mengestimasi Rata-rata (*Mean*)

Saat mengestimasi rata-rata populasi, ukuran sampel yang diperlukan ditentukan dengan rumus berikut:

$$n = \left(\frac{z \times \sigma}{E}\right)^2$$

Di mana:

- Z = Z-skor yang bersesuaian dengan tingkat kepercayaan yang diinginkan
- σ = simpangan baku populasi
- E = margin kesalahan (*margin of error*) yang diinginkan

Contoh Pengerjaan

Hitung ukuran sampel yang diperlukan untuk mengestimasi rata-rata populasi dengan ketentuan:

- Tingkat kepercayaan = 95% → Z = 1,96
- Simpangan baku populasi = 15
- Margin kesalahan = 2

Perhitungan:

$$n = \left(\frac{1.96 \times 15}{2}\right)^2 = \left(\frac{29.4}{2}\right)^2 = (14.7)^2 = 216.09$$
$$n \approx 217$$

Interpretasi:

Anda memerlukan sampel dengan setidaknya 217 pengembangan teknologi masa depan untuk mengestimasi rata-rata populasi dengan margin kesalahan sebesar 2 pada tingkat kepercayaan 95%.

2. Ukuran Sampel untuk Mengestimasi Proporsi

Saat mengestimasi proporsi populasi (misalnya, persentase pelanggan yang lebih menyukai suatu produk), rumusnya adalah:

$$n = \frac{z^2 p(1-p)}{E^2}$$

Di mana:

- p = proporsi populasi estimasi
- E = margin kesalahan (*margin of error*)
- Z = skor-Z untuk tingkat kepercayaan yang diinginkan

Catatan Penting

Jika proporsi yang sebenarnya tidak diketahui, gunakan:

p = 0,5

Nilai ini menghasilkan ukuran sampel maksimum yang diperlukan, sehingga menjamin estimasi yang konservatif dan aman.

Bereksperimen dan Membangun Intuisi (Latihan Excel)

Untuk memperdalam pemahaman Anda:

- Ubah tingkat kepercayaan (contoh: 90%, 95%, 99%)
- Sesuaikan margin kesalahan (*margin of error*)
- Ubah standar deviasi atau proporsi

Amati bagaimana perubahan-perubahan tersebut memengaruhi ukuran sampel yang diperlukan.

Wawasan Utama: Ukuran sampel tidak ditentukan secara sembarangan melainkan didorong oleh kebutuhan akan presisi, tingkat ketidakpastian, dan tingkat kepercayaan tertentu.

Perspektif Seorang Data Scientist

Penentuan ukuran sampel bukan sekadar perhitungan teknis, melainkan sebuah keputusan strategis. Seorang data scientist yang andal akan bertanya:

- Seberapa presisi hasil yang saya butuhkan?
- Tingkat ketidakpastian seperti apa yang dapat diterima?
- Apakah saya memiliki sumber daya yang memadai untuk mencapai tingkat akurasi tersebut?

Pada akhirnya, tujuannya bukanlah mengumpulkan data sebanyak mungkin, melainkan mengumpulkan data yang cukup untuk membuat keputusan yang dapat diandalkan dan dapat dipertanggungjawabkan.

11.5 DISTRIBUSI SAMPLING

Pada bagian sebelumnya, kita telah membahas bagaimana sampel digunakan untuk mengestimasi karakteristik populasi. Kini muncul pertanyaan krusial:

Seberapa andalkah estimasi yang didasarkan pada sampel tersebut?

Untuk menjawabnya, kita akan memperkenalkan salah satu konsep terpenting dalam statistika: distribusi sampling.

Konsep dan Pentingnya

Distribusi sampling adalah distribusi probabilitas dari suatu statistik (seperti rata-rata atau proporsi) yang dihitung dari berbagai kemungkinan sampel dengan ukuran yang sama, yang diambil dari suatu populasi. Bayangkan proses berikut yang dilakukan berulang kali:

- Mengambil sampel berukuran n
- Menghitung suatu statistik (misalnya, rata-rata sampel)

- Mencatat hasilnya

Jika kita memplot semua nilai yang telah dihitung tersebut, kita akan memperoleh distribusi sampling.

Mengapa hal ini penting?

Karena distribusi sampling memungkinkan kita untuk:

- Memahami bagaimana statistik sampel bervariasi dari satu sampel ke sampel lainnya
- Mengukur ketidakpastian
- Membuat pernyataan probabilistik mengenai estimasi yang diperoleh

Pada dasarnya, distribusi sampling menjadi jembatan antara statistika deskriptif dan statistika inferensial.

Distribusi Sampling dari Rata-rata Sampel

Distribusi sampling yang paling penting adalah distribusi sampling dari rata-rata sampel ($\bar{x}$). Jika kita mengambil banyak sampel dan menghitung rata-ratanya, distribusi yang dihasilkan akan memiliki tiga karakteristik utama:

1. Rata-rata dari Distribusi Sampling

$$\mu_{\bar{x}} = \mu$$

Rata-rata dari distribusi sampling untuk rata-rata sampel sama dengan rata-rata populasi. Hal ini menunjukkan bahwa rata-rata sampel merupakan estimator tak bias bagi rata-rata populasi.

2. Simpangan Baku dari Distribusi Sampling

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Nilai ini dikenal sebagai galat baku (*standard error*/SE) dari rerata. Nilai ini:

- Mengukur seberapa besar variasi rerata sampel di sekitar rerata populasi
- Menurun seiring dengan bertambahnya ukuran sampel

Dengan demikian, sampel yang lebih besar menghasilkan estimasi yang lebih presisi karena rerata sampelnya mengelompok lebih rapat di sekitar rerata populasi yang sebenarnya.

Distribusi Sampling dari Proporsi Sampel

Untuk data kategorikal, kita sering berfokus pada proporsi sampel (p̂). Jika kita menghitung proporsi "kejadian sukses" untuk semua kemungkinan sampel berukuran n, distribusi sampling yang dihasilkan memiliki karakteristik sebagai berikut:

1. Rerata dari Distribusi Sampling

$$\mu_{\hat{p}} = p$$

Rerata dari distribusi sampling sama dengan proporsi populasi p, yang menunjukkan bahwa proporsi sampel merupakan estimator tak bias bagi proporsi populasi.

2. Galat Baku (*Standard Error*/SE) dari Proporsi Sampel

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Galat baku (*standard error*) ini mengukur variabilitas proporsi sampel di sekitar proporsi populasi yang sebenarnya. Seperti halnya pada rata-rata sampel, ukuran sampel yang lebih besar akan menghasilkan galat baku yang lebih kecil, yang berujung pada estimasi yang lebih presisi.

Teorema Limit Pusat (*Central Limit Theorem*/Clt)

Teorema Limit Pusat (*Central Limit Theorem*/CLT) bisa dibilang merupakan teorema paling penting dalam statistika. Teorema ini menjelaskan mengapa Distribusi Normal memegang peranan sentral dalam inferensi statistika bahkan ketika populasi asalnya tidak berdistribusi normal. CLT menyediakan landasan teoretis yang memungkinkan kita menggunakan data sampel untuk membuat pernyataan yang dapat diandalkan mengenai populasi.

Pernyataan: Rata-rata Sampel Mendekati Distribusi Normal seiring dengan Bertambahnya

CLT menyatakan bahwa untuk ukuran sampel yang cukup besar, distribusi sampling dari rata-rata sampel akan mendekati distribusi normal, terlepas dari bentuk distribusi populasinya, dengan syarat populasi tersebut memiliki varians yang terhingga.

Lebih lanjut:

- Rata-rata dari distribusi sampling sama dengan rata-rata populasi (μ).
- Simpangan baku dari distribusi sampling sama dengan galat baku (*standard error*): σ / √n.

Dengan kata lain: Meskipun populasi bersifat menceng (*skewed*) atau tidak beraturan, distribusi rata-rata sampel akan cenderung membentuk kurva berbentuk lonceng yang dapat diprediksi seiring dengan bertambahnya ukuran sampel.

Syarat Utama untuk CLT

Agar Teorema Limit Pusat (*Central Limit Theorem*/CLT) dapat diterapkan, beberapa syarat harus dipenuhi:

- **Pengambilan sampel acak**: Pengembangan teknologi masa depan harus dipilih secara acak.
- **Independensi**: Setiap pengembangan teknologi masa depan harus bersifat independen satu sama lain.
- **Ukuran sampel yang memadai**: Sampel yang lebih besar menghasilkan pendekatan yang lebih baik terhadap distribusi normal (pedoman umum yang sering digunakan adalah n ≥ 30).

Persyaratan Ukuran Sampel

Seberapa besar ukuran sampel yang diperlukan agar CLT dapat bekerja secara efektif? Meskipun konteks tetap berpengaruh, beberapa pedoman umum meliputi:

- n ≥ 30: Cocok untuk sebagian besar populasi.
- n ≥ 15: Sering kali sudah cukup jika populasi berdistribusi mendekati normal.
- n ≥ 100: Disarankan untuk populasi yang memiliki kemencengan (*skewness*) tinggi.
- Ukuran berapa pun: Jika populasi itu sendiri berdistribusi normal, maka distribusi sampling dari rata-rata akan berdistribusi normal untuk ukuran sampel (n) berapa pun.

Tiga Poin Utama CLT

1. Distribusi sampling dari rata-rata sampel akan mendekati distribusi normal seiring dengan bertambahnya ukuran sampel, terlepas dari bentuk distribusi populasinya.
2. Rata-rata dari distribusi sampling sama dengan rata-rata populasi: $\bar{\mu}x = \mu$.
3. Simpangan baku dari distribusi sampling (dikenal sebagai galat baku atau *standard error*) adalah: $\bar{\sigma}x = \sigma/\sqrt{n}$.

Mengapa CLT Penting

Teorema Limit Pusat merupakan landasan bagi inferensi statistik. Teorema ini memungkinkan kita untuk:

- Menyusun selang kepercayaan (*confidence intervals*).
- Melakukan pengujian hipotesis.
- Mengestimasi parameter populasi dengan tingkat ketidakpastian yang terukur.
- Mengambil keputusan berbasis probabilitas berdasarkan data sampel.

Tanpa CLT, kita harus mengetahui distribusi yang tepat dari setiap populasi yang diteliti—sebuah persyaratan yang tidak realistis dalam praktiknya.

11.6 DISTRIBUSI Z DAN T DALAM INFERENSI STATISTIK

Pada bagian sebelumnya, Teorema Limit Pusat (CLT) menetapkan sebuah hasil yang sangat penting: Terlepas dari bagaimana bentuk distribusi populasinya, distribusi sampling dari rata-rata sampel akan mendekati distribusi normal seiring dengan bertambahnya ukuran sampel. Hasil ini menjadi mesin teoretis bagi inferensi statistik. Namun, hal ini segera memunculkan pertanyaan praktis:

> *Jika distribusi sampel kira-kira berdistribusi normal, distribusi normal manakah yang harus kita gunakan dan bagaimana cara menggunakannya untuk membuat pernyataan mengenai parameter populasi yang tidak diketahui?*

Untuk menjawabnya, kita memerlukan tiga hal yang belum diperkenalkan secara formal dalam konteks inferensi:

1. Pemahaman yang jelas tentang bagaimana Distribusi Normal Standar (Z) beralih dari alat deskriptif menjadi alat inferensial khususnya, konsep nilai kritis (*critical values*).
2. Pemahaman yang tepat mengenai apa arti tingkat kepercayaan (*confidence levels*—90%, 95%, 99%) secara statistik.
3. Pengenalan terhadap distribusi-t (*t-distribution*) kerabat dekat distribusi normal yang menjadi sangat penting ketika standar deviasi populasi σ tidak diketahui (kondisi yang hampir selalu terjadi dalam praktik).

Konsep-konsep ini memberikan makna bagi segala hal yang akan dibahas selanjutnya margin kesalahan (*margin of error*), selang kepercayaan (*confidence intervals*), dan estimasi.

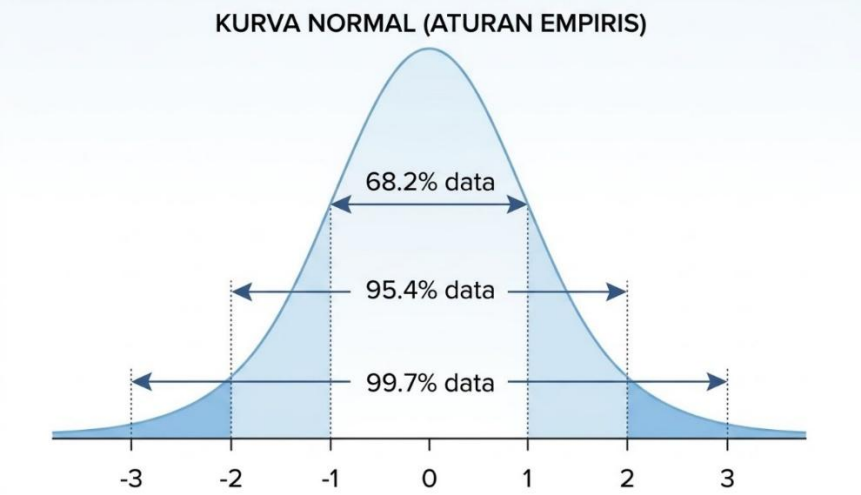
Dari Distribusi Normal Standar menuju Inferensi

Ringkasan Singkat

Pada Bab 10, Anda telah mempelajari bahwa variabel X apa pun yang berdistribusi normal dengan rata-rata μ dan standar deviasi σ dapat distandarisasi menggunakan rumus skor-Z (*Z-score*):

$$Z = \frac{X - \mu}{\sigma}$$

Transformasi ini mengubah distribusi normal apa pun menjadi Distribusi Normal Standar: suatu distribusi normal dengan nilai rata-rata 0 dan simpangan baku 1, yang ditulis sebagai $Z \sim N(0, 1)$.



Gambar 11.2 Distribusi Normal Standar

Anda juga telah mempelajari cara menggunakan distribusi Normal Standar untuk menemukan probabilitas luas di bawah kurva antara dua nilai Z. Itu adalah penggunaan deskriptif dari distribusi Normal Standar. Sekarang, dalam konteks inferensi, ia memainkan peran yang berbeda.

Peran Inferensial dari Z

Ingat kembali apa yang dikatakan Teorema Batas Pusat (CLT) tentang distribusi sampling dari rata-rata sampel $\bar{X}$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ (approximately, for large } n)$$

Ini berarti kita dapat menstandarisasi $\bar{X}$ sama seperti kita menstandarisasi variabel apa pun yang berdistribusi normal:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Z-score ini memberi tahu kita:

- Seberapa jauh rata-rata sampel dari rata-rata populasi yang sebenarnya
- Dalam satuan kesalahan baku (*standard error*)

Karena nilai Z ini mengikuti distribusi normal standar, kita sekarang dapat:

- Menghitung probabilitas terkait estimasi sampel
- Mengukur ketidakpastian
- Membangun selang kepercayaan (*confidence intervals*)

Pergeseran Utama

Sebelumnya:

> *"Berapa probabilitas untuk mengamati suatu nilai di atas skor tertentu?"*

Sekarang:

> *"Berapa probabilitas bahwa estimasi sampel kita berada dalam jarak tertentu dari nilai populasi yang sebenarnya?"*

Secara matematis, caranya sama namun interpretasinya kini bersifat inferensial.

Memperkenalkan Nilai Kritis (*Critical Values*)

Pada Bab 10, Anda menggunakan Z untuk mencari probabilitas yang berkaitan dengan nilai tertentu. Dalam inferensi, Anda sering bekerja dengan arah sebaliknya: Anda memulai dengan suatu probabilitas, lalu mencari nilai Z yang bersesuaian dengannya. Definisi — Nilai Kritis (Z*): Nilai kritis adalah Z-score yang menandai batas di mana probabilitas (luas area) tertentu terletak pada ekor distribusi Normal Standar. Pertimbangkan pertanyaan berikut: Di antara dua nilai Z berapakah 95% bagian tengah dari distribusi Normal Standar berada? Karena distribusi Normal Standar bersifat simetris terhadap nol, jawabannya adalah: 95% bagian tengah terletak di antara −Z* dan +Z*, dengan sisa 5% terbagi rata di kedua ekor distribusi—masing-masing 2,5% pada setiap ekor.

Contoh: Kepercayaan 95%

Kita menginginkan 95% bagian tengah dari distribusi normal standar.

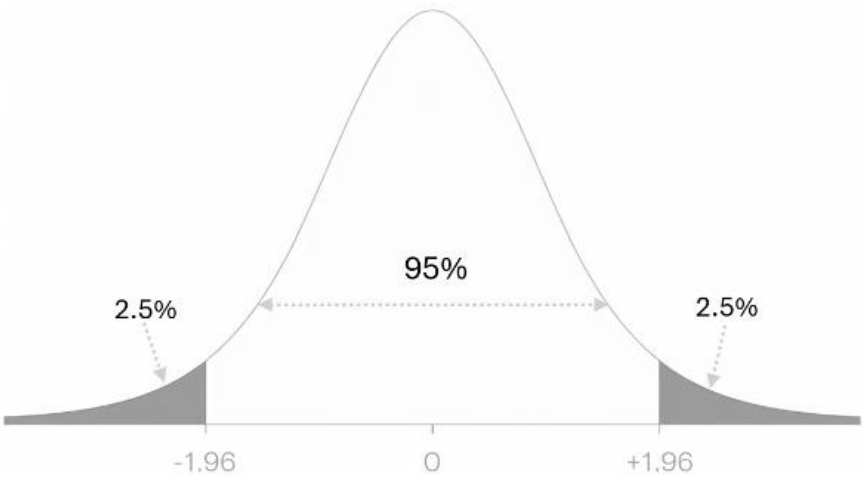
- Luas total di luar = α = 0,05
- Setiap ekor = $\alpha/2$ = 0,025

Jadi:

$$P(Z < Z*) = 0{,}975$$

Menggunakan Excel:

$$= NORM.S.INV(0{,}975)$$
$$Z* = \ 1{,}96$$



Gambar 11.3: Dua nilai kritis ekor (α = 0,05)

Jadi, 95% bagian tengah terletak di antara: −1,96 dan +1,96 Nilai-nilai ini, ±1,96, adalah nilai kritis untuk kepercayaan 95%. Tabel 11.1 merangkum nilai kritis untuk tiga tingkat kepercayaan yang paling sering digunakan:

Tabel 11.1: Nilai kritis untuk tiga tingkat kepercayaan yang paling sering digunakan

Tingkat Keyakinan	α (Tingkat Signifikansi)	α/2	Z* (Nilai Kritis)
90%	0,10	0,05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.576

Membaca tabel: Simbol α (alfa) melambangkan total probabilitas di bagian ekor distribusi — yaitu probabilitas berada di luar wilayah kepercayaan pusat. Karena interval kepercayaan bersifat dua sisi, α dibagi sama rata ke kedua sisi ekor, sehingga menghasilkan α/2 pada masing-masing ekor. Nilai kritis Z* adalah skor-Z yang membatasi tepat α/2 pada ekor bagian atas.

Excel: =NORM.S.INV(0,975) menghasilkan 1,96 (untuk tingkat kepercayaan 95%). Rumus umumnya adalah =NORM.S.INV(1 − α/2).

Konsep nilai kritis inilah yang mengubah pernyataan mengenai probabilitas menjadi batas-batas suatu interval kepercayaan. Anda akan sering menggunakan nilai-nilai ini dalam setiap rumus di bagian selanjutnya.

Tingkat Kepercayaan Dalam Estimasi

Sebelum menerapkan tingkat kepercayaan ke dalam rumus apa pun, sangat penting untuk memahami dengan tepat apa artinya dan yang tak kalah penting, apa yang bukan merupakan artinya.

Makna Formal Tingkat Kepercayaan

Tingkat kepercayaan adalah probabilitas, yang dinyatakan dalam persentase, bahwa prosedur yang digunakan untuk menyusun interval kepercayaan akan menghasilkan interval yang benar-benar memuat parameter populasi yang sesungguhnya jika prosedur tersebut diulangi untuk berbagai sampel yang berbeda. Ini adalah perbedaan yang halus namun krusial. Saat kita menyatakan sedang menyusun interval kepercayaan 95%, kita tidak bermaksud mengatakan:

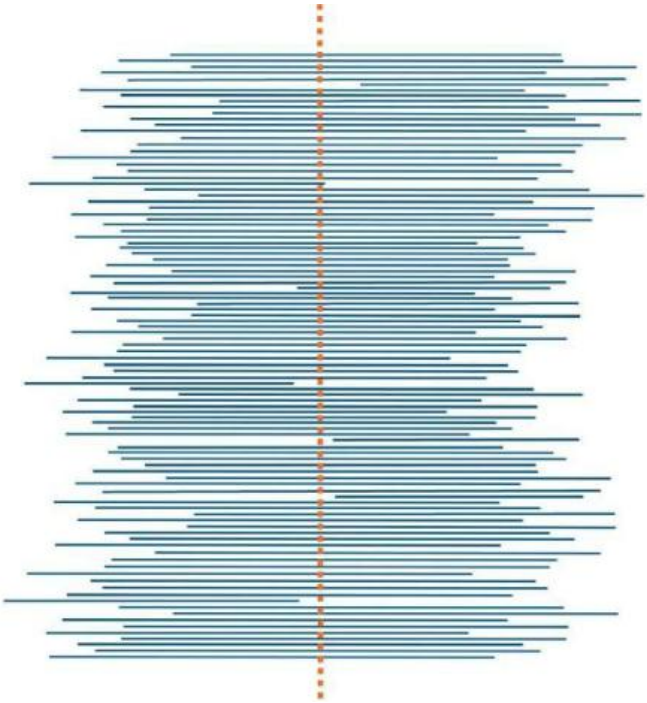
> *"Terdapat probabilitas 95% bahwa rata-rata populasi yang sesungguhnya berada di dalam interval spesifik ini."*

Rata-rata populasi μ adalah angka yang tetap (meskipun tidak diketahui). Nilai tersebut bisa berada di dalam interval atau tidak. Apa yang sebenarnya kita katakan adalah:

> *"Jika kita mengambil berbagai sampel acak yang berbeda dari populasi ini dan menyusun interval kepercayaan dari masing-masing sampel tersebut menggunakan metode yang sama, maka sekitar 95% dari interval-interval tersebut akan memuat rata-rata populasi yang sesungguhnya."*

Intuisi Visual

Bayangkan Anda mengambil 100 sampel acak yang berbeda dari populasi yang sama, lalu menghitung selang kepercayaan (*confidence interval*) 95% untuk setiap sampel tersebut. Setiap batang horizontal mewakili satu selang kepercayaan. Garis putus-putus vertikal menunjukkan nilai rata-rata populasi yang sebenarnya (namun tidak diketahui), yaitu μ. Jika prosedur ini dilakukan dengan benar, maka sekitar 95 dari 100 selang tersebut akan memotong garis putus-putus (artinya, selang tersebut mencakup μ), sementara sekitar 5 selang lainnya tidak mencakup μ sama sekali. Sebanyak 5% selang yang tidak mencakup μ ini berkaitan dengan $\alpha = 0{,}05$ yaitu probabilitas di mana selang tersebut gagal memuat parameter yang sebenarnya. Nilai ini sering disebut sebagai tingkat signifikansi (*significance level*). Gambar 11.4 mengilustrasikan hal ini:



Gambar 11.4 Visualisasi selang kepercayaan 95%

Hubungan Utama: Tingkat Kepercayaan (*Confidence Level*) = 1 − α

- Tingkat kepercayaan 90% → α = 0,10 → terdapat peluang 10% selang tidak mencakup μ
- Tingkat kepercayaan 95% → α = 0,05 → terdapat peluang 5% selang tidak mencakup μ
- Tingkat kepercayaan 99% → α = 0,01 → terdapat peluang 1% selang tidak mencakup μ

Pertimbangan dalam Memilih Tingkat Kepercayaan

Tingkat kepercayaan yang lebih tinggi akan menghasilkan selang yang lebih lebar. Secara intuitif, hal ini masuk akal: agar Anda lebih yakin bahwa selang yang Anda buat mencakup nilai yang sebenarnya, Anda harus memperlebar selang tersebut.

Tabel 11.2 Pertimbangan dalam memilih tingkat kepercayaan

Tingkat Kepercayaan	Lebar Interval	Presisi
90%	Paling sempit	Tertinggi (namun tingkat keyakinan paling rendah)
95%	Sedang	Sedang (paling sering digunakan)
99%	Paling lebar	Terendah (namun tingkat keyakinan paling tinggi)

Konvensi Praktis: Tingkat kepercayaan 95% adalah yang paling banyak digunakan dalam studi ilmu sosial, bisnis, dan kesehatan. Pemilihan tingkat kepercayaan harus mencerminkan konsekuensi (biaya) dari kesalahan dalam suatu penerapan tertentu. Uji coba medis sering kali menuntut tingkat kepercayaan 99%, sementara survei riset pasar sering menggunakan tingkat kepercayaan 90%.

Distribusi-T

Sejauh ini dalam bab ini, kita secara implisit mengasumsikan bahwa simpangan baku populasi, σ, telah diketahui. Dalam rumus berbasis Z yang diturunkan dari Teorema Limit Pusat

(*Central Limit Theorem*/CLT), σ muncul di bagian penyebut sebagai bagian dari galat baku (*standard error*):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Namun, dalam praktiknya, σ hampir tidak pernah diketahui. Dalam pemeriksaan data di dunia nyata, kita mengumpulkan sampel justru karena populasinya tidak diketahui termasuk variabilitasnya. Solusi yang wajar adalah mengganti σ dengan simpangan baku sampel, s. Akan tetapi, penggantian ini menimbulkan ketidakpastian tambahan: berbeda dengan σ, nilai s dapat bervariasi dari satu sampel ke sampel lainnya. Variabilitas ini sangat terasa terutama pada sampel berukuran kecil. Jika kita sekadar menyubstitusikan s ke dalam rumus berbasis distribusi Z, kita akan meremehkan ketidakpastian tambahan ini, yang mengakibatkan interval kepercayaan menjadi terlalu sempit dan terlalu optimis. Masalah ini dipecahkan pada tahun 1908 oleh William Sealy Gosset, yang menerbitkan karyanya dengan menggunakan nama samaran "*Student*." Solusi yang ia kembangkan dikenal sebagai distribusi-t *Student* (*Student's t-distribution*).

Apa Itu Distribusi-t?

Distribusi-t adalah sekumpulan distribusi probabilitas yang simetris dan berbentuk lonceng, menyerupai distribusi Normal Standar namun memiliki ekor yang lebih tebal; artinya, distribusi ini memberikan probabilitas lebih besar pada nilai-nilai yang jauh dari pusat. Probabilitas tambahan di bagian ekor ini mencerminkan ketidakpastian ekstra yang muncul akibat penggunaan s untuk mengestimasi σ. Distribusi-t dicirikan oleh satu parameter saja: derajat kebebasan (*degrees of freedom*/df). Derajat Kebebasan (df): Dalam konteks mengestimasi rata-rata populasi berdasarkan sampel berukuran n, derajat kebebasannya adalah:

$$\mathrm{df} = n - 1$$

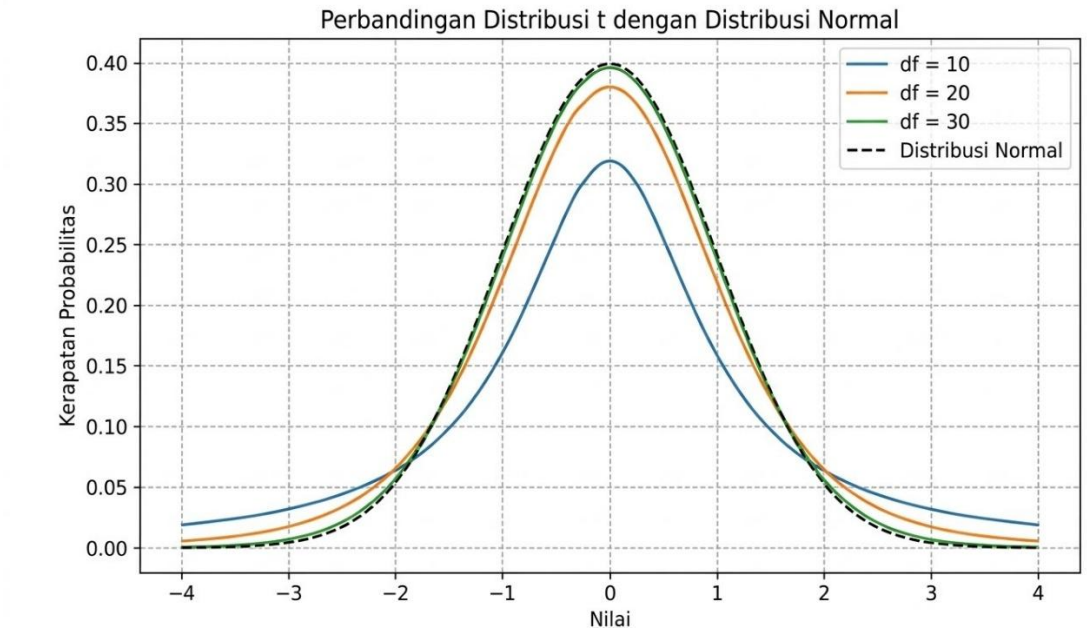
Secara intuitif, derajat kebebasan (*degrees of freedom*) merepresentasikan jumlah informasi independen yang tersedia setelah data digunakan untuk mengestimasi suatu parameter. Saat menghitung s, kita terlebih dahulu menghitung $\bar{X}$ dari data yang sama, yang kemudian "menggunakan" satu derajat kebebasan oleh karena itu menjadi $n - 1$.

Karakteristik Utama Distribusi-t

1. **Simetris dan berbentuk lonceng**
   Seperti distribusi Normal Standar, distribusi ini berpusat di angka nol dan bersifat unimodal (memiliki satu puncak).
2. **Memiliki ekor yang lebih tebal dibandingkan distribusi normal**
   Hal ini menghasilkan nilai kritis yang lebih besar dan selang kepercayaan (*confidence interval*) yang lebih lebar sehingga memperhitungkan ketidakpastian dengan lebih tepat.

3. **Konvergen menuju distribusi normal**
   Seiring dengan bertambahnya derajat kebebasan ($df \to \infty$), distribusi-t akan mendekati distribusi Normal Standar.
4. **Bentuk bergantung pada ukuran sampel**
   Sampel yang lebih kecil (derajat kebebasan lebih rendah) menghasilkan sebaran yang lebih lebar dan ekor yang lebih tebal; sampel yang lebih besar menghasilkan bentuk yang mendekati distribusi normal.
5. **Setiap nilai derajat kebebasan menentukan kurva yang berbeda:**
   Berbeda dengan keluarga distribusi normal (yang ditentukan oleh μ dan σ), setiap nilai derajat kebebasan menghasilkan distribusi-t yang unik.

Inilah sebabnya mengapa tabel distribusi-t disusun berdasarkan derajat kebebasan. Gambar di bawah ini mengilustrasikan hubungan antara distribusi-t dan distribusi Normal Standar:



Gambar 11.5 Comparisow antara distribusi-t dan distribusi kerja

Nilai-Nilai Kritis dan Implikasinya

Tabel 11.3: Derajat menyalakan dan nilai kritik mereka pada 95% CI

Distribusi	**df**	**Nilai Kritis (IK 95%)**
Normal Standar (Z)	∞	1,96
Distribusi-t	30	2,042
Distribusi-t	20	2,086
Distribusi-t	10	2,228
Distribusi-t	5	2,571
Distribusi-t	3	3,182

Perhatikan bagaimana nilai kritis t meningkat secara signifikan seiring dengan berkurangnya derajat kebebasan. Untuk sampel yang sangat kecil, distribusi-t memerlukan interval yang jauh lebih lebar guna mempertahankan tingkat kepercayaan yang sama.

Statistik-T dan Penerapannya

Ketika simpangan baku populasi σ tidak diketahui dan digantikan oleh simpangan baku sampel s, statistik uji yang telah distandarisasi tidak lagi berupa skor-Z, melainkan mengikuti distribusi-t:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Statistik ini, yang disebut sebagai statistik-t, mengikuti distribusi-t dengan $df = n - 1$, dengan asumsi bahwa populasi asal berdistribusi mendekati normal (atau n cukup besar sehingga Teorema Limit Pusat/CLT dapat diterapkan).

Nilai Kritis *t* dan Notasi

Nilai kritis dari distribusi-t ditulis sebagai berikut:

$$t^* \text{ or } t_{\alpha/2,\text{df}}$$

Contoh:

- t0.025,19 sesuai dengan interval kepercayaan 95% dengan n = 20

Mencari Nilai Kritis *t* di Excel

Excel menyediakan dua fungsi untuk bekerja dengan distribusi t:

Tabel 11.4 Menghitung nilai kritis t menggunakan Excel

Tujuan	Fungsi Excel	Contoh
Menentukan nilai kritis t*	=T.INV.2T(α, df)	=T.INV.2T(0,05; 19) → 2,093 untuk IK 95%, n=20
Menentukan nilai p (p-value) untuk statistik-t	=T.DIST.2T(t, df)	=T.DIST.2T(2,093; 19) → 0,05

Catatan Excel: T.INV.2T(α, df) menghasilkan nilai kritis dua arah (*two-tailed*) yaitu, nilai t* sedemikian rupa sehingga total luas di kedua ujung (ekor) distribusi sama dengan α. Fungsi inilah yang akan Anda gunakan untuk interval kepercayaan. Selalu pastikan derajat kebebasan (*degrees of freedom*) Anda sebelum menerapkan rumus tersebut.

Kapan Menggunakan Z vs. t

Aturan Keputusan

```
Apakah σ diketahui?

├── YA → Gunakan distribusi Z
│
└── TIDAK → Gunakan simpangan baku sampel (s)
        │
        ├── n ≥ 30 → Z dapat diterima, t lebih disukai
        │
        └── n < 30 → Gunakan distribusi t
```

Panduan Praktis: Dalam sebagian besar studi nyata, nilai σ tidak diketahui. Ini berarti distribusi-t merupakan alat standar untuk mengestimasi rata-rata populasi. Distribusi-Z digunakan dalam kasus yang relatif jarang, yaitu ketika nilai σ diketahui contohnya, saat menggunakan instrumen pengukuran yang sudah mapan dengan standar deviasi jangka panjang yang telah diketahui, atau saat bekerja dengan proporsi. Jika ragu, gunakan distribusi-t.

Distribusi-t untuk Proporsi?

Saat mengestimasi proporsi populasi bukan rata-rata distribusi sampling dari proporsi sampel ($\hat{p}$) akan mendekati distribusi normal (berdasarkan Teorema Limit Pusat/CLT, untuk sampel berukuran besar), dan galat baku (*standard error*) dihitung menggunakan nilai $\hat{p}$ itu sendiri, bukan menggunakan nilai σ yang tidak diketahui. Dalam kasus ini, distribusi-Z digunakan secara langsung; distribusi-t tidak berlaku untuk estimasi proporsi.

Merangkum Semuanya: Kerangka Kerja Pengambilan Keputusan

Setiap masalah estimasi dalam bab ini dapat diselesaikan dengan menjawab tiga pertanyaan kunci:

Tabel 11.5: Kerangka kerja pengambilan keputusan untuk masalah estimasi

Pertanyaan	**Menentukan**
Berapa tingkat kepercayaannya?	Nilai α dan nilai kritis
Apakah $\boldsymbol{\sigma}$ diketahui?	Gunakan Z atau t
Berapa ukuran sampel $\mathbf{n}$?	Derajat kebebasan df = n − 1

11.7 ESTIMASI TITIK DAN ESTIMASI INTERVAL

Setelah memahami distribusi sampling, kita kini dapat membahas tujuan utama dari statistika inferensial: Mengestimasi parameter populasi yang tidak diketahui menggunakan data sampel.

Terdapat dua pendekatan mendasar:

- Estimasi Titik (*Point Estimation*) → satu estimasi terbaik
- Estimasi Interval (*Interval Estimation*) → rentang nilai yang mungkin (masuk akal)

Estimasi Titik

Estimasi titik adalah satu nilai numerik yang dihitung dari sampel dan digunakan untuk mengestimasi parameter populasi yang tidak diketahui.

Contoh Umum

- Rata-rata sampel ($\bar{x}$) → mengestimasi rata-rata populasi (μ)
- Proporsi sampel ($\hat{p}$) → mengestimasi proporsi populasi (p)

Gagasan Utama: Estimasi titik memberikan satu tebakan terbaik, namun tidak menunjukkan seberapa akurat tebakan tersebut.

Estimator Tak Bias (*Unbiased Estimators*)

Sebuah estimator dikatakan tak bias jika nilai harapannya (*expected value*) sama dengan nilai sebenarnya dari parameter populasi yang ingin diestimasi.

- **Estimator**: Suatu statistik (seperti rata-rata sampel) yang digunakan untuk mendekati parameter populasi yang tidak diketahui (seperti rata-rata populasi).
- **Sifat Tak Bias (*Unbiasedness*)**: Sebuah estimator dikatakan tak bias jika nilai harapannya yang dirata-ratakan dari banyak sampel acak yang berulang sama dengan parameter populasi yang sebenarnya.

Contoh Estimator Tak Bias

- Rata-rata Sampel ($\bar{x}$): Rata-rata sampel merupakan estimator tak bias untuk rata-rata populasi (μ).

$$E(\bar{x}) = \mu$$

- Proporsi Sampel ($\hat{p}$): Proporsi sampel merupakan estimator tak bias untuk proporsi populasi (p).

$$E(\hat{p}) = p$$

Varians Sampel: Varians sampel yang dihitung dengan n sebagai penyebut merupakan estimator bias untuk varians populasi (σ^2). Namun, varians sampel yang dihitung dengan $n - 1$ sebagai penyebut rumus standar yang digunakan dalam statistika merupakan estimator tak bias untuk σ^2. Penyesuaian ini dikenal sebagai koreksi Bessel (*Bessel's correction*), yang meningkatkan akurasi estimasi.

Sifat-Sifat Estimator yang Baik

Selain sifat tak bias, para ahli statistika sering mencari sifat-sifat lain yang diinginkan:

- **Efisiensi**: Estimator yang efisien memiliki varians terkecil di antara semua estimator tak bias untuk parameter yang sama. Artinya, distribusi sampling-nya terkonsentrasi secara rapat di sekitar parameter yang sebenarnya, sehingga menghasilkan estimasi yang lebih presisi.
- **Konsistensi**: Estimator yang konsisten akan mendekati parameter populasi yang sebenarnya seiring dengan bertambahnya ukuran sampel. Baik rata-rata sampel maupun proporsi sampel merupakan estimator yang konsisten.

Menghitung Estimasi Titik di Excel

Dalam praktiknya, estimasi titik dapat dihitung dengan mudah menggunakan Excel:

- **Rata-rata sampel:**
  =AVERAGE(rentang)
- **Varians sampel:**
  =VAR.S(rentang)
- **Simpangan baku sampel:**
  =STDEV.S(rentang)
- **Proporsi sampel:**
  =COUNTIF(rentang,"kriteria") / COUNT(rentang) atau
  =COUNTIF(rentang,"kriteria") / COUNTA(rentang)

Tips: Jika data Anda menggunakan angka 1 (sukses) dan 0 (gagal), proporsi sampel dapat dihitung dengan menggunakan:

$$= AVERAGE(rentang)$$

Fungsi-fungsi ini memberikan estimasi titik untuk parameter populasi yang bersesuaian berdasarkan data sampel Anda.

Margin Of Error (MOE)

Estimasi titik (*point estimate*) saja tidaklah cukup. Dua sampel yang diambil dari populasi yang sama akan menghasilkan hasil yang sedikit berbeda. Hal ini memunculkan pertanyaan penting:

Seberapa jauh estimasi kita dari nilai yang sebenarnya?

Di sinilah margin of error menjadi sangat penting. Margin of error mengukur besarnya perbedaan yang mungkin terjadi antara estimasi sampel dan parameter populasi yang sebenarnya akibat pengambilan sampel secara acak.

Definisi

Margin of error adalah penyimpangan maksimum yang diharapkan dari suatu estimasi sampel terhadap nilai populasi yang sebenarnya pada tingkat kepercayaan tertentu. Dalam inferensi statistik, *margin of error* (MOE) mengukur seberapa besar estimasi berbasis sampel (seperti rata-rata atau proporsi) diperkirakan akan bervariasi dari parameter populasi yang sebenarnya akibat pengambilan sampel secara acak.

Interpretasi dalam Konteks Nyata

Jika sebuah laporan menyatakan:

"52% pemilih mendukung Kandidat A, dengan margin of error ±3%"

Ini berarti:

- Estimasi titiknya adalah 52%
- Nilai yang sebenarnya kemungkinan berada di antara 49% dan 55%

- Angka ±3% mencerminkan ketidakpastian dalam pengambilan sampel

Menghitung Margin of Error

Margin of error (MOE) dihitung menggunakan nilai kritis dari distribusi sampling dan kesalahan baku (*standard error*) dari estimator tersebut.

MOE = z* SE (untuk sampel besar dan σ yang diketahui)

MOE = t* SE (untuk sampel kecil dengan σ yang tidak diketahui)

Contoh:

- Rata-rata untuk sampel besar dengan σ yang diketahui (berbasis z)

$$\mathrm{MOE} = z \cdot \frac{\sigma}{\sqrt{n}}$$

- Rata-rata untuk sampel kecil dengan σ yang tidak diketahui (berbasis distribusi t)

$$\mathrm{MOE} = t \cdot \frac{s}{\sqrt{n}}$$

- Proporsi

$$\mathrm{MOE} = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Interpretasi:

Saat mengestimasi nilai rata-rata, margin kesalahan (*margin of error* atau MOE) menentukan seberapa jauh estimasi tersebut membentang di kedua sisi nilai estimasi.

- MOE yang lebih kecil → presisi yang lebih tinggi
- MOE yang lebih besar → ketidakpastian yang lebih besar

Faktor-faktor yang Memengaruhi Margin Kesalahan

1. Ukuran Sampel (n)

- Sampel yang lebih besar menghasilkan margin kesalahan yang lebih kecil.
- Hubungan ini mengikuti aturan akar kuadrat:

Untuk mengurangi margin kesalahan hingga setengahnya, Anda harus melipatgandakan ukuran sampel sebanyak empat kali lipat.

2. Tingkat Kepercayaan (*Confidence Level*)

- Tingkat kepercayaan yang lebih tinggi memerlukan nilai kritis yang lebih besar.
- Interval kepercayaan 99% lebih lebar daripada interval 95% karena interval tersebut harus mencakup parameter yang sebenarnya dengan tingkat keandalan yang lebih tinggi.

3. Variabilitas Populasi (σ atau p)

- Variabilitas yang lebih besar menyebabkan margin kesalahan yang lebih besar.

- Untuk proporsi, nilai yang mendekati 0,5 menghasilkan margin kesalahan terbesar, sedangkan proporsi yang mendekati 0 atau 1 menghasilkan margin kesalahan yang lebih kecil.

Perspektif Seorang Ilmuwan Data (*Data Scientist*)

Estimasi titik (*point estimate*) tanpa margin kesalahan adalah estimasi yang tidak lengkap. Sebagai seorang ilmuwan data, tanggung jawab Anda bukan sekadar memberikan estimasi, melainkan juga mengomunikasikan:

- Seberapa akurat estimasi tersebut
- Seberapa besar ketidakpastian yang ada
- Seberapa besar tingkat keyakinan yang seharusnya dimiliki pemangku kepentingan terhadap hasil tersebut

Pergeseran ini dari nilai tunggal menuju ketidakpastian yang terukur adalah hal yang mengubah data menjadi wawasan yang dapat diandalkan untuk pengambilan keputusan.

Interval Kepercayaan (*Confidence Intervals*)

Kita telah mengetahui bahwa meskipun estimasi titik memberikan satu tebakan terbaik, estimasi tersebut tidak mencakup ketidakpastian; hal ini dikarenakan sampel yang berbeda akan menghasilkan estimasi yang sedikit berbeda pula. Untuk mengatasi masalah ini, kita menggunakan interval kepercayaan (*confidence intervals* atau CI). Interval kepercayaan menyediakan rentang nilai yang masuk akal, di mana parameter populasi yang sebenarnya kemungkinan besar berada di dalamnya.

Bentuk Umum

Interval Kepercayaan = Estimasi Titik ± Margin Kesalahan

Struktur ini menyoroti dua komponen:

- Estimasi pusat (contoh: $\bar{x}$, $\hat{p}$)
- Ukuran ketidakpastian (margin kesalahan)

Memahami Tingkat Kepercayaan

Sebuah kesalahpahaman yang umum terjadi adalah:

“Interval kepercayaan 95% berarti terdapat peluang 95% bahwa parameter tersebut berada di dalam interval tersebut.”

Pemahaman ini keliru. Interpretasi yang Benar

- Parameter populasi bersifat tetap (konstan)
- Interval bersifat acak (bervariasi dari satu sampel ke sampel lainnya)

Jika kita mengulangi proses pengambilan sampel berkali-kali dan menyusun selang kepercayaan 95% setiap kalinya:

- Sekitar 95% dari selang-selang tersebut akan memuat parameter yang sebenarnya

Selang Kepercayaan untuk Rata-rata Populasi

Rumus yang digunakan bergantung pada apakah simpangan baku populasi (σ) diketahui atau tidak.

1. Saat σ Diketahui (Distribusi Z)

Jika simpangan baku populasi diketahui dan populasi berdistribusi normal atau ukuran sampel cukup besar (≥30) berdasarkan Teorema Limit Pusat maka selang kepercayaan untuk rata-rata populasi adalah

$$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Di mana:

- Z**α**/2 = nilai kritis dari distribusi normal standar

Digunakan ketika:

- Simpangan baku populasi diketahui
- Ukuran sampel besar (n ≥ 30) atau populasi berdistribusi mendekati normal

2. Ketika σ Tidak Diketahui (Distribusi-t)

Jika simpangan baku populasi tidak diketahui (skenario yang paling umum terjadi), kita mengganti σ dengan simpangan baku sampel s, lalu menggunakan distribusi-t Student. Distribusi ini memperhitungkan ketidakpastian tambahan yang muncul akibat estimasi variabilitas dari sampel.

$$\bar{x} \pm t_{\alpha/2,\mathrm{df}} \cdot \frac{s}{\sqrt{n}}$$

Di mana:

- t**α**/2, df = nilai kritis dari distribusi-t dengan derajat kebebasan (*degrees of freedom*) sebesar n-1
- $\frac{s}{\sqrt{n}} =$ estimasi kesalahan baku (*standard error*) dari rata-rata

Digunakan ketika:

- Simpangan baku populasi tidak diketahui (sebagian besar kasus nyata)
- Ukuran sampel kecil

Gunakan saat:

- Simpangan baku populasi tidak diketahui (sebagian besar kasus nyata)
- Ukuran sampel kecil

Wawasan Utama: Distribusi-t memperhitungkan ketidakpastian tambahan saat mengestimasi σ.

Selang Kepercayaan untuk Proporsi Populasi

Untuk proporsi populasi, distribusi sampling dari proporsi sampel bersifat mendekati normal jika ukuran sampel cukup besar (umumnya diperiksa menggunakan $n\hat{p} \geq 5$ dan $n(1 - \hat{p}) \geq 5$). Selang kepercayaannya adalah:

$$\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Di mana:

- $\hat{p}$ = proporsi sampel
- $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = estimasi galat baku (*standard error*) dari proporsi

Selang Kepercayaan di Excel

Excel menyediakan fungsi bawaan yang mempermudah perhitungan margin kesalahan (*margin of error*) dan selang kepercayaan, baik untuk nilai rata-rata maupun proporsi. Alat-alat ini meniadakan kebutuhan untuk memeriksa tabel statistik secara manual dan memungkinkan proses yang cepat dari data mentah hingga estimasi selang.

1. Selang Kepercayaan untuk Nilai Rata-rata (σ Diketahui atau Sampel Besar)

Jika simpangan baku populasi diketahui (atau jika ukuran sampel cukup besar untuk membenarkan penggunaan pendekatan distribusi normal), Excel menggunakan:

$$= CONFIDENCE.NORM(alpha, standard_dev, size)$$

Fungsi ini menghasilkan margin kesalahan $Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Contoh:

- **Pertanyaan Sains Data**: "Berapa rata-rata waktu sebenarnya (dalam menit) yang dibutuhkan oleh tim dukungan pelanggan kami untuk menyelesaikan tiket teknis di seluruh organisasi?"
- **Data**: Seorang ilmuwan data mengambil sampel besar (n = 100) dari tiket yang telah diselesaikan. Ia menghitung rata-rata sampel menggunakan fungsi =AVERAGE(*range*) dan memperoleh rata-rata waktu penyelesaian sebesar 45 menit. Berdasarkan data historis perusahaan, diketahui bahwa standar deviasi populasi (σ) adalah 15 menit.

1. Menghitung Margin Kesalahan (*Margin of Error*)

Untuk menentukan tingkat presisi dari rata-rata sampel, kita menggunakan fungsi CONFIDENCE.NORM dengan tingkat kepercayaan 95% (α = 0,05).

- Rumus Excel: =CONFIDENCE.NORM(0,05, 15, 100)
- Hasil (Margin Kesalahan): 2,94

2. Menyusun Interval Kepercayaan (*Confidence Interval*)

Dengan menggunakan rata-rata sampel (45), kita menerapkan margin kesalahan (2,94) untuk menentukan rentangnya:

- Interval Kepercayaan = Estimasi Titik ± Margin Kesalahan
- Batas Bawah: 45 - 2,94 = 42,06
- Batas Atas: 45 + 2,94 = 47,94

Interpretasi Hasil

"Kami memiliki tingkat kepercayaan 95% bahwa rata-rata waktu penyelesaian yang sebenarnya untuk semua tiket teknis di perusahaan berada di antara 42,06 dan 47,94 menit. Dengan kata lain, jika kita mengulangi proses pengambilan sampel ini sebanyak 100 kali, kita

memperkirakan 95 dari interval yang dihasilkan akan mencakup rata-rata populasi yang sebenarnya." Dukungan Keputusan: Baik tujuannya untuk memperbaiki waktu penyelesaian maupun menyesuaikan pola penempatan staf, informasi ini membantu proses pengambilan keputusan perusahaan.

2. Interval Kepercayaan untuk Rata-rata (σ Tidak Diketahui, Distribusi-t)

Contoh: Pemeriksaan Pengeluaran Pelanggan untuk Aplikasi Baru

- **Pertanyaan Sains Data**: "Berdasarkan peluncuran uji coba yang baru saja kita lakukan, berapakah estimasi rata-rata pengeluaran bulanan seorang pelanggan premium, dan seberapa besar ketidakpastian dalam estimasi tersebut?"
- **Data**: Anda mengambil sampel sebanyak 40 pelanggan premium (n = 40).
  - Anda menghitung Rata-rata sampel ($\bar{x}$): = AVERAGE(rentang) = Rp 1.200.000
  - Anda menghitung Standar Deviasi sampel (s): =STDEV.S(rentang) = Rp 142.000 (Catatan: Karena ini berasal dari sampel, kita menggunakan s, bukan σ.)

1. Menghitung Margin Kesalahan (*Margin of Error*) di Excel

Karena kita bekerja dengan standar deviasi populasi yang tidak diketahui, kita menggunakan fungsi CONFIDENCE.T pada tingkat kepercayaan 95% (α = 0,05).

- Rumus Excel: =CONFIDENCE.T(0,05; 14,2; 40)
- Hasil (Margin Kesalahan): 4,54

2. Menyusun Interval Kepercayaan (*Confidence Interval*)

Kita menerapkan margin kesalahan ini pada rata-rata sampel kita:

- Batas Bawah: 120 - 4,54 = 115,46
- Batas Atas: 120 + 4,54 = 124,54

Menafsirkan Hasil

"Kita memiliki tingkat kepercayaan 95% bahwa rata-rata pengeluaran bulanan yang sebenarnya untuk seluruh pelanggan premium berada di kisaran Rp 1.154.600 hingga Rp 1.245.400. Mengingat model bisnis kita memerlukan rata-rata pengeluaran setidaknya Rp 1.100 agar tetap menguntungkan, interval ini menunjukkan dengan tingkat kepastian statistik yang tinggi bahwa strategi penetapan harga kita saat ini telah mencapai tujuannya."

3. Interval Kepercayaan untuk Proporsi

Excel tidak menyediakan fungsi bawaan khusus untuk interval kepercayaan proporsi, namun semua komponen yang diperlukan tersedia:

- NORM.S.INV(probabilitas) → nilai kritis Z (*Z-critical value*)
- Aritmatika dasar untuk kesalahan standar (*standard error*)

Untuk proporsi sampel $\hat{p}$:

$$\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\mathrm{ME} = Z_{\alpha/2} \times \mathrm{SE}$$

$$\mathrm{CL} = \hat{p} + \mathrm{ME}$$

Contoh: Pengendalian Kualitas untuk Merek *E-commerce*

- **Pertanyaan Sains Data**: "Berapa proporsi pengisi daya portabel yang cacat dari kiriman terbaru kami yang berjumlah 500 unit?"
- **Data**: Seorang manajer gudang memilih 500 pengisi daya secara acak (n = 500) untuk diuji. Dari jumlah tersebut, ditemukan 10 pengisi daya yang cacat (x = 10).

Langkah 1: Proporsi Sampel (p̂) – Berapakah "Estimasi Titik" (*Point Estimate*) kita?

- Perhitungan: 10 / 500 = 0,02
- Interpretasi: Dalam sampel spesifik ini, 2% dari pengisi daya tersebut cacat (p̂ = 0,02)

Langkah 2: Tingkat Kepercayaan (Confidence Level) & Z-Score

Z-kritis untuk Tingkat Kepercayaan 95% (α = 0,05): =NORM.S.INV(1 - 0,05/2) = 1,96

Langkah 3: Margin Kesalahan (Margin of Error/ME)

Perhitungan Manual:

$$\mathrm{ME} = Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.02 \times (1-0.02)}{500}}$$

- Hasil: 0,012 (atau 1,2%)

Langkah 4: Menyusun Interval Kepercayaan (Confidence Interval)

- Batas Bawah: 0,02 - 0,012 = 0,008 (0,8%)
- Batas Atas: 0,02 + 0,012 = 0,032 (3,2%)

Interpretasi Hasil

"Berdasarkan sampel kami, kami memiliki tingkat keyakinan 95% bahwa tingkat kerusakan yang sebenarnya untuk keseluruhan pengiriman berada di antara 0,8% dan 3,2%."

Penerapan Interval Kepercayaan

Interval kepercayaan digunakan secara luas di berbagai bidang untuk mengukur ketidakpastian dan mendukung pengambilan keputusan. Contohnya:

- **Survei Kepuasan Pelanggan**: Sebuah perusahaan melaporkan skor kepuasan rata-rata sebesar 4,2 (dari skala 5) dengan interval kepercayaan 95% sebesar (4,0; 4,4). Hal ini menunjukkan bahwa skor kepuasan rata-rata yang sebenarnya untuk seluruh pelanggan kemungkinan besar berada dalam rentang tersebut.
- **Uji Klinis**: Sebuah obat baru menurunkan tekanan darah dengan rata-rata 10 mmHg, dengan interval kepercayaan 99% sebesar (8 mmHg; 12 mmHg). Interval ini memberikan estimasi yang presisi mengenai efek obat tersebut.
- **Jajak Pendapat Pemilihan Umum**: Jajak pendapat sering kali melaporkan hasil seperti "Kandidat A memperoleh dukungan 48% ± 3%." Margin kesalahan (*margin of error*) ini berkaitan langsung dengan interval kepercayaan untuk proporsi populasi.

Interval kepercayaan memberikan ringkasan yang lebih kaya dan informatif dibandingkan estimasi titik tunggal. Interval ini mengukur ketidakpastian, menunjukkan tingkat presisi, serta membantu analis dan pengambil keputusan dalam memahami rentang nilai yang mungkin bagi suatu parameter populasi.

Latihan: Margin Kesalahan / Interval Kepercayaan

Latihan-latihan ini dirancang untuk membantu Anda beralih dari teori ke praktik dengan membangun perangkat interaktif di Excel. Anda akan menghitung margin kesalahan, menyusun interval kepercayaan, serta mengeksplorasi bagaimana faktor-faktor kunci memengaruhi presisi statistik.

Latihan Excel 1: Membuat Kalkulator Margin Kesalahan

Dalam latihan ini, Anda akan membuat perangkat Excel yang dapat digunakan kembali untuk menghitung margin kesalahan, baik untuk nilai rata-rata maupun proporsi.

Bagian A: Margin Kesalahan untuk Rata-rata Populasi

Langkah 1: Menyiapkan Lembar Kerja

Masukkan label, nilai, dan rumus berikut ke dalam sel yang ditentukan:

CELL	LABEL / VALUE / FORMULA
A1	Margin of Error Rata-Rata Populasi
A3	Ukuran Sampel (n)
B3	100
A4	Simpangan Baku Populasi (σ)
B4	15
A5	Alfa (α)
B5	0.05
A6	Margin of Error
B6	=CONFIDENCE.NORM(B5, B4, B3)

Penjelasan

- CONFIDENCE.NORM() menghasilkan margin kesalahan (*margin of error*) untuk suatu nilai rata-rata dengan menggunakan distribusi normal.
- Interval kepercayaan (*confidence interval*) kemudian dirumuskan sebagai berikut:

$$CI = \bar{x} \pm Margin\ Kesalahan$$

Bagian B: Margin Kesalahan untuk Proporsi Populasi

Excel tidak menyediakan fungsi bawaan untuk proporsi, sehingga kita perlu menyusun rumusnya secara manual.

Langkah 1: Menyiapkan Lembar Kerja

Pada lembar kerja yang sama, masukkan label, nilai, dan rumus berikut ke dalam sel-sel yang telah ditentukan.

Cell	Label / Value / Formula
A8	Margin of Error Proporsi Populasi
A10	Ukuran Sampel (n)
B10	100
A11	Proporsi Sampel ($\hat{p}$)

B11	0.52
A12	Alfa (α)
B12	0.05
A13	Margin of Error
B13	=NORM.S.INV(1 - B12/2) * SQRT((B11*(1 - B11)) / B10)

Penjelasan

Rumus di B13 mengimplementasikan:

- NORM.S.INV(1 - α/2) menghitung nilai kritis Z (z)
- Suku akar kuadrat menghitung kesalahan standar (SE)
- MOE = z * SE

Interval Kepercayaan

$$CI = \hat{p} \pm ME$$

Latihan Excel 2: Membandingkan Skenario Pengambilan Sampel

Latihan ini membantu Anda membangun intuisi tentang bagaimana ukuran sampel dan tingkat kepercayaan memengaruhi presisi.

Langkah 1: Mendefinisikan Skenario

Menggunakan lembar kerja yang sama, masukkan nilai-nilai berikut ke dalam baris dan kolom yang ditentukan. Entri ini memperluas struktur yang Anda buat di Latihan 1.

Tabel 11.6 Pengambilan sampel berbasis Meaw

Baris	Label Baris	Kolom C	Kolom D	Kolom E
3	Ukuran Sampel (n)	400	400	1600
4	Simpangan Baku Populasi	15	15	15
5	Alpha (α)	0,05	0,01	0,05

Tabel 11.7 Perhitungan berbasis proporsi

Baris	Label Baris	Kolom C	Kolom D	Kolom E
10	Ukuran sampel (n)	400	400	1600
11	Proporsi sampel ($\hat{p}$)	0.52	0.52	0.52
12	Alfa (α)	0.05	0.01	0.05

Langkah 2: Menghitung Margin Kesalahan (*Margin of Error*)

- Salin rumus margin kesalahan untuk nilai rata-rata dari sel B6 ke sel C6, D6, dan E6.
- Salin rumus margin kesalahan untuk proporsi dari sel B13 ke sel C13, D13, dan E13. Langkah ini akan menghasilkan tabel perbandingan lengkap untuk semua skenario.

Langkah 3: Memeriksa Hasil

Renungkan pertanyaan-pertanyaan berikut untuk memperdalam pemahaman Anda:

- Bagaimana peningkatan ukuran sampel memengaruhi margin kesalahan?
- Apa dampak dari peningkatan tingkat kepercayaan (*confidence level*)?

- Pada titik mana Anda melihat penurunan manfaat (*diminishing returns*) dari penggunaan sampel yang lebih besar?

Wawasan ini mencerminkan pertimbangan yang dilakukan analis saat merancang survei, eksperimen, dan pengujian A/B.

Menafsirkan Margin Kesalahan: Panduan Praktis untuk Presisi

Memahami hasil sama pentingnya dengan menghitungnya.

Gunakan tolok ukur berikut saat menafsirkan margin kesalahan (terutama untuk proporsi):

- ±1–2% → Presisi sangat baik
- ±3–5% → Presisi baik (umum dalam survei dan jajak pendapat)
- ±5–10% → Dapat diterima untuk pemeriksaan eksploratif
- > ±10% → Presisi rendah (pertimbangkan untuk menambah ukuran sampel)

Latihan-latihan ini menggambarkan satu kebenaran penting:

Presisi tidak diperoleh secara cuma-cuma—presisi harus dicapai melalui data.

Sebagai seorang ilmuwan data (*data scientist*), peran Anda adalah menyeimbangkan antara:

- Persyaratan akurasi
- Keterbatasan sumber daya
- Urgensi pengambilan keputusan

Memahami perilaku margin kesalahan memungkinkan Anda merancang studi yang tidak hanya kuat secara statistik, tetapi juga layak untuk dilaksanakan secara praktis.

11.8 RANGKUMAN ESTIMASI DAN INFERENSI STATISTIK

Dalam bab ini, kita telah mengeksplorasi salah satu fondasi terpenting dalam pemikiran statistik: bagaimana menggunakan sampel untuk menarik kesimpulan yang dapat diandalkan mengenai populasi. Proses ini merupakan inti dari sains data, mengingat bekerja dengan populasi secara keseluruhan sering kali tidak praktis atau bahkan mustahil dilakukan. Kita mulai dengan membedakan antara populasi dan sampel, dengan menekankan bahwa sementara populasi mencakup seluruh elemen yang diteliti, sampel menyediakan cara yang efisien dan mudah dikelola untuk memperoleh wawasan. Kita memperkenalkan perbedaan mendasar antara parameter (karakteristik populasi) dan statistik (estimasi berbasis sampel), yang menjadi landasan bagi inferensi statistik.

Selanjutnya, kita menelaah berbagai metode pengambilan sampel beserta implikasinya. Kita mempelajari bahwa metode pengambilan sampel probabilitas seperti pengambilan sampel acak sederhana, berstrata, sistematis, dan klister memungkinkan inferensi statistik yang valid karena didasarkan pada pemilihan secara acak. Sebaliknya, metode pengambilan sampel non-probabilitas seperti pengambilan sampel berdasarkan kemudahan (*convenience sampling*), tujuan tertentu (*purposive sampling*), dan kuota berguna untuk studi eksploratif atau kualitatif, namun tidak mendukung generalisasi terhadap populasi.

Kita kemudian membahas aspek krusial mengenai penentuan ukuran sampel, dengan menyoroti bagaimana faktor-faktor seperti presisi (margin kesalahan), variabilitas, tingkat

kepercayaan, dan kendala praktis memengaruhi jumlah data yang diperlukan. Kita melihat bahwa sampel yang lebih besar umumnya menghasilkan estimasi yang lebih akurat dan andal, meskipun membutuhkan biaya dan waktu yang lebih besar. Konsep distribusi sampling menjadi jembatan teoretis antara data sampel dan inferensi populasi. Kita mempelajari bahwa statistik seperti rata-rata sampel dan proporsi sampel memiliki distribusinya sendiri, yang dicirikan oleh nilai rata-rata dan kesalahan baku (*standard error*). Yang terpenting, Teorema Limit Pusat (*Central Limit Theorem* atau CLT) menunjukkan bahwa untuk sampel yang cukup besar, distribusi sampling ini akan mendekati distribusi normal terlepas dari bentuk distribusi populasi aslinya. Hasil yang kuat ini memungkinkan kita untuk menerapkan penalaran berbasis probabilitas dalam berbagai skenario dunia nyata.

Kita kemudian menghubungkan Teorema Limit Pusat, distribusi probabilitas, dan inferensi statistik dengan memperkenalkan peran praktis dari distribusi Normal Standar (Z), distribusi t, nilai kritis, dan tingkat kepercayaan. Kita menelaah bagaimana distribusi t memperhitungkan ketidakpastian tambahan ketika simpangan baku populasi tidak diketahui, sehingga memberikan kerangka kerja yang jelas dalam memilih antara distribusi Z atau t untuk keperluan estimasi. Secara keseluruhan, konsep-konsep ini membentuk dasar untuk menyusun margin kesalahan dan selang kepercayaan (*confidence interval*) secara tepat. Berdasarkan landasan tersebut, kita memperkenalkan estimasi titik (*point estimation*) di mana satu nilai tunggal digunakan untuk mengestimasi parameter populasi serta membahas karakteristik estimator yang baik, seperti sifat tidak bias (*unbiasedness*), efisiensi, dan konsistensi. Namun, kita menekankan bahwa sebuah estimasi tunggal belumlah lengkap tanpa adanya ukuran ketidakpastian. Hal ini mengarah pada konsep margin kesalahan (*margin of error*), yang mengukur besaran penyimpangan yang mungkin terjadi antara estimasi sampel dan parameter populasi yang sebenarnya. Kami mempelajari bagaimana margin kesalahan bergantung pada ukuran sampel, tingkat kepercayaan, serta variabilitas, dan bagaimana hal tersebut secara langsung menentukan presisi dari estimasi yang kita buat. Terakhir, kami membahas interval kepercayaan (*confidence intervals*), yang menyajikan rentang nilai yang mungkin bagi suatu parameter populasi. Kami memperjelas interpretasi yang tepat mengenai tingkat kepercayaan serta mempelajari cara menyusun interval kepercayaan, baik untuk nilai rata-rata maupun proporsi, dengan menggunakan distribusi Z atau distribusi t, tergantung pada konteksnya. Melalui latihan berbasis Excel, kami menerapkan konsep-konsep ini ke dalam perangkat praktis untuk pemeriksaan data di dunia nyata.

BAB 12
DASAR-DASAR PENGUJIAN HIPOTESIS

12.1 LANDASAN PENGUJIAN HIPOTESIS

Pada bab-bab sebelumnya, kita telah membangun landasan statistik yang kuat mempelajari cara mendeskripsikan data melalui statistik deskriptif, mengukur ketidakpastian menggunakan probabilitas, mengambil sampel yang representatif, serta mengestimasi parameter populasi menggunakan statistik sampel. Perangkat-perangkat ini telah mempersiapkan kita untuk langkah krusial berikutnya: membuat keputusan berbasis bukti mengenai populasi dengan menggunakan data. Hal ini membawa kita pada salah satu kerangka kerja yang paling ampuh dan banyak digunakan dalam statistik: pengujian hipotesis.

Pengujian hipotesis menyediakan pendekatan terstruktur untuk mengevaluasi klaim, menguji asumsi, dan menentukan apakah pola yang teramati dalam data mencerminkan efek nyata atau sekadar akibat dari variasi acak. Ini merupakan jembatan antara data dan pengambilan keputusan, yang memungkinkan kita untuk beralih dari sekadar pengembangan teknologi masa depan menuju penarikan kesimpulan (inferensi) dengan penuh keyakinan dan ketelitian. Di akhir bab ini, Anda akan mampu:

- Memahami logika dan tujuan dari pengujian hipotesis
- Mendefinisikan serta membedakan dengan jelas antara hipotesis nol dan hipotesis alternatif
- Menerapkan proses enam langkah dalam melakukan pengujian hipotesis
- Mengidentifikasi dan menafsirkan kesalahan Tipe I dan Tipe II
- Membedakan antara uji satu arah (*one-tailed*) dan uji dua arah (*two-tailed*)
- Membedakan antara signifikansi statistik dan signifikansi praktis
- Memahami kekuatan statistik (*statistical power*) dan perannya dalam pengambilan keputusan
- Menghubungkan pengujian hipotesis dengan selang kepercayaan (*confidence intervals*)

Beberapa konsep ini mungkin terasa asing atau cukup menantang pada awalnya. Hal tersebut sangatlah wajar. Pengujian hipotesis dibangun berdasarkan gagasan-gagasan yang telah Anda pelajari sebelumnya, dan pemahaman Anda akan semakin mendalam seiring dengan eksplorasi kita melalui penjelasan yang jernih, penalaran intuitif, serta contoh-contoh praktis. Pelajarilah bab ini dengan rasa ingin tahu dan kesabaran setiap konsep merupakan langkah untuk mulai berpikir layaknya seorang ilmuwan data (*data scientist*).

12.2 KONSEP HIPOTESIS STATISTIK

Dari Pertanyaan Studi hingga Pernyataan yang Dapat Diuji

Setiap penyelidikan data dimulai dengan sebuah pertanyaan. Misalnya:

1. Apakah pupuk baru meningkatkan hasil panen?
2. Apakah pelanggan lebih puas dengan kebijakan layanan baru?

3. Apakah orang yang rutin berolahraga mempunyai tekanan darah yang lebih rendah?

Meskipun pertanyaan-pertanyaan ini penting, namun belum cocok untuk pemeriksaan statistik. Untuk memeriksanya menggunakan data, kita harus menerjemahkannya ke dalam pernyataan yang dapat diuji. Hipotesis adalah klaim yang jelas, spesifik, dan terukur mengenai parameter populasi seperti rata-rata, proporsi, atau perbedaan yang dapat dievaluasi menggunakan data. Ini mengubah pertanyaan studi umum menjadi bentuk yang dapat diperiksa melalui metode statistik.

Misalnya:

- **Pertanyaan Studi**: Apakah kafein meningkatkan konsentrasi?
- **Hipotesis yang Dapat Diuji**: Rata-rata skor perhatian individu yang mengonsumsi kafein lebih tinggi dibandingkan mereka yang tidak.

Hipotesis yang dirumuskan dengan baik memiliki tiga karakteristik penting:

- **Kejelasan**: Tepat dan tidak ambigu
- **Keterukuran**: Melibatkan kuantitas yang dapat diamati dan dicatat
- **Kepalsuan**: Hal ini dapat didukung atau ditantang dengan menggunakan data empiris

Sifat-sifat ini memastikan bahwa hipotesis dapat diuji secara bermakna dan bukan sekadar didiskusikan.

Contoh di Seluruh Domain

Hipotesis muncul di hampir setiap bidang di mana data digunakan untuk menginformasikan keputusan:

- **Bisnis**: Rata-rata skor kepuasan pelanggan meningkat setelah diperkenalkannya dukungan online.
- **Kesehatan**: Obat baru menurunkan kadar gula darah lebih efektif dibandingkan pengobatan standar.
- **Ilmu Pengetahuan Sosial**: Rumah tangga perkotaan menghabiskan lebih banyak uang untuk pendidikan dibandingkan rumah tangga pedesaan.

Dalam setiap kasus, data dapat dikumpulkan dan diperiksa untuk menentukan apakah bukti mendukung klaim tersebut. Kemampuan untuk beralih dari pertanyaan ke evaluasi berbasis bukti inilah yang membuat pengujian hipotesis menjadi alat yang ampuh.

12.3 LOGIKA PENGUJIAN HIPOTESIS

Metode Ilmiah dan Inferensi Statistik

Pertimbangkan dokter mengevaluasi obat baru. Pengamatan awal mungkin menunjukkan bahwa hal ini menurunkan tekanan darah tetapi bagaimana kita dapat menentukan apakah efek ini asli dan bukan hanya karena variasi acak? Demikian pula, seorang pemimpin bisnis mungkin mengamati peningkatan produktivitas setelah program pelatihan tetapi apakah peningkatan tersebut benar-benar disebabkan oleh program tersebut? Ini adalah jenis pertanyaan yang dirancang untuk dijawab oleh pengujian hipotesis. Pengujian hipotesis adalah kerangka formal untuk mengambil keputusan tentang populasi dengan menggunakan data sampel. Hal ini memungkinkan kita untuk beralih dari bukti yang diamati ke kesimpulan yang terinformasi, sehingga membentuk hubungan penting antara pemeriksaan

data dan pengambilan keputusan di dunia nyata. Pada intinya, pengujian hipotesis mengikuti logika metode ilmiah:

1. Ajukan pertanyaan → Apakah obat ini menurunkan tekanan darah?
2. Merumuskan hipotesis → Obat tersebut menurunkan tekanan darah setidaknya 10 poin
3. Mengumpulkan data → Melakukan studi terhadap sampel pasien
4. Pemeriksaan data → Terapkan metode statistik untuk mengevaluasi bukti
5. Menarik kesimpulan → Tentukan apakah bukti mendukung hipotesis

Proses terstruktur ini memastikan bahwa kesimpulan tidak didasarkan pada intuisi atau pengamatan yang terisolasi, namun pada bukti yang sistematis dan dapat direproduksi.

Membuat Keputusan dalam Ketidakpastian

Prinsip mendasar dalam statistik adalah bahwa keputusan berdasarkan data sampel pada dasarnya tidak pasti. Ketidakpastian ini muncul karena sampel yang berbeda dari populasi yang sama dapat menghasilkan hasil yang berbeda, sebuah konsep yang dikenal sebagai variabilitas pengambilan sampel. Misalnya, jika Anda mengukur tekanan darah 30 pasien hari ini dan kelompok berbeda yang terdiri dari 30 pasien besok, kemungkinan besar hasilnya akan berbeda bahkan jika obatnya tidak memberikan efek nyata. Variasi alami ini menyulitkan untuk membedakan antara efek asli dan fluktuasi acak. Pengujian hipotesis mengatasi tantangan ini dengan memberikan cara terstruktur untuk mengevaluasi apakah perbedaan yang diamati cukup besar untuk dianggap bermakna secara statistik, dan bukan hanya sekedar kebetulan. Alih-alih mencari kepastian mutlak, hal ini membantu kita membuat keputusan yang bersifat probabilistic mengukur seberapa kuat bukti yang ada dan seberapa yakin kita terhadap kesimpulan kita.

12.4 HIPOTESIS: NULL (H_0) DAN ALTERNATIF (H_1)

Uji hipotesis adalah kerangka terstruktur untuk menggunakan data sampel guna mengevaluasi klaim tentang parameter populasi seperti rata-rata (μ), proporsi (p), atau perbedaan antar parameter ($\mu_1 - \mu_2$). Pada intinya, hal ini melibatkan perumusan dua pernyataan yang bersaing dan menggunakan data untuk menentukan mana yang lebih baik didukung oleh bukti. Kedua pernyataan tersebut adalah hipotesis nol (H_0) dan hipotesis alternatif (H_1 atau H_a).

Hipotesis Nol (H_0): Status Quo

Dalam pengujian hipotesis, kita memulai dengan asumsi bahwa tidak ada efek atau perbedaan yang berarti. Asumsi ini berfungsi sebagai tolok ukur untuk membandingkan data yang kita amati. Hipotesis nol adalah pernyataan yang menyatakan bahwa tidak ada perbedaan, tidak ada perubahan, atau bahwa suatu parameter bernilai sama dengan angka tertentu. Ini merupakan asumsi dasar, status quo, atau pernyataan yang menunjukkan tidak adanya efek maupun hubungan. Dalam praktiknya, H_0 dianggap benar kecuali terdapat bukti kuat dalam data yang menunjukkan sebaliknya. Pendekatan yang hati-hati ini membantu mencegah kesimpulan yang terburu-buru atau tidak berdasar.

Contoh hipotesis nol:

- H_0: $\mu = 100$ (rata-rata populasi sama dengan 100)
- H_0: $\mu_1 = \mu_2$ (rata-rata dua populasi adalah sama)
- H_0: $p = 0{,}5$ (proporsi populasi sama dengan 0,5)

Tujuan utama dari pengujian hipotesis adalah untuk menilai apakah terdapat bukti yang cukup untuk menolak hipotesis nol.

Hipotesis Alternatif (H_1 atau H_a): Klaim Studi

Hipotesis alternatif (H_1 atau H_a) mewakili klaim yang ingin kita buktikan. Hipotesis ini secara langsung bertentangan dengan hipotesis nol dan mencerminkan adanya efek, perbedaan, atau hubungan. Penting untuk ditekankan bahwa pengujian statistik tidak membuktikan hipotesis alternatif. Sebaliknya, kita menentukan apakah bukti yang ada cukup kuat untuk menolak H_0 demi mendukung hipotesis alternatif tersebut. Jika bukti tidak mencukupi, kita hanya "gagal menolak" H_0 kita tidak menyatakan bahwa H_0 adalah benar.

Hipotesis alternatif dapat berbentuk dua jenis utama:

1. **Dua Arah (Non-Direksional)**

- H_1: $\mu \neq 100$
- Digunakan ketika kita ingin mendeteksi adanya perbedaan, tanpa memandang arah perbedaannya.

2. **Satu Arah (Direksional)**

- H_1: $\mu > 100$
- H_1: $\mu < 100$
- Digunakan ketika kita secara khusus tertarik pada arah perubahan tertentu.

Contoh Skenario

Tabel 12.1: Contoh Hipotesis

Situasi	**H_o**	**H_o**	**Jenis**
Menguji apakah suatu obat baru dapat menurunkan tekanan darah	μ = 120 mmHg	μ = 120 mmHg	Satu sisi
Menguji apakah terdapat perbedaan antara dua metode pengajaran	$\mu_1 = \mu_2$	$\mu_1 = \mu_2$	Dua sisi
Menguji apakah suatu proses baru dapat meningkatkan hasil	μ = 500 satuan	μ = 500 satuan	Satu sisi

Karakteristik Utama Hipotesis

Agar suatu pengujian hipotesis dapat dianggap valid, H_o dan H_1 harus memenuhi ketentuan berikut:

- **Komplementer**: Keduanya mencakup seluruh kemungkinan hasil yang ada.
- **Saling lepas (*mutually exclusive*)**: Keduanya tidak dapat benar secara bersamaan.
- **Ditetapkan sebelumnya**: Hipotesis harus dirumuskan sebelum pengumpulan atau pemeriksaan data dilakukan untuk menghindari bias.

Analogi Ruang Sidang:

Cara yang berguna untuk membangun intuisi mengenai pengujian hipotesis adalah dengan membayangkan situasi di ruang sidang:

- **Terdakwa**: Di ruang sidang, terdakwa dianggap tidak bersalah sampai terbukti bersalah. Anggapan tidak bersalah ini sejajar dengan hipotesis nol (H_0), yang kita asumsikan benar kecuali terdapat bukti kuat yang menunjukkan sebaliknya.
- **Jaksa Penuntut**: Peran jaksa adalah menyajikan bukti untuk mencoba membuktikan kesalahan terdakwa. Hal ini mencerminkan hipotesis alternatif (H_1), yang mewakili klaim yang ingin kita dukung melalui bukti-bukti yang ada.
- **Bukti**: Data yang dikumpulkan dari sampel berfungsi layaknya bukti yang disajikan di pengadilan.
- **Hakim atau Juri**: Mereka mengevaluasi bukti dan memutuskan apakah bukti tersebut cukup kuat untuk membatalkan anggapan tidak bersalah. Dalam pengujian hipotesis, kita mengevaluasi data untuk menentukan apakah terdapat bukti yang cukup untuk menolak H_0 dan menerima H_1.
- **"Melampaui Keraguan yang Wajar" (*Beyond a Reasonable Doubt*)**: Standar hukum ini mencerminkan tingkat kepastian yang diperlukan untuk menjatuhkan vonis bersalah. Dalam statistika, konsep yang setara adalah tingkat signifikansi (α) ambang batas yang kita gunakan untuk memutuskan apakah bukti yang ada cukup kuat untuk menolak hipotesis nol.

Penting untuk dicatat bahwa vonis "tidak bersalah" bukan berarti terdakwa terbukti tidak bersalah; hal itu hanya menunjukkan bahwa bukti yang ada tidak cukup untuk menjatuhkan vonis bersalah. Demikian pula, kegagalan dalam menolak hipotesis nol tidak membuktikan bahwa H_0 adalah benar, hal itu hanya menunjukkan bahwa data sampel tidak memberikan bukti yang cukup kuat untuk menyimpulkan sebaliknya.

12.5 KONSEP DASAR PENGUJIAN HIPOTESIS

Sebelum menerapkan pengujian hipotesis, penting untuk memahami konsep-konsep kunci yang mendasari keseluruhan proses tersebut. Meskipun beberapa gagasan ini mungkin terasa baru, banyak di antaranya yang dibangun berdasarkan konsep yang telah diperkenalkan pada bab-bab sebelumnya terutama mengenai distribusi sampling dan estimasi. Seiring Anda mempelajari berbagai contoh, gagasan-gagasan ini akan menjadi semakin intuitif.

Jenis-jenis Uji Statistik

Inti dari setiap pengujian hipotesis adalah statistik uji sebuah nilai numerik tunggal yang merangkum seberapa jauh hasil sampel Anda menyimpang dari apa yang diharapkan jika hipotesis nol (H_0) benar. Untuk menafsirkan nilai ini, kita membandingkannya dengan suatu distribusi probabilitas. Pemilihan distribusi dan dengan demikian jenis uji statistik yang digunakan bergantung pada karakteristik data serta parameter yang sedang diteliti. Anda akan menjumpai beberapa jenis pengujian, termasuk uji-z, uji-t, uji-F, dan uji chi-square (χ^2).

Memahami Nama Uji Statistik: Analogi Alat Ukur

Bayangkan uji statistik sebagai berbagai alat ukur yang ada di sebuah bengkel. Anda akan menggunakan pita pengukur untuk jarak yang jauh dan penggaris kecil untuk jarak yang pendek. Dengan cara yang sama, setiap uji statistik merupakan alat khusus yang dirancang untuk jenis data atau pertanyaan tertentu. Berikut adalah poin utamanya:

Setiap uji statistik dinamai berdasarkan distribusi probabilitas yang digunakan untuk mengevaluasi statistik ujinya.

Setelah Anda menghitung statistik uji dari data Anda, pertanyaan utamanya adalah:

"Seberapa tidak lazim nilai ini jika hipotesis nol benar?"

Untuk menjawabnya, Anda membandingkan statistik tersebut dengan suatu distribusi probabilitas. Situasi yang berbeda memerlukan distribusi yang berbeda pula itulah sebabnya terdapat beragam jenis uji statistik.

Uji Statistik yang Umum Digunakan

Berikut adalah gambaran umum mengenai uji statistik yang paling sering digunakan:

1. **Uji-z (z-test)**

   Distribusi yang digunakan: Distribusi normal standar (distribusi-z)

   Kapan digunakan:

   - Simpangan baku populasi (σ) diketahui
   - Ukuran sampel besar (biasanya $n \geq 30$)
   - Menguji proporsi populasi (dengan sampel besar)

Contoh: Menguji apakah rata-rata waktu tunggu pelanggan berbeda dari tolok ukur yang telah diketahui.

2. **Uji-t (t-test)**

Distribusi yang digunakan: Distribusi-t (mirip dengan distribusi normal, tetapi dengan ekor yang lebih tebal) Kapan digunakan:

- Simpangan baku populasi tidak diketahui (kasus yang paling umum)
- Ukuran sampel kecil hingga sedang
- Membandingkan rata-rata antara satu atau dua kelompok

Contoh: Mengevaluasi apakah metode pengajaran baru dapat meningkatkan kinerja mahasiswa.

3. **Uji-F (F-test)**

Distribusi yang digunakan: Distribusi-F Kapan digunakan:

- Membandingkan varians antar kelompok
- Menguji perbedaan rata-rata di antara tiga kelompok atau lebih (ANOVA)
- Menilai signifikansi model secara keseluruhan dalam regresi

Contoh: Menentukan apakah kepuasan pelanggan berbeda di berbagai lokasi toko.

4. **Uji Chi-square (χ^2)**

Distribusi yang digunakan: Distribusi Chi-square Kapan digunakan:

- Memeriksa data kategorikal (jumlah atau frekuensi)
- Menguji kecocokan (*goodness-of-fit*) atau independensi antar variabel

Contoh: Menentukan apakah preferensi pelanggan berbeda dari ekspektasi peluang acak, atau apakah terdapat hubungan antara jenis kelamin dan pilihan produk.

Menyesuaikan Uji Statistik dengan Parameter

Jenis parameter yang diuji juga menentukan uji statistik serta distribusi yang bersesuaian mana yang tepat untuk digunakan:

- **Rata-rata (μ)**: Uji-z atau uji-t (tergantung pada ukuran sampel dan apakah σ diketahui)
- **Proporsi (p):** Uji-z (dengan sampel besar)
- **Data kategorikal:** Uji Chi-square (χ^2)
- **Varians:** Uji-F

Anda tidak perlu menghafal aturan-aturan ini seketika. Nantinya dalam bab ini, Anda akan mempelajari pendekatan terstruktur untuk memilih uji statistik yang tepat.

Tingkat Signifikansi (α): Menetapkan Ambang Batas Keputusan

Tingkat signifikansi (α) menentukan ambang batas untuk memutuskan apakah bukti yang ada cukup kuat untuk menolak H_0. Nilai ini merepresentasikan probabilitas maksimum terjadinya kesalahan Tipe I yaitu menolak hipotesis nol yang sebenarnya benar. Pilihan yang paling umum digunakan adalah $\alpha = 0{,}05$, yang berarti kita menerima kemungkinan sebesar 5% untuk menolak hipotesis nol yang sebenarnya benar. Nilai lain yang juga sering digunakan meliputi:

- α = 0,01 (lebih ketat; digunakan dalam konteks yang memiliki risiko tinggi)
- α = 0,10 (lebih fleksibel; terkadang digunakan dalam studi eksploratif)

Menentukan α melibatkan sebuah kompromi (*trade-off*):

- α yang lebih rendah → lebih sedikit hasil positif palsu (*false positive*), namun risiko melewatkan efek nyata menjadi lebih besar.
- α yang lebih tinggi → sensitivitas lebih besar, namun risiko terjadinya hasil positif palsu meningkat.

Penting: Tingkat signifikansi harus ditetapkan sebelum memeriksa data untuk menghindari bias.

Statistik Uji dan Kriteria Penolakan Hipotesis

Statistik uji meringkas data sampel Anda ke dalam satu nilai tunggal yang mengukur seberapa jauh statistik sampel tersebut dari nilai yang dinyatakan dalam hipotesis nol, yang diukur dalam satuan kesalahan baku (*standard error*). Statistik uji yang umum digunakan:

1. Statistik-z
2. Statistik-t
3. Statistik-χ^2 (*chi-square*)
4. Statistik-F

Rumus umum untuk statistik-z dan statistik-t:

$$\text{Statistik Uji} = \frac{\text{Statistik Sampel} - \text{Parameter Hipotesis}}{\text{Kesalahan Baku}}$$

Sebagai contoh, dalam uji-z:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Nilai ini menunjukkan seberapa jauh rata-rata sampel dari rata-rata populasi yang dihipotesiskan, diukur dalam satuan kesalahan baku (*standard error*).

Nilai Kritis dan Wilayah Penolakan

Nilai kritis menentukan batas antara hasil yang wajar (tipikal) dan hasil yang ekstrem di bawah asumsi H_0. Jika statistik uji berada di luar batas ini, maka statistik tersebut jatuh ke dalam wilayah penolakan artinya, terdapat bukti yang cukup kuat untuk menolak hipotesis nol. Konsep utama:

- **Wilayah penolakan**: Nilai-nilai ekstrem yang sangat kecil kemungkinannya terjadi di bawah H_0.
- **Wilayah non-penolakan**: Nilai-nilai yang konsisten dengan H_0.
- **Nilai kritis**: Batas yang ditentukan, bergantung pada distribusi yang dipilih dan tingkat signifikansi (α).

Distribusi yang berbeda menghasilkan nilai kritis yang berbeda pula karena perbedaan bentuk distribusinya:

- Distribusi t: Bergantung pada derajat kebebasan (*degrees of freedom*); akan mendekati distribusi normal seiring dengan bertambahnya ukuran sampel.
- Distribusi χ^2 dan F: Memiliki kemiringan ke kanan (*right-skewed*); nilai kritisnya selalu bernilai positif.

Arah Ekor: Uji Satu Arah vs. Uji Dua Arah

Struktur wilayah penolakan bergantung pada hipotesis alternatif. Pada uji satu arah, seluruh wilayah penolakan terletak pada satu sisi (satu ekor), sehingga nilai kritis berfungsi sebagai batas untuk mendeteksi efek ke arah tertentu. Sebagai contoh, dengan $\alpha = 0{,}05$, nilai kritis z adalah 1,645 untuk uji satu arah ke kanan, atau –1,645 untuk uji satu arah ke kiri. Pada uji dua arah, wilayah penolakan terbagi ke dalam kedua sisi (kedua ekor), sehingga digunakan dua nilai kritis; untuk $\alpha = 0{,}05$, nilai kritis z adalah –1,96 dan +1,96, yang mencakup hasil ekstrem di kedua arah. Tabel 12.2 menyajikan nilai kritis z untuk tingkat signifikansi (α) yang umum digunakan, berdasarkan arah uji z.

Tabel 12.2 Nilai kritis z untuk tingkat signifikansi yang umum digunakan

Alfa (α)	Z-kritis (uji sisi kiri)	Zkritis Ekor Kanan	Rentang kritis Z dua arah
0,01	-2,326	2,326	± 2,576
0,05	-1,645	1,645	± 1,96
0,1	-1,28	1,28	± 1,645

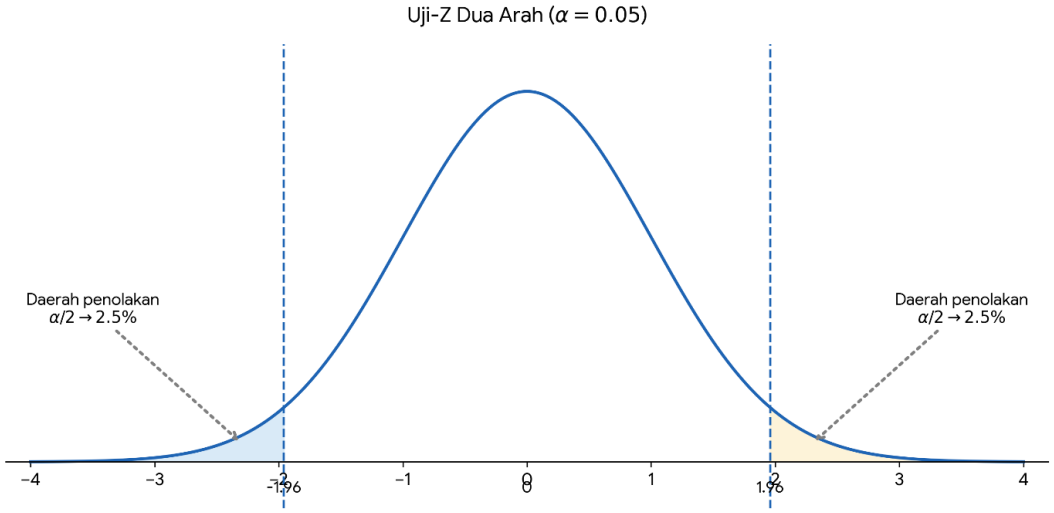
- **Uji satu arah sisi kiri**: Tolak H_0 jika statistik uji lebih kecil daripada nilai kritis negatif.
- **Uji satu arah sisi kanan**: Tolak H_0 jika statistik uji lebih besar daripada nilai kritis positif.

- **Uji dua arah**: Tolak H_0 jika statistik uji berada di luar interval yang ditentukan oleh ± nilai kritis. Representasi visual sering kali membuat konsep-konsep ini lebih mudah dipahami.

1. Uji Dua Arah

Digunakan ketika kita menguji perbedaan ke kedua arah (contoh: $\mu \neq \mu_0$).

- Wilayah penolakan terbagi di kedua ujung (ekor) distribusi.
- Untuk $\alpha = 0{,}05$, bagian tengah sebesar 95% dari distribusi merupakan wilayah di mana kita gagal menolak H_0.
- Bagian luar sebesar 5% (masing-masing 2,5% di setiap ujung) merupakan wilayah di mana kita menolak H_0.



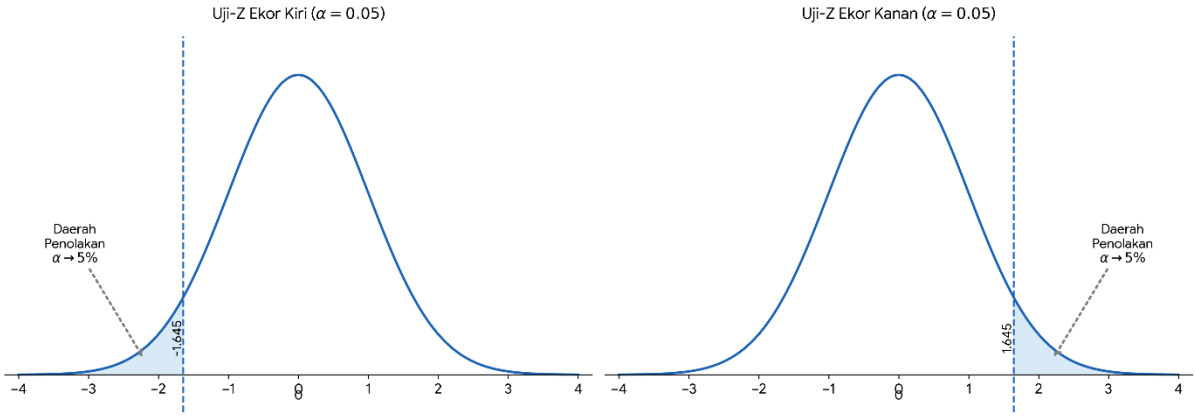
Gambar 12.1 Dua wilayah penolakan (α = 0,05)

Hanya uji yang didasarkan pada distribusi simetris (z dan t) yang dapat bersifat dua arah (*two-tailed*). Uji yang didasarkan pada distribusi χ^2 dan F secara inheren bersifat satu arah (*one-tailed*) karena sifatnya yang tidak negatif.

2. Uji Satu Arah (*One-Tailed Test*)

Digunakan ketika pertanyaan studi menetapkan suatu arah (misalnya, $\mu > \mu 0$ atau $\mu < \mu 0$).

- Seluruh wilayah penolakan terletak pada satu sisi (satu ekor distribusi).
- Untuk α = 0,05, wilayah penolakan mencakup 5% bagian bawah (satu arah kiri) atau 5% bagian atas (satu arah kanan) dari distribusi tersebut.
- Sebesar 95% sisanya merupakan wilayah di mana kita gagal menolak H_0.



Gambar 12.2 Wilayah penolakan satu arah kiri dan satu arah kanan (α = 0,05)

Uji yang didasarkan pada distribusi asimetris, seperti uji chi-square (χ^2) dan uji F, pada dasarnya bersifat satu arah (*one-tailed*) karena statistik ujinya tidak dapat bernilai negatif.

Nilai-p (p-value): Mengukur Bukti untuk Menentang H_0

Nilai-p adalah probabilitas untuk memperoleh hasil yang sama ekstremnya dengan (atau lebih ekstrem daripada) hasil sampel yang diamati, dengan asumsi bahwa hipotesis nol (H_0) adalah benar. Nilai ini memberikan ukuran langsung mengenai seberapa kompatibel data tersebut dengan H_0:

- Nilai-p besar (misalnya, 0,40): Data konsisten dengan H_0 → bukti lemah
- Nilai-p kecil (misalnya, 0,01): Data tidak mungkin terjadi di bawah asumsi H_0 → bukti kuat

Aturan keputusan:

- Jika nilai-p ≤ α → Tolak H_0
- Jika nilai-p > α → Gagal menolak H_0

Pendekatan nilai-p menawarkan alternatif yang fleksibel dan banyak digunakan dibandingkan metode nilai kritis (*critical value*), meskipun keduanya akan menghasilkan kesimpulan yang sama jika diterapkan dengan benar.

12.6 PROSES PENGUJIAN HIPOTESIS: PROSEDUR ENAM LANGKAH

Pengujian hipotesis mengikuti urutan langkah yang jelas dan logis. Untuk mengilustrasikan proses ini, kita akan menggunakan satu contoh yang konsisten.

Skenario Contoh

Sebuah perusahaan mengklaim bahwa berat rata-rata sereal dalam kemasannya adalah 500 g. Sebuah kelompok hak konsumen mencurigai bahwa berat rata-rata yang sebenarnya kurang dari 500 g. Untuk menyelidikinya, sejumlah kotak sereal diambil sebagai sampel dan diperiksa.

Langkah 1: Menyatakan Hipotesis (H_0 dan H_1)

Hipotesis Nol (H_0): Rata-rata populasi sama dengan 500 g. (Tidak ada perbedaan, status quo)

$$H_0: \mu = 500$$

Hipotesis Alternatif (H_1): Rata-rata populasi kurang dari 500 g. (Klaim yang sedang diuji)

$$H_1: \mu < 500$$

Karena klaim tersebut berfokus pada pendeteksian penurunan, maka ini merupakan uji satu arah (sisi kiri).

Langkah 2: Tentukan Tingkat Signifikansi (α)

Tingkat signifikansi (α) menetapkan ambang batas untuk memutuskan kapan bukti dianggap cukup kuat guna menolak H_0. Pilihan yang umum digunakan meliputi:

- α = 0,01 (1%) → lebih konservatif
- α = 0,05 (5%) → paling sering digunakan
- α = 0,10 (10%) → kurang konservatif. Untuk contoh ini, kita menggunakan α = 0,05.

Langkah 3: Pilih Uji Statistik yang Sesuai

Pemilihan uji statistik bergantung pada:

Jenis data (numerik atau kategorikal)

- Jumlah sampel (satu, dua, atau lebih kelompok)
- Parameter yang diteliti (rata-rata, proporsi, varians)
- Ukuran sampel
- Apakah standar deviasi populasi (σ) diketahui atau tidak

Panduan:

- Gunakan uji-z (z-test) jika σ diketahui atau untuk sampel berukuran besar
- Gunakan uji-t (t-test) jika σ tidak diketahui (kasus yang paling umum)
- Gunakan uji χ^2 (chi-square) untuk data kategorikal
- Gunakan uji-F (F-test) untuk membandingkan beberapa varians atau rata-rata kelompok

Dalam contoh ini:

- Data numerik
- Satu sampel
- σ tidak diketahui
- Ukuran sampel $n = 25 (< 30)$

→ Kita menggunakan uji-t satu sampel (*one-sample t-test*).

Langkah 4: Hitung Statistik Uji

Statistik uji merangkum seberapa jauh hasil sampel menyimpang dari hipotesis nol, yang diukur dalam satuan kesalahan baku (*standard error*).

Untuk uji-t satu sampel:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Di mana:

- $\bar{x}$ = rata-rata sampel
- $\mu 0$ = rata-rata yang dihipotesiskan
- s = simpangan baku sampel
- n = ukuran sampel

Data sampel:

- $\bar{x}$ = 495
- s = 10
- n = 25

$$t = \frac{495 - 500}{\frac{10}{\sqrt{25}}} = \frac{-5}{2} = -2.5$$
$$t = -2.5$$

Langkah 5: Menentukan Aturan Keputusan (p-value vs. Nilai Kritis)

Terdapat dua pendekatan yang setara pendekatan p-value dan pendekatan nilai kritis. Meskipun menggunakan kerangka kerja keputusan yang berbeda, kedua metode ini akan selalu menghasilkan kesimpulan yang sama jika didasarkan pada tingkat signifikansi yang sama. Anda akan mempelajari setiap pendekatan secara mendetail pada subbagian berikutnya.

Pendekatan A: Metode p-Value

p-value merepresentasikan probabilitas untuk memperoleh statistik uji yang sama ekstremnya dengan hasil yang didapat, dengan asumsi H_0 adalah benar.

Menggunakan Excel:

$$= T.DIST(-2.5, 24, TRUE) \approx 0{,}010$$

Aturan keputusan:

- Jika p ≤ α → Tolak H_0
- Jika p > α → Gagal menolak H_0.

Karena p = 0,01 < 0,05, maka kita menolak H_0.

Pendekatan B: Metode Nilai Kritis

Nilai kritis menentukan batas wilayah penolakan.

- α = 0,05
- Derajat kebebasan (*degrees of freedom*): df = n − 1 = 24

Nilai kritis untuk uji sisi kiri (*left-tailed test*)

=T.INV(0.05, 24) ≈ −1,711

Aturan keputusan:

- Jika t ≤ nilai kritis → Tolak H_0

Karena t = −2,5 < −1,711, maka statistik uji berada di dalam wilayah penolakan → tolak H_0.

Langkah 6: Membuat Keputusan dan Menyatakan Kesimpulan

Keputusan Statistik: Tolak H_0 (berdasarkan kedua metode)

Interpretasi Praktis: Terdapat bukti yang cukup pada tingkat signifikansi 5% untuk menyimpulkan bahwa rata-rata berat sereal kurang dari 500 g.

Pendekatan P-Value Vs. Pendekatan Nilai Kritis (Uji Z & Uji T)

Kedua pendekatan ini secara matematis setara dan selalu menghasilkan kesimpulan yang sama jika diterapkan dengan benar.

1. Pendekatan p-Value

Langkah-langkah

- Hitung statistik uji.
- Hitung p-value: probabilitas untuk memperoleh hasil yang sama ekstremnya dengan hasil sampel jika H_0 adalah benar.
- Tolak H_0 jika $p < \alpha$.

Kekuatan: Memberikan ukuran bukti yang tepat untuk menentang H_0.

Menghitung Nilai-p (p-value) dengan Excel

Tabel 12.3 Fungsi Nilai-p untuk Uji-Z

Jenis Tes	Formulir Hipotesis	Fungsi Excel	Penjelasan
Uji Satu Arah Kiri	H_1: $\mu < \mu_o$	=NORM.S.DIST(z, TRUE)	Menghasilkan probabilitas di sebelah kiri z. Memberikan nilai-p (p-value) secara langsung.
Uji Satu Arah Kanan	H_1: $\mu > \mu_o$	=1 - NORM.S.DIST(z, TRUE)	Menghasilkan probabilitas di sebelah kanan z.
Uji Dua Arah	H_1: $\mu \neq \mu_o$	=2 * (1 - NORM.S.DIST(ABS(z), TRUE))	Melipatgandakan probabilitas ekor yang lebih kecil untuk memperhitungkan kedua sisi.

Wawasan Utama (Uji-Z)

- NORM.S.DIST(z, TRUE) selalu memberikan probabilitas kumulatif (sisi kiri)
- Anda perlu menyesuaikannya bergantung pada arah hipotesis

Tabel 12.4 Fungsi p-value untuk Uji-t

Jenis Tes	Formulir Hipotesis	Fungsi Excel	Penjelasan
Uji Satu Arah Kiri	H_1: $\mu < \mu_o$	=T.DIST(t, df, TRUE)	Mengembalikan probabilitas di sebelah kiri t.
Uji Satu Arah Kanan	H_1: $\mu > \mu_o$	=T.DIST.RT(t, df)	Mengembalikan probabilitas di sebelah kanan t.

Uji Dua Arah	$H_1: \mu \neq \mu_o$	=T.DIST.2T(ABS(t), df)	Mengembalikan probabilitas gabungan dari kedua ujung (kedua sisi).

Ringkasan Konseptual

Arah menentukan rumus

- Ekor kiri → gunakan kumulatif (kiri)
- Ekor kanan → kurangi dari 1 (z) atau gunakan .RT (t)
- Dua ekor → gandakan probabilitas ekor

Tabel 12.5 Perbedaan Utama Antara Z dan T kecuali

Fitur	**Uji-Z**	**Uji-t (T-Test)**
Distribusi	Normal Standar	Distribusi-t (t-distribution)
Fungsi Excel	NORM.S.DIST	T.DIST, T.DIST.RT, T.DIST.2T
Derajat Kebebasan	Tidak diperlukan	Diperlukan (df = n - 1)

2. Pendekatan Nilai Kritis

Langkah-langkah

- Tentukan nilai α terlebih dahulu.
- Tentukan nilai kritis dari distribusi yang sesuai.
- Hitung statistik uji.
- Tolak H_0 jika statistik uji berada dalam wilayah penolakan.

Menghitung Nilai Kritis dengan Excel

Tabel 12.6 Fungsi Nilai Kritis untuk Uji Z

Jenis Uji	**Bentuk Hipotesis**	**Fungsi Excel**	**Contoh (α = 0,05)**	**Penjelasan**
Uji Satu Arah Kiri	$H_1: \mu < \mu_o$	=NORM.S.INV(α)	=NORM.S.INV(0,05)	Menghasilkan skor-z dengan luas α di sebelah kiri
Uji Satu Arah Kanan	$H_1: \mu > \mu_o$	=NORM.S.INV(1 - α)	=NORM.S.INV(1 - 0,05)	Menghasilkan skor-z dengan luas α di sebelah kanan
Uji Dua Arah	$H_1: \mu \neq \mu_o$	=NORM.S.INV(1 - α/2)	=NORM.S.INV(1 - 0,05/2)	Membagi α menjadi dua ekor (masing-masing sebesar α/2 di setiap sisi)

Tabel 12.7 Fungsi Nilai Kritis untuk Uji-t

Jenis Uji	**Bentuk Hipotesis**	**Fungsi Excel**	**Contoh α =0,05, df=24**	**Penjelasan**
Uji Satu Arah Kiri	H_1: $\mu < \mu_o$	=T.INV(α, df)	=T.INV(0,05; 24)	Menghasilkan skor-t dengan luas α di sebelah kiri
Uji Satu Arah Kanan	H_1: $\mu > \mu_o$	=T.INV(1 - α, df)	=T.INV(1 - 0,05; 24)	Menghasilkan skor-t dengan luas α di sebelah kanan
Uji Dua Arah	H_1: $\mu \neq \mu_o$	=T.INV.2T(α, df)	=T.INV.2T(0,05; 24)	Menghasilkan nilai kritis positif; gunakan ±

Ringkasan Konseptual (Penting untuk Pemahaman)

1. Gagasan Utama di Balik Nilai Kritis (*Critical Values*)

Nilai kritis adalah titik batas yang memisahkan:

- Hasil yang mungkin terjadi (gagal menolak H_0)
- Hasil yang tidak mungkin terjadi (menolak H_0)

2. Arah Menentukan Rumus

Arah	**Logika**
Ujung kiri	Gunakan α secara langsung → ujung bawah
Ujung kanan	Gunakan (1 − α) → ujung atas
Dua ujung	Bagi α → gunakan (1 − α/2)

3. Mengapa Uji Dua Arah Menggunakan α/2? Karena:

- Total area penolakan = α
- Dibagi sama rata ke dalam dua ekor (sisi)
- Setiap ekor mendapatkan α/2

4. Perbedaan Utama: Z vs T dalam Excel

Fitur	**Uji Z**	**Uji-t**
Fungsi Excel	NORM.S.INV()	T.INV(), T.INV.2T()
Memerlukan df?	Tidak	Ya (df = n - 1)
Output (Dua arah)	± nilai secara manual	Nilai positif (± tersirat)

Perbedaan Utama dalam Pola Pikir

- **Pendekatan Nilai Kritis (*Critical Value Approach*)**: "Apakah statistik uji saya berada di dalam wilayah penolakan?"
- **Pendekatan Nilai-p (*p-Value Approach*)**: "Seberapa ekstrem hasil yang saya peroleh dengan asumsi bahwa H_0 adalah benar?"

Kedua perspektif ini menghasilkan kesimpulan yang sama, keduanya hanya menyajikan pengambilan keputusan dengan cara yang berbeda.

12.7 JENIS KESALAHAN DAN KEKUATAN STATISTIK (*STATISTICAL POWER*)

Dalam pengujian hipotesis, keputusan dibuat berdasarkan data sampel. Akibatnya, selalu ada kemungkinan untuk menarik kesimpulan yang salah. Dua jenis kesalahan dapat terjadi:

Kesalahan Tipe I (α): Positif Palsu (*False Positive*)

Kesalahan Tipe I terjadi ketika kita menolak hipotesis nol yang sebenarnya benar. Dengan kata lain, kita menyimpulkan bahwa terdapat suatu efek, padahal kenyataannya tidak ada.

- **Probabilitas**: P(Kesalahan Tipe I) = α
- I**nterpretasi**: Tingkat signifikansi (α) mencerminkan risiko yang bersedia kita terima untuk melakukan kesalahan ini.

Contoh: Menyatakan suatu obat efektif, padahal sebenarnya obat tersebut tidak memiliki efek nyata.

Kesalahan Tipe II (β): Negatif Palsu (*False Negative*)

Kesalahan Tipe II terjadi ketika kita gagal menolak hipotesis nol yang sebenarnya salah. Ini berarti kita menyimpulkan bahwa tidak ada efek, padahal kenyataannya terdapat suatu efek.

- **Probabilitas**: P(Kesalahan Tipe II) = β

Contoh: Tidak mengakui efektivitas suatu obat yang sebenarnya memang efektif.

Kekuatan Statistik (1 – β)

Kekuatan statistik adalah probabilitas untuk menolak hipotesis nol yang salah secara tepat. Hal ini mencerminkan kemampuan uji tersebut untuk mendeteksi efek nyata jika efek itu memang ada.

$$Kekuatan\ (Power)\ =\ 1\ -\ \beta\ =\ P(Menolak\ H_0\ \mid\ H_0\ adalah\ salah)$$

Uji dengan kekuatan statistik yang tinggi lebih mungkin untuk mengidentifikasi efek yang bermakna.

Faktor-Faktor yang Memengaruhi Kekuatan Statistik (*Statistical Power*)

Beberapa faktor memengaruhi kekuatan suatu pengujian:

- **Ukuran Sampel (n)**: Sampel yang lebih besar mengurangi kesalahan standar (*standard error*), sehingga meningkatkan kekuatan pengujian.

- **Besaran Efek (*Effect Size*)**: Perbedaan nyata yang lebih besar akan lebih mudah dideteksi.
- **Variabilitas (σ)**: Variabilitas yang lebih rendah meningkatkan kejelasan sinyal dan memperkuat pengujian.
- **Tingkat Signifikansi (α)**: Nilai α yang lebih tinggi meningkatkan kekuatan pengujian, namun juga meningkatkan risiko kesalahan Tipe I (*Type I error*).
- **Arah Pengujian**: Pengujian satu arah (*one-tailed test*) yang ditentukan dengan tepat memiliki kekuatan lebih besar dalam arah yang telah ditetapkan tersebut.

Pertukaran (*Trade-off*) antara α dan β

Untuk ukuran sampel yang tetap:

- Menurunkan α → mengurangi kesalahan Tipe I, tetapi meningkatkan β (mengurangi kekuatan pengujian).
- Meningkatkan α → meningkatkan kekuatan pengujian, tetapi meningkatkan risiko kesalahan Tipe I.

Untuk mengurangi kedua jenis kesalahan tersebut secara bersamaan, tingkatkan ukuran sampel.

Memilih Tingkat Signifikansi yang Tepat

Pemilihan α harus mencerminkan konteks studi:

- Keputusan dengan konsekuensi besar (misalnya, dalam bidang kedokteran atau keuangan): Gunakan α yang lebih kecil (contoh: 0,01).
- Studi eksploratif: Nilai α yang lebih besar (contoh: 0,10) mungkin dapat diterima.
- Praktik terbaik: Selalu berikan justifikasi untuk nilai α yang dipilih berdasarkan konsekuensi nyata dari kesalahan yang mungkin terjadi.

12.8 SIGNIFIKANSI STATISTIK VS. SIGNIFIKANSI PRAKTIS

Salah satu pembedaan terpenting dalam statistik adalah antara signifikansi statistik dan signifikansi praktis.

- **Signifikansi Statistik**: Menilai apakah efek yang diamati kemungkinan besar bukan terjadi karena kebetulan (berdasarkan nilai p atau p-value).
- **Signifikansi Praktis**: Mengevaluasi apakah besaran efek tersebut cukup besar untuk memiliki makna dalam konteks dunia nyata.

Contoh Ilustratif

Misalkan sebuah program diet baru diuji pada satu juta peserta dan menghasilkan penurunan berat badan rata-rata sebesar 0,5 pon (sekitar 0,23 kg) selama satu tahun, dengan nilai p yang sangat kecil (misalnya, $p < 0{,}000001$).

- Signifikan secara statistik? Ya — efek tersebut kemungkinan besar bukan terjadi karena kebetulan.
- Signifikan secara praktis? Tidak — efek tersebut terlalu kecil untuk dianggap bermakna.

Wawasan Utama

Nilai p yang kecil menunjukkan bukti yang kuat, bukan dampak yang besar. Praktik terbaik:

- Selalu laporkan ukuran efek (*effect size*) bersamaan dengan nilai p (*p-values*)

- Interpretasikan hasil dalam konteks kepentingan di dunia nyata

12.9 SELANG KEPERCAYAAN DAN PENGUJIAN HIPOTESIS

Selang kepercayaan (*confidence interval*) dan pengujian hipotesis saling berkaitan erat dan sering kali bersifat melengkapi. Dalam banyak kasus, selang kepercayaan memberikan informasi yang lebih kaya dibandingkan pengujian hipotesis saja.

Hubungan Utama

Untuk uji dua arah (*two-tailed test*) pada tingkat signifikansi α:

- Jika selang kepercayaan (pada tingkat kepercayaan 1-α) mencakup nilai hipotesis nol → Gagal menolak H_0
- Jika selang tersebut tidak mencakup nilai hipotesis nol → Tolak H_0

Contoh 1: Menguji apakah μ = 100 pada α = 0,05.

Selang kepercayaan 95% Anda: (97, 103)

- Karena 100 berada di dalam selang ini → gagal menolak H_0 pada α = 0,05
- Nilai p (p-value) akan > 0,05

Contoh 2: Menguji apakah μ = 100 pada α = 0,05.

Selang kepercayaan 95% Anda: (105, 115)

- Karena 100 TIDAK berada di dalam selang ini → tolak H_0 pada α = 0,05
- Nilai p (p-value) akan < 0,05

Mengapa Hubungan Ini Berlaku

Selang kepercayaan merepresentasikan semua nilai parameter yang konsisten dengan data yang diamati pada tingkat kepercayaan tertentu. Jika nilai hipotesis nol berada di dalam rentang ini, maka nilai tersebut dianggap masuk akal (*plausibel*); jika tidak, maka hipotesis nol ditolak.

Menggunakan Selang Kepercayaan untuk Uji Dua Arah

Langkah 1: Hitung Selang Kepercayaan

Selang Kepercayaan = Estimasi Titik ± Margin Kesalahan (*Margin of Error*).

Untuk rata-rata populasi dengan σ yang tidak diketahui:

$$\mathrm{CI} = \bar{x} \pm t_{\alpha/2,\,df} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

Langkah 2: Periksa Apakah Nilai Hipotesis Nol Berada di Dalam Selang Tersebut

Langkah 3: Ambil Keputusan dan Interpretasikan Hasilnya

Contoh Pengerjaan

Pertanyaan Pembelajaran: Apakah metode pengajaran baru mengubah nilai tes?

Rata-rata historis: $\mu_0 = 75$

Data Sampel:

- n = 25
- x^- = 78,5
- s = 8,2

- α = 0,05

Langkah 1: Hitung Komponen

$$(CI = \bar{x} \pm ME)$$

$\bar{x}$ = 78,5

ME = CONFIDENCE.T(alpha, standard_dev, size)
= CONFIDENCE.T(0,05; 8,2; 25)
= 3,38

Langkah 2: Susun Interval Kepercayaan (*Confidence Interval*)

- Batas bawah: 78,5 − 3,38 = 75,12
- Batas atas: 78,5 + 3,38 = 81,88
  CI 95%: (75,12; 81,88)

Langkah 3: Keputusan

Karena nilai nol (75) berada di luar interval tersebut, kita menolak H_0 pada α = 0,05.

Langkah 4: Interpretasi

Kita memiliki tingkat kepercayaan 95% bahwa nilai rata-rata yang sebenarnya berada di antara 75,12 dan 81,88. Hal ini menunjukkan bahwa metode baru tersebut meningkatkan skor, dengan estimasi peningkatan sekitar 3,5 poin.

Rekomendasi Praktik Terbaik

- Laporkan interval kepercayaan bersamaan dengan nilai p (p-value)
- Interpretasikan interval menggunakan bahasa yang sederhana dan mudah dipahami
- Pertimbangkan lebar interval untuk menilai presisi
- Evaluasi signifikansi praktis dengan melihat rentang nilai yang mungkin (*plausible values*)

Contoh Pelaporan yang Efektif

"Program baru ini meningkatkan produktivitas dengan rata-rata 12%, CI 95% (8%, 16%), dan efek ini signifikan secara statistik $(t(48) = 3{,}21;\ p = 0{,}002)$. Hal ini menunjukkan adanya peningkatan yang berarti bagi operasional bisnis."

Ringkasan Implementasi Excel

Tabel 12.8 Fungsi Excel yang Relevan

Kuantitas	**Rumus Excel**
Rerata Sampel	=AVERAGE(range)
Simpangan Baku Sampel	=STDEV.S(range)
Ukuran Sampel	=COUNT(range)
Margin Kesalahan	=CONFIDENCE.T(alpha, standard_dev, size)
Batas Bawah IK	=mean - ME
Batas Atas IK	=mean + ME

12.10 PEMILIHAN UJI STATISTIK BERDASARKAN JENIS DATA DAN DESAIN PEMBELAJARAN

Pengujian hipotesis bukanlah proses yang berlaku untuk semua situasi. Pertanyaan pembelajaran yang berbeda memerlukan perangkat statistik yang berbeda pula. Memilih uji yang tepat ibarat memilih alat yang pas untuk tugas tertentu: menggunakan alat yang salah mungkin tetap menghasilkan sesuatu, namun dapat berujung pada kesimpulan yang tidak efisien, menyesatkan, atau tidak valid. Oleh karena itu, pemilihan uji statistik yang tepat merupakan langkah krusial dalam setiap analisis data. Hal ini memastikan bahwa kesimpulan Anda akurat, bermakna, dan selaras dengan tujuan pembelajaran Anda.

Faktor-Faktor Utama yang Perlu Dipertimbangkan

Saat memilih uji statistik, pertimbangkan hal-hal berikut:

- **Jenis Data**: Apakah variabel Anda bersifat numerik (contoh: tinggi badan, pendapatan) atau kategorikal (contoh: jenis kelamin, ya/tidak, wilayah)?
- **Jumlah Kelompok atau Sampel**: Apakah Anda menganalisis satu kelompok, dua kelompok, atau tiga kelompok atau lebih?
- **Hubungan Antarsampel**: Apakah kelompok-kelompok tersebut bersifat independen atau berpasangan (dependen)?
- **Parameter yang Diamati**: Apakah Anda menguji rata-rata, proporsi, varians, atau hubungan antarvariabel?
- **Asumsi Data**: Apakah data terdistribusi secara normal? Apakah varians antar-kelompok bersifat serupa?

Faktor-faktor ini menjadi dasar dalam menentukan uji statistik yang sesuai.

Uji Hipotesis Satu Sampel

Uji satu sampel membandingkan satu sampel dengan nilai populasi yang telah diketahui atau dihipotesiskan.

Contoh penggunaan umum:

- Pengendalian kualitas: Apakah produk yang dihasilkan memenuhi standar yang ditetapkan?
- Tolok ukur kinerja: Apakah suatu sistem menunjukkan kinerja yang berbeda dari standar industri?
- Evaluasi akademik: Apakah nilai mahasiswa berbeda dari rata-rata yang diharapkan?

Dalam kasus-kasus ini, tujuannya adalah untuk menentukan apakah sampel memberikan bukti yang cukup bahwa parameter populasi berbeda dari nilai yang telah diketahui.

Uji Hipotesis Dua Sampel

Uji dua sampel membandingkan dua kelompok untuk menentukan apakah terdapat perbedaan yang signifikan di antara keduanya.

Skenario umum:

- Perlakuan (treatment) vs. kontrol: Membandingkan hasil antara dua intervensi.
- Sebelum vs. sesudah: Mengukur perubahan dari waktu ke waktu.
- Perbandingan metode: Mengevaluasi pendekatan alternatif.
- Perbedaan antar-kelompok: Membandingkan karakteristik di antara populasi yang berbeda.

Pertimbangan utama di sini adalah apakah sampel bersifat independen atau berpasangan, karena hal ini akan menentukan uji statistik yang tepat untuk digunakan.

Sampel Independen vs. Sampel Berpasangan

1. Sampel Independen

Definisi: Kedua kelompok tidak saling berkaitan, dan observasi dalam satu kelompok tidak memiliki hubungan dengan observasi di kelompok lainnya.

Karakteristik utama:

- Individu yang berbeda di setiap kelompok
- Tidak ada pasangan alami antar observasi
- Ukuran sampel bisa berbeda
- Observasi bersifat independen (bebas satu sama lain)

Contoh:

- Membandingkan rata-rata tinggi badan pria dan wanita
- Menguji dua jenis obat pada kelompok pasien yang berbeda
- Membandingkan kinerja mahasiswa di sekolah yang berbeda

2. Sampel Berpasangan (Dependen)

Definisi: Observasi saling terkait secara bermakna, baik karena berasal dari subjek yang sama maupun karena sengaja dipasangkan.

Karakteristik utama:

- Subjek yang sama diukur dua kali, atau berupa pasangan yang disesuaikan (*matched pairs*)
- Setiap observasi memiliki pasangan alami
- Ukuran sampel sama (berpasangan)
- Observasi bersifat dependen (saling bergantung)

Contoh:

- Mengukur tekanan darah sebelum dan sesudah pengobatan
- Membandingkan hasil pada pasangan yang disesuaikan (misalnya, anak kembar atau peserta yang dipasangkan berdasarkan karakteristik tertentu)

Memahami perbedaan ini sangat penting karena akan menentukan apakah harus menggunakan uji untuk sampel independen atau sampel berpasangan.

Tiga Kelompok atau Lebih: ANOVA (Uji-F)

Saat membandingkan tiga kelompok atau lebih, melakukan beberapa uji-t secara terpisah akan meningkatkan risiko terjadinya kesalahan Tipe I (*Type I error*). Sebagai gantinya, digunakan Analisis Varians (ANOVA) yang didasarkan pada uji-F.

Tujuan: Untuk menentukan apakah setidaknya terdapat satu rata-rata kelompok yang berbeda dari kelompok lainnya.

Contoh skenario:

- Apakah berbagai metode pengajaran menghasilkan capaian mahasiswa yang berbeda?
- Apakah tingkat kepuasan pelanggan berbeda di beberapa cabang perusahaan?

Hasil ANOVA yang signifikan menunjukkan bahwa setidaknya terdapat satu kelompok yang berbeda, namun diperlukan uji lanjutan (*post-hoc test*) untuk mengidentifikasi kelompok mana saja yang memiliki perbedaan tersebut.

Data Kategorikal: Uji Chi-Square (χ^2)

Saat menangani data kategorikal, pengujian hipotesis berfokus pada frekuensi (jumlah kejadian) alih-alih rata-rata numerik.

Uji Chi-Square yang umum digunakan:

- **Uji Kesesuaian (*Goodness-of-Fit Test*)**: Menentukan apakah frekuensi yang teramati sesuai dengan frekuensi yang diharapkan.
  Contoh: Apakah preferensi pelanggan terdistribusi secara merata di antara berbagai pilihan produk?
- **Uji Independensi (*Test of Independence*)**:
  Menilai apakah dua variabel kategorikal saling berkaitan.
  Contoh: Apakah terdapat hubungan antara jenis kelamin dan preferensi produk?

Memilih Uji yang Tepat: Panduan Praktis

Panduan langkah demi langkah berikut dapat membantu Anda memilih uji yang sesuai:

Langkah 1: Identifikasi Jenis Data

- Data kategorikal → Gunakan uji χ^2 (chi-square)
- Data numerik → Lanjutkan ke Langkah 2

Langkah 2: Tentukan Jumlah Kelompok atau Variabel

- Satu kelompok → Lanjutkan ke Langkah 3
- Dua kelompok → Lanjutkan ke Langkah 4
- Tiga kelompok atau lebih → Gunakan ANOVA (uji-F)
- Dua variabel numerik → Gunakan korelasi atau regresi

Langkah 3: Uji Satu Sampel (Rata-rata)

- Jika σ diketahui dan ukuran sampel besar → Uji-z satu sampel
- Jika σ tidak diketahui → Uji-t satu sampel

Langkah 4: Uji Dua Sampel (Rata-rata)

- Sampel independen → Uji-t dua sampel independen
- Sampel berpasangan → Uji-t sampel berpasangan

Uji untuk Proporsi

- Satu kelompok → Uji-z satu sampel untuk proporsi
- Dua kelompok → Uji-z dua sampel untuk proporsi

Wawasan Akhir

Kerangka kerja ini mencakup uji statistik yang paling sering digunakan. Meskipun bidang statistika sangat luas, menguasai proses pengambilan keputusan ini akan memberikan landasan yang kuat untuk analisis yang lebih lanjut.

Selalu ingat: Uji statistik harus sesuai dengan data, dan data harus sesuai dengan pertanyaan pembelajaran.

12.11 UJI HIPOTESIS STATISTIK: PROSEDUR, KESALAHAN, DAN INTERPRETASI HASIL

Dalam bab ini, kita telah membahas pengujian hipotesis sebagai kerangka kerja mendasar untuk mengambil keputusan berbasis data mengenai populasi dengan menggunakan bukti dari sampel. Berdasarkan konsep-konsep sebelumnya seperti statistika deskriptif, probabilitas, pengambilan sampel, dan estimasi parameter populasi kita telah mempelajari cara menguji klaim secara formal serta mengevaluasi apakah hasil yang diamati mungkin terjadi karena kebetulan atau mencerminkan efek nyata yang mendasarinya.

Kita mengawalinya dengan mendefinisikan hipotesis sebagai pernyataan yang jelas dan dapat diuji mengenai parameter populasi, serta menunjukkan bagaimana pertanyaan pembelajaran diterjemahkan ke dalam klaim statistik. Hal ini mengarah pada perumusan hipotesis nol (H_0) dan hipotesis alternatif (H_1), yang bersama-sama menjadi landasan bagi seluruh pengujian hipotesis. Perlu ditekankan bahwa pengujian hipotesis bukanlah upaya untuk membuktikan suatu klaim dengan kepastian mutlak, melainkan untuk menilai bukti dan mengambil keputusan di tengah ketidakpastian.

Selanjutnya, kita menelaah logika pengujian hipotesis yang berakar pada metode ilmiah, serta memperkenalkan konsep-konsep kunci seperti tingkat signifikansi (α), statistik uji, nilai kritis, dan nilai-p (p-value). Elemen-elemen ini bekerja bersama untuk membantu menentukan apakah bukti dari sampel cukup kuat untuk menolak hipotesis nol. Kita juga membahas peran wilayah penolakan (*rejection regions*) dan bagaimana keputusan dapat berbeda bergantung pada apakah uji tersebut merupakan uji satu arah (sisi kiri atau sisi kanan) atau uji dua arah. Kemudian, kita menyajikan prosedur pengujian hipotesis enam langkah—sebuah pendekatan terstruktur yang memandu Anda mulai dari merumuskan hipotesis hingga mengambil serta menafsirkan keputusan dalam konteks tertentu. Baik pendekatan nilai-p maupun pendekatan nilai kritis telah diperkenalkan; keduanya akan menghasilkan kesimpulan yang sama, namun menawarkan perspektif yang berbeda mengenai bukti statistik.

Bagian penting lainnya dalam bab ini berfokus pada pemahaman mengenai kesalahan dalam pengambilan keputusan. Kita membedakan antara kesalahan Tipe I (positif palsu) dan kesalahan Tipe II (negatif palsu), serta memperkenalkan konsep kekuatan statistik (*statistical power*, 1 − β) sebagai ukuran kemampuan suatu uji untuk mendeteksi efek yang nyata. Kita juga mendiskusikan implikasi praktis dari pemilihan tingkat signifikansi yang berbeda serta kompromi yang tak terelakkan dalam upaya meminimalkan kesalahan-kesalahan tersebut. Kami telah memperjelas perbedaan penting antara signifikansi statistik dan signifikansi praktis, dengan menekankan bahwa hasil yang signifikan secara statistik tidak selalu bermakna dalam konteks dunia nyata. Hal ini menegaskan perlunya mempertimbangkan ukuran efek (*effect size*) di samping nilai-p (*p-value*) saat menafsirkan hasil pembelajaran.

Hubungan antara selang kepercayaan (*confidence intervals*) dan pengujian hipotesis juga telah dibahas. Kami menunjukkan bagaimana selang kepercayaan memberikan rentang nilai yang mungkin bagi suatu parameter dan bagaimana selang tersebut dapat digunakan sebagai metode alternatif untuk pengambilan keputusan. Kaitan ini memperdalam pemahaman serta menyajikan cara yang lebih informatif dalam mengomunikasikan hasil pembelajaran. Terakhir, kami membahas cara memilih uji statistik yang tepat berdasarkan

karakteristik data, jumlah kelompok, dan tujuan pembelajaran. Dengan memahami perbedaan antara uji satu sampel, dua sampel, dan banyak kelompok serta perbedaan antara data numerik dan kategorikal Anda kini memiliki kerangka kerja praktis untuk menentukan pendekatan analisis yang tepat. Di akhir bab ini, Anda telah membangun landasan konseptual dan praktis yang kuat dalam pengujian hipotesis. Anda kini siap untuk menerapkan prinsip-prinsip tersebut pada masalah di dunia nyata, menafsirkan hasil secara bertanggung jawab, serta mengambil keputusan yang tepat berdasarkan bukti sebagai seorang ilmuwan data (*data scientist*).

BAB 13
PROSEDUR STATISTIK UNTUK PENGUJIAN HIPOTESIS

13.1 PEMILIHAN DAN PENERAPAN UJI STATISTIK PARAMETRIK DAN NONPARAMETRIK

Pada bab sebelumnya, kita telah menetapkan prinsip-prinsip dasar pengujian hipotesis, termasuk perumusan hipotesis nol dan hipotesis alternatif, peran tingkat signifikansi, interpretasi nilai-p (p-value), serta implikasi dari kesalahan Tipe I dan Tipe II. Berlandaskan prinsip-prinsip tersebut, kini kita beralih pada penerapan praktis dari konsep-konsep ini melalui prosedur statistik tertentu.

Memilih uji statistik yang tepat sangatlah penting untuk menarik kesimpulan yang valid dan dapat diandalkan dari data. Pemilihan yang keliru dapat menghasilkan kesimpulan yang menyesatkan, penggunaan sumber daya yang tidak efisien, dan pengambilan keputusan yang cacat. Sebagai panduan, uji statistik secara garis besar dibagi menjadi dua kategori utama:

- **Uji Parametrik**: Uji ini mengasumsikan bahwa data mengikuti distribusi tertentu paling umum adalah distribusi normal dan biasanya diterapkan pada data berskala interval atau rasio. Jika asumsi-asumsinya terpenuhi, uji parametrik sangat efisien dan memiliki kekuatan statistik yang tinggi. Contoh yang umum meliputi uji-z, uji-t, uji-F, analisis varians (ANOVA), dan analisis regresi.
- **Uji Non-Parametrik**: Uji ini memerlukan lebih sedikit asumsi mengenai distribusi populasi dan cocok digunakan untuk data ordinal atau data yang tidak berdistribusi normal. Uji ini menjadi alternatif yang tangguh (*robust*) ketika asumsi-asumsi parametrik tidak terpenuhi. Contohnya meliputi uji Mann–Whitney U, uji peringkat bertanda Wilcoxon (*Wilcoxon signed-rank test*), uji Kruskal–Wallis, dan uji chi-square untuk independensi.

Kedua kategori ini secara bersama-sama menyediakan kerangka kerja yang fleksibel untuk menganalisis berbagai jenis data dan skenario pembelajaran.

Meskipun terdapat banyak sekali uji statistik, bab ini berfokus pada empat prosedur dasar yang menjadi tulang punggung dalam sebagian besar alur kerja analisis data: uji-z, uji-t, uji chi-square, dan uji-F. Sementara uji-z, uji-t, dan uji-F termasuk dalam kelompok uji parametrik, uji chi-square menghadirkan pendekatan yang bebas distribusi (*distribution-free*), yang menjadi pintu masuk ke metode non-parametrik. Bab ini menyajikan panduan praktis dan terstruktur mengenai uji-uji penting tersebut, dengan menekankan pada kapan harus menggunakan setiap metode, asumsi apa saja yang harus dipenuhi, serta bagaimana cara menginterpretasikan hasilnya dalam konteks dunia nyata.

13.2 GAMBARAN UMUM UJI HIPOTESIS YANG UMUM DIGUNAKAN

Penting untuk dipahami bahwa setiap uji statistik utama uji-z, uji-t, uji chi-square (χ^2), dan uji-F memiliki beberapa variasi khusus, yang masing-masing dirancang untuk menjawab jenis pertanyaan pembelajaran tertentu. Memahami variasi-variasi ini sangat penting untuk menentukan uji yang tepat dalam praktik.

Uji-Z (Z-Tests)

Uji-z didasarkan pada distribusi normal standar dan biasanya digunakan ketika simpangan baku populasi (σ) diketahui serta ukuran sampel tergolong besar.

Variasi yang umum meliputi:

- Uji-z satu sampel untuk rata-rata: Membandingkan rata-rata sampel dengan rata-rata populasi yang diketahui atau dihipotesiskan (σ diketahui, ukuran sampel besar).
- Uji-z satu sampel untuk proporsi: Membandingkan proporsi sampel dengan proporsi populasi yang dihipotesiskan.
- Uji-z dua sampel untuk rata-rata (sampel independen): Membandingkan rata-rata dari dua sampel besar yang independen (σ diketahui).
- Uji-z dua sampel untuk proporsi: Membandingkan proporsi antara dua kelompok yang independen.

Uji-t (T-Tests)

Uji-t didasarkan pada distribusi-t Student dan digunakan ketika simpangan baku populasi tidak diketahui, terutama pada ukuran sampel yang kecil. Uji ini memperhitungkan ketidakpastian tambahan yang muncul akibat variabilitas sampel.

Variasi utamanya meliputi:

- Uji-t satu sampel (*one-sample t-test*): Membandingkan rata-rata sampel dengan rata-rata populasi yang dihipotesiskan ketika σ (simpangan baku populasi) tidak diketahui.
- Uji-t dua sampel independen (*independent two-sample t-test*): Membandingkan rata-rata dari dua kelompok yang saling bebas:
  - Asumsi varians sama (*pooled t-test*)
  - Asumsi varians tidak sama (*Welch's t-test*)
- Uji-t sampel berpasangan (*paired-sample t-test*/sampel dependen): Membandingkan rata-rata dari kelompok yang sama yang diukur pada dua titik waktu berbeda atau dari pasangan yang telah dicocokkan (*matched pairs*).

Uji Chi-Kuadrat (χ^2)

Uji chi-kuadrat (χ^2) dirancang untuk data kategorikal dan berfokus pada hubungan atau perbedaan dalam distribusi frekuensi. Berbeda dengan uji-z dan uji-t yang menganalisis data numerik (rata-rata dan proporsi), uji χ^2 membandingkan frekuensi yang teramati dengan frekuensi yang diharapkan.

Tiga variasi utamanya adalah:

- Uji Kecocokan (*Goodness-of-Fit Test*): Menilai apakah suatu variabel kategorikal tunggal mengikuti distribusi tertentu.
- Uji Independensi (*Test of Independence*): Menentukan apakah dua variabel kategorikal saling berhubungan atau independen.
- Uji Homogenitas (*Test of Homogeneity*): Mengevaluasi apakah distribusi suatu variabel kategorikal sama di berbagai kelompok.

Semua uji chi-kuadrat bersifat satu arah ke kanan (*right-tailed*) dan bergantung pada besarnya penyimpangan antara frekuensi yang teramati dan frekuensi yang diharapkan, yang diringkas dalam statistik χ^2.

Uji-F (F-Tests)

Uji-F digunakan untuk membandingkan varians atau mengevaluasi perbedaan di antara rata-rata beberapa kelompok. Uji ini didasarkan pada rasio dua varians dan mengikuti distribusi-F, yang bernilai non-negatif dan umumnya menceng ke kanan (*right-skewed*).
Penerapan yang umum meliputi:

- Uji Kesamaan Varians (*Test for Equality of Varians*): Menentukan apakah dua populasi memiliki varians yang sama.
- Analisis Varians (ANOVA): Menguji apakah rata-rata dari tiga kelompok atau lebih berbeda secara signifikan.
- Uji-F Signifikansi Regresi: Menilai apakah model regresi mampu menjelaskan sebagian besar variabilitas dalam variabel dependen. Nilai F yang lebih besar menunjukkan bukti yang lebih kuat untuk menolak hipotesis nol.

Pada bagian-bagian selanjutnya, setiap pengujian ini akan dibahas secara mendetail, dengan fokus pada kapan harus menggunakannya, bagaimana cara menghitungnya, serta bagaimana menafsirkan hasilnya melalui contoh-contoh praktis di dunia nyata.

Semua pengujian hipotesis dalam bab ini akan dilakukan pada tingkat signifikansi α = 0,05. Ambang batas ini mencerminkan keseimbangan yang diterima secara luas antara risiko melakukan kesalahan Tipe I dan kebutuhan untuk mendeteksi efek yang bermakna dalam data. Namun, perlu dipahami bahwa pemilihan tingkat signifikansi tidak bersifat universal. Dalam praktiknya, pemilihan ini harus didasarkan pada konteks masalah, konsekuensi dari kesalahan pengambilan keputusan, serta standar yang berlaku dalam bidang terkait. Oleh karena itu, Anda disarankan untuk meninjau kembali contoh-contoh praktis dalam bab ini dengan menggunakan tingkat signifikansi yang berbeda (seperti α = 0,01 atau α = 0,10) dan mengamati bagaimana kesimpulannya mungkin berubah.

13.3 UJI-Z (*Z-TEST*): PENERAPAN PRAKTIS

Uji-Z digunakan untuk menguji hipotesis mengenai rata-rata atau proporsi populasi ketika kondisi tertentu terpenuhi terutama jika simpangan baku populasi diketahui, atau jika ukuran sampel cukup besar sehingga Teorema Limit Pusat (*Central Limit Theorem*) dapat diterapkan. Pengujian ini mengandalkan distribusi normal standar untuk menentukan signifikansi statistik.

a) Uji-Z Satu Sampel untuk Rata-rata (One-Sample Z-Test for the Mean)

Pengujian ini digunakan untuk menentukan apakah rata-rata sampel berbeda secara signifikan dari rata-rata populasi yang diketahui atau dihipotesiskan, dengan ketentuan bahwa simpangan baku populasi (σ) diketahui dan ukuran sampel cukup besar ($n \geq 30$).
Contoh: Pengendalian Kualitas dalam Skenario Manufaktur:

Sebuah pabrik cokelat memproduksi cokelat batangan yang seharusnya memiliki berat tepat 100 gram. Proses produksi tersebut memiliki simpangan baku yang diketahui sebesar 8 gram. Seorang pemeriksa pengendalian kualitas mengambil sampel acak sebanyak 36 batang cokelat dan menemukan rata-rata berat sebesar 97,5 gram. Apakah proses produksi tersebut melenceng dari target?

Langkah 1: Menyatakan Hipotesis

- H_0: μ = 100 (proses produksi sesuai target)
- H_1: μ ≠ 100 (proses produksi melenceng dari target) Ini merupakan uji dua arah (*two-tailed test*).

Langkah 2: Tetapkan Tingkat Signifikansi

α = 0,05

Langkah 3: Periksa Kondisi

- Uji untuk rata-rata ✓
- Satu sampel ✓
- σ diketahui (σ = 8) ✓
- Sampel besar (n = 36 ≥ 30) ✓

Kondisi terpenuhi untuk uji-z satu sampel bagi rata-rata.

Langkah 4: Hitung Statistik Uji

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Di mana:

- $\bar{x} = 97.5$
- $\mu_0 = 100$
- $\sigma = 8$
- $n = 36$

Kesalahan Standar (SE):

$$SE = \frac{8}{\sqrt{36}} = \frac{8}{6} = 1.33$$

Statistik-Z:

$$z = \frac{97.5 - 100}{1.33} = -1.88$$

Langkah 5: Menentukan Nilai Kritis dan *p-value*

- Nilai kritis (dua arah, α = 0,05): ±1,96
  Dapat dihitung menggunakan =NORM.S.INV(1- α /2)
- p-value = 2*NORM.S.DIST(-1,88, TRUE) = 0,0602

Langkah 6: Keputusan dan Kesimpulan

- Pendekatan nilai kritis: |−1,88| < 1,96 → Gagal menolak H_0
- Pendekatan *p-value*: 0,0602 > 0,05 → Gagal menolak H_0

Kesimpulan: Tidak terdapat cukup bukti untuk menyimpulkan bahwa proses produksi melenceng dari target.

Tabel 13.1 Nilai Kritis Z yang Umum Digunakan (sebagai referensi cepat)

α	Ekor kiri	Ekor kanan	Dua arah
0,01	−2,326	2,326	±2,576
0,05	−1,645	1,645	±1,96
0,10	−1,28	1,28	±1,645

b) Uji-Z Satu Sampel untuk Proporsi

Digunakan untuk menguji apakah proporsi sampel berbeda dari proporsi populasi yang dihipotesiskan, biasanya ketika:

- Sampel berukuran besar, dan
- ($np_0 \geq 5$ dan $n(1-p_0) \geq 5$)

Contoh: Skenario Tingkat Konversi Situs Web:

Sebuah situs web e-commerce memiliki tingkat konversi historis sebesar 12%. Setelah desain ulang, 75 dari 500 pengunjung melakukan pembelian. Apakah tingkat konversi tersebut meningkat?

Langkah 1: Nyatakan Hipotesis

- $H_0: p = 0{,}12$
- $H_1: p > 0{,}12$ (uji satu arah kanan)

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

- Uji untuk proporsi ✓
- Satu sampel ✓
- $n = 500$ ✓
- $\hat{p} = 0{,}15$
- $np_0 = 60 \geq 5$ ✓
- $n(1 - p_0) = 440 \geq 5$ ✓

Kondisi terpenuhi untuk uji-Z satu sampel untuk proporsi

Langkah 4: Hitung Statistik Uji

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

$$SE = \sqrt{\frac{0.12 \times 0.88}{500}} = 0.0145$$

$$z = \frac{0.15 - 0.12}{0.0145} = 2.07$$

Langkah 5: Tentukan Nilai Kritis dan p-value

- Nilai kritis untuk uji sisi kanan pada α = 0,05: 1,645
  Dapat dihitung dengan rumus =NORM.S.INV(1- α)

- p-value: =1-NORM.S.DIST(2,07, TRUE) ≈ 0,0192

Langkah 6: Keputusan

- z = 2,07 > 1,645 → Tolak H_0
- p-value < 0,05 → Tolak H_0

Kesimpulan: Terdapat bukti yang cukup bahwa desain situs web yang baru meningkatkan tingkat konversi hingga di atas 12%.

c) Uji Z Dua Sampel untuk Rata-rata (Sampel Independen)

Digunakan untuk membandingkan rata-rata dari dua kelompok independen ketika:

- Simpangan baku populasi diketahui, atau
- Ukuran kedua sampel besar ($n_1,\ n_2 \geq 30$)

Contoh: Membandingkan Metode Pengajaran Skenario:

Sebuah sekolah ingin membandingkan dua metode pengajaran. Metode A diterapkan pada 40 mahasiswa (skor rata-rata: 78, σ = 12), dan Metode B diterapkan pada 35 mahasiswa (skor rata-rata: 82, σ = 10). Apakah terdapat perbedaan yang signifikan di antara kedua metode tersebut?

Langkah 1: Nyatakan Hipotesis

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

Metode A: $n_1 = 40, \bar{x}_1 = 78, \sigma_1 = 12$

Metode B: $n_2 = 35, \bar{x}_2 = 82, \sigma_2 = 10$

- Uji untuk rata-rata
- Dua sampel independen ✓
- Simpangan baku populasi diketahui ✓
- Sampel berukuran besar ✓

Kondisi terpenuhi untuk melakukan uji Z dua sampel bagi rata-rata.

Langkah 4: Hitung Statistik Uji

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$SE = \sqrt{\frac{144}{40} + \frac{100}{35}} = \sqrt{6.46} = 2.54$$

$$z = \frac{78-82}{2.54} = -1.574$$

Langkah 5: Tentukan Nilai Kritis dan Nilai-p (p-value)

- Nilai kritis dua arah (*two-tailed*): ±1,96
- Nilai-p: =2*NORM.S.DIST(-1,574, TRUE) = 0,115

Langkah 6: Keputusan dan Kesimpulan

Gagal menolak H_0.

Kesimpulan: Tidak terdapat cukup bukti untuk menyimpulkan bahwa kedua metode pengajaran tersebut memiliki tingkat efektivitas yang berbeda.

d) Uji Z Dua Sampel untuk Proporsi

Digunakan untuk membandingkan proporsi dari dua kelompok independen apabila:

- Kedua sampel berukuran besar, dan
- $\hat{n}ip \geq 5$ serta $ni(1-\hat{p}) \geq 5$ untuk setiap kelompok

Contoh: Skenario Perbandingan Kampanye Pemasaran

Perusahaan A menguji dua kampanye pemasaran melalui surel (email). Kampanye 1 dikirim kepada 800 orang, dengan 96 orang mengeklik tautan tersebut. Kampanye 2 dikirim kepada 750 orang, dengan 105 orang mengeklik tautan tersebut. Apakah terdapat perbedaan yang signifikan dalam tingkat klik (*click-through rates*)?

Langkah 1: Nyatakan Hipotesis

- $H_0: p_1 = p_2$
- $H_1: p_1 \neq p_2$

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

a) Hitung proporsi sampel dan proporsi gabungan (pooled proportion).

Kampanye 1: 96 klik dari 800 orang → $\hat{p}_1 = 0{,}12$

Kampanye 2: 105 klik dari 750 orang → $\hat{p}_2 = 0{,}14$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{96 + 105}{800 + 750} = 0.1297$$

b) Periksa Kondisi

- Uji untuk proporsi ✓
- Dua sampel ✓
- $n_1\hat{p} = 800 \times 0{,}1297 = 103{,}76 \geq 5$ ✓
- $n_1(1-\hat{p}) = 800 \times 0{,}8703 = 696{,}24 \geq 5$ ✓
- $n_2\hat{p} = 750 \times 0{,}1297 = 97{,}28 \geq 5$ ✓
- $n_2(1-\hat{p}) = 750 \times 0{,}8703 = 652{,}72 \geq 5$ ✓

Kondisi terpenuhi untuk uji-z dua sampel bagi proporsi.

Langkah 4: Statistik Uji

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.12 - 0.14}{\sqrt{0.1297(1-0.1297)\left(\frac{1}{800} + \frac{1}{750}\right)}} = \frac{-0.02}{\sqrt{0.1123 \times 0.0026}}$$

$$z = -1.17$$

Langkah 5: Menentukan Nilai Kritis dan Nilai-p (p-value)

- Nilai kritis untuk uji dua arah (α = 0,05): ±1,96
- Nilai-p = 2*NORM.S.DIST(-1,17, TRUE) = 0,2420

Langkah 6: Keputusan dan Kesimpulan

Gagal menolak H_0.

Kesimpulan: Tidak terdapat cukup bukti untuk menyimpulkan bahwa kedua kampanye memiliki tingkat klik-tayang (*click-through rate*) yang berbeda.

Menggunakan Data Analysis ToolPak di Excel

Excel menyediakan cara yang praktis untuk melakukan uji-z dua sampel bagi rata-rata (*two-sample z-test for means*).

Langkah-langkah:

1. Susun data ke dalam dua kolom
2. Buka: Data → Data Analysis → Z-Test: Two Sample for Means
3. Masukkan parameter yang diperlukan:
   - Rentang variabel
   - Perbedaan rata-rata yang dihipotesiskan (biasanya 0)
   - Varians yang diketahui
   - Tingkat signifikansi (α)
4. Pilih lokasi output
5. Interpretasikan hasil:
   - Statistik-z
   - Nilai-p (uji satu arah dan dua arah)
   - Nilai kritis

Aturan Keputusan:

- Jika nilai-p ≤ α → Tolak H_0
- Jika nilai-p > α → Gagal menolak H_0

Pendekatan terstruktur ini memastikan bahwa uji-z diterapkan dengan benar dan diinterpretasikan dengan penuh keyakinan dalam skenario analisis data di dunia nyata.

13.4 UJI-T (*T-TESTS*): PENERAPAN PRAKTIS

Uji-t digunakan untuk membandingkan rata-rata sampel ketika standar deviasi populasi tidak diketahui. Uji ini sangat berguna untuk ukuran sampel kecil, di mana ketidakpastian tambahan harus diperhitungkan. Berbeda dengan uji-z yang mengandalkan distribusi normal standar, uji-t menggunakan distribusi-t Student, yang menyesuaikan variabilitas sampel melalui derajat kebebasan (*degrees of freedom*/df).

a) Uji-t Satu Sampel (One-Sample T-Test)

Uji ini digunakan untuk memeriksa apakah rata-rata sampel berbeda dari rata-rata populasi yang dihipotesiskan, dengan kondisi:

- Standar deviasi populasi tidak diketahui, dan
- Ukuran sampel kecil (biasanya $n < 30$).

Contoh: Skenario Waktu Layanan di Kedai Kopi:
Seorang manajer mengklaim bahwa rata-rata waktu layanan adalah 3 menit. Seorang pelanggan mencatat waktu layanan untuk 12 pesanan yang dipilih secara acak:

$$[2.5,\ 3.2,\ 2.8,\ 3.5,\ 2.9,\ 3.1,\ 2.7,\ 3.4,\ 2.6,\ 3.0,\ 3.3,\ 2.8]$$

Apakah rata-rata waktu layanan yang sebenarnya berbeda dari 3 menit?
Langkah 1: Hipotesis

- $H_0: \mu = 3$
- $H_1: \mu \neq 3$ (uji dua arah)

Langkah 2: Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

- Satu sampel ✓
- Uji untuk rata-rata ✓
- σ tidak diketahui ✓
- Sampel kecil ($n = 12$) ✓

Kondisi terpenuhi untuk uji-t satu sampel.
Langkah 4: Hitung Statistik Uji

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Menggunakan Excel:

- $\bar{x} = 2.98$
- $s = 0.32$
- $n = 12, \quad df = 11$

$$SE = \frac{0.32}{\sqrt{12}} = 0.0924$$

$$t = \frac{2.98-3}{0.0924} = -0.216$$

Langkah 5: Nilai Kritis dan p-value

- Nilai kritis (df = 11, α = 0,05): =T.INV.2T(0,05; 11)
  = ±2,201
- p-value: =T.DIST.2T(0,216; 11)
  = 0,833

Langkah 6: Keputusan

- $|t| < 2{,}201$ → Gagal menolak H_0
- p-value > 0,05 → Gagal menolak H_0

Kesimpulan: Tidak terdapat cukup bukti untuk menyimpulkan bahwa rata-rata waktu layanan berbeda dari 3 menit.

b) Uji-t Dua Sampel Independen (*Independent Two-Sample T-Test*)

Asumsi Varians Sama (*Pooled T-Test*)

Digunakan saat membandingkan rata-rata dari dua kelompok independen dengan kondisi:

- Simpangan baku populasi tidak diketahui,
- Varians diasumsikan sama,
- Biasanya ukuran sampel kecil.

Contoh: Membandingkan Metode Belajar.

Skenario:

Seorang guru ingin membandingkan dua metode belajar. Kelompok 1 (metode tradisional) terdiri dari 8 mahasiswa dengan nilai: $[78, 82, 75, 80, 77, 79, 81, 76]$. Kelompok 2 (metode baru) terdiri dari 10 mahasiswa dengan nilai: $[85, 88, 83, 87, 84, 86, 89, 82, 85, 88]$. Asumsikan varians sama. Apakah terdapat perbedaan yang signifikan?

Langkah 1: Hipotesis

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$ (uji dua arah/*two-tailed test*)

Langkah 2: Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

Hitung statistik ringkasan (*summary statistics*) di Excel:

- Group 1 (Traditional): $n_1 = 8, \quad \overline{x_1} = 78.5, \quad s_1^2 = 6$
- Group 2 (New Method): $n_2 = 10, \quad \overline{x_2} = 85.7, \quad s_2^2 = 5.344$

Periksa Kondisi:

- Pengujian untuk nilai rata-rata ✓
- Dua sampel independen ✓
- Ukuran sampel kecil ✓
- Simpangan baku populasi (σ) tidak diketahui ✓
- Varians sama ✓

Kondisi terpenuhi untuk uji-t dua sampel independen (dengan asumsi varians sama)

Langkah 4: Hitung Statistik Uji

Varians gabungan (*pooled variance*):

- Karena kita mengasumsikan varians populasi yang sama, kita menggabungkan (mengumpulkan) varians sampel menggunakan rata-rata tertimbang

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = 5.631$$

$$s_p = \sqrt{s_p^2} = \sqrt{5.631} = 2.374$$

Statistik t

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{78.5 - 85.7}{2.374 \times \sqrt{\frac{1}{8} + \frac{1}{10}}}$$

$$t = \frac{-7.2}{1.125} = -6.40$$

Derajat Kebebasan

$$df = n_1 + n_2 - 2 = 8 + 10 - 2 = 16$$

Langkah 5: Nilai Kritis dan p-value

- Nilai kritis: =T.INV.2T(0.05,16) = ±2,12
- p-value: =T.DIST.2T(6,4;16) = < 0,001

Langkah 6: Keputusan

$Tolak\ H_0$.

Kesimpulan:

Terdapat bukti kuat mengenai perbedaan yang signifikan di antara metode-metode pembelajaran tersebut, dengan metode baru menghasilkan skor yang lebih tinggi.

Melakukan Pengujian menggunakan ToolPak di Excel

Gunakan:

Data → Data Analysis → t-Test: *Two-Sample Assuming Equal Variances* (Uji-t: Dua Sampel dengan Asumsi Varians Sama)

Varians Tidak Sama (Uji-t Welch)

Digunakan untuk membandingkan rata-rata dari dua kelompok independen ketika varians tidak diketahui dan tidak diasumsikan sama.

Contoh: Membandingkan Waktu Reaksi

Skenario: Peneliti membandingkan waktu reaksi antara dua kelompok usia. Dewasa muda ($n_1 = 12$): rata-rata = 0,25 detik, $s_1 = 0{,}05$. Orang dewasa yang lebih tua ($n_2 = 15$): rata-rata = 0,32 detik, $s_2 = 0{,}12$. Uji apakah terdapat perbedaan yang signifikan.

Langkah 1: Hipotesis

- H_0: $\mu_1 = \mu_2$ (tidak ada perbedaan dalam waktu reaksi)
- H_1: $\mu_1 \neq \mu_2$ (terdapat perbedaan dalam waktu reaksi)
- Uji dua arah (*two-tailed test*)

Langkah 2: Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Memeriksa Kondisi

Statistik Ringkasan yang Diberikan

- Kelompok 1: $n_1 = 12, \bar{x}1 = 0{,}25, s1 = 0{,}05$
- Kelompok 2: $n_2 = 15, \bar{x}2 = 0{,}32, s2 = 0{,}12$

Memeriksa kondisi:

- Uji untuk rata-rata ✓
- Dua sampel independen ✓
- Ukuran sampel kecil ✓
- Simpangan baku populasi (σ) tidak diketahui ✓
- Varians tidak sama ✓
- Kondisi terpenuhi untuk uji-t dua sampel independen (dengan varians tidak sama)

Langkah 4: Statistik Uji

Hitung nilai t:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{0.25 - 0.32}{\sqrt{\frac{0.0025}{12} + \frac{0.0144}{15}}}$$

$$t = -2.0479$$

Hitung rumus derajat kebebasan Welch:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Perhitungan:

- df = [(0,0025/12) + (0,0144/15)]² / [(0,0025/12)²/11 + (0,0144/15)²/14]
- df = 1,365x10⁻⁶ / (3,94571x10⁻⁹ + 6,58286x10⁻⁸)
- df = 1,365x10⁻⁶ / 6,97743x10⁻⁸ df = 19,56 ≈ 20

Langkah 5: Nilai Kritis dan p-value

- Nilai kritis: ±2,086
- p-value ≈ 0,054

Langkah 6: Keputusan

Gagal menolak H_0 (kasus yang berada di ambang batas).

Kesimpulan: Bukti menunjukkan adanya perbedaan, namun tidak signifikan secara statistik pada α = 0,05.

Melakukan Pengujian dengan ToolPak Excel

Gunakan:

Data → Data Analysis → t-Test: Two-Sample Assuming Unequal Variances (Uji-t: Dua Sampel dengan Asumsi Varians Tidak Sama)

c) Uji-t Sampel Berpasangan (Sampel Dependen)

Uji ini digunakan untuk membandingkan dua pengukuran yang saling berkaitan pada subjek yang sama (sebelum/sesudah) atau pasangan yang dicocokkan (*matched pairs*). Uji ini memeriksa apakah rata-rata selisih antarpasangan berbeda secara signifikan dari nol.

Contoh: Program Penurunan Berat Badan

Skenario: Seorang ahli gizi ingin menguji efektivitas program penurunan berat badan. Ia mengukur berat badan 10 peserta sebelum dan sesudah mengikuti program selama 8 minggu:

Tabel 13.2 Kumpulan data penurunan berat badan

Orang / Subjek	Sebelum	Sesudah	Selisih (Selisih = Sesudah - Sebelum)
1	180	175	-5
2	165	162	-3
3	195	188	-7
4	172	170	-2
5	158	155	-3
6	210	202	-8
7	178	176	-2
8	185	180	-5
9	167	164	-3
10	203	196	-7

Langkah 1: Hipotesis

- H_0: $\mu d = 0$ (tidak ada penurunan berat badan rata-rata)
- H_1: $\mu d < 0$ (terdapat penurunan berat badan rata-rata)

Uji sisi kiri (kita mengharapkan adanya penurunan berat badan, sehingga selisihnya bernilai negatif)

Langkah 2: Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Memeriksa Kondisi

- Selisih dan Statistik Ringkasan
- Selisih: d = [−5, −3, −7, −2, −3, −8, −2, −5, −3, −7]
- Rata-rata selisih: $\bar{d}$ = −4,5
- Simpangan baku selisih: sd = 2,224
- n = 10
- df = 9

Memeriksa kondisi:

- Uji untuk nilai rata-rata ✓
- Dua sampel dependen (berpasangan) ✓
- Ukuran sampel kecil ✓
- Simpangan baku populasi (σ) tidak diketahui ✓

Kondisi terpenuhi untuk uji-t sampel berpasangan (*paired samples t-test*)

Langkah 4: Statistik Uji

$$t = \frac{\bar{d} - \mu_{d0}}{s_d/\sqrt{n}}$$

$\mu d0$ = perbedaan rata-rata yang dihipotesiskan (biasanya 0)

$$t = \frac{-4.5 - 0}{2.224/\sqrt{10}} = -6.398$$

$$t = -6.40$$

Langkah 5: Nilai Kritis dan p-value

- Nilai kritis (df = 9, α = 0,05): =T.INV.2T(0,05, 9) = −1,833
- p-value: =T.DIST(-6,4, 9, TRUE) = 0,000062636 (sangat kecil)

Langkah 6: Keputusan

Tolak H_0.

Kesimpulan: Terdapat bukti kuat bahwa program penurunan berat badan tersebut efektif. Rata-rata, peserta mengalami penurunan berat badan ($\bar{d}$ = −4,5 pon).

Melakukan Pengujian menggunakan ToolPak di Excel

Gunakan:

Data → Data Analysis → t-Test: *Paired Two Sample for Means* (Uji-t: Dua Sampel Berpasangan untuk Rata-rata)

ToolPak Excel menyediakan tiga pilihan utama uji-t:

- *Paired Two Sample for Means* (Dua Sampel Berpasangan untuk Rata-rata) → Kelompok yang sama (sebelum–sesudah)
- *Two-Sample Assuming Equal Variances* (Dua Sampel dengan Asumsi Varians Sama) → Kelompok independen, variabilitas serupa
- *Two-Sample Assuming Unequal Variances* (Dua Sampel dengan Asumsi Varians Tidak Sama) → Kelompok independen, variabilitas berbeda

13.5 UJI CHI-SQUARE (X^2): PENERAPAN PRAKTIS

Sejauh ini, kita telah berfokus pada pengujian hipotesis untuk data numerik, seperti rata-rata dan proporsi. Namun, banyak masalah di dunia nyata melibatkan data kategorikal—data yang terbagi ke dalam kelompok atau kategori tertentu. Sebagai contoh, kita mungkin ingin menguji apakah preferensi pelanggan sesuai dengan klaim perusahaan, atau apakah dua variabel kategorikal saling berkaitan. Untuk menjawab pertanyaan semacam itu, kita menggunakan uji Chi-Square (χ^2), yang dirancang khusus untuk data kategorikal. Gagasan utama di balik pengujian ini sederhana: Bandingkan apa yang Anda amati dalam data dengan apa yang diharapkan seandainya hipotesis nol benar.

Statistik Chi-Square

Semua uji χ^2 menghitung satu nilai ringkasan yang disebut statistik Chi-Square, yang mengukur perbedaan antara frekuensi yang diamati dan frekuensi yang diharapkan:

$$\chi^2 = \sum \frac{(O-E)^2}{E} v$$

Di mana:

- O = Frekuensi yang diamati (*observed frequency*)
- E = Frekuensi yang diharapkan (*expected frequency*)
- $\sum$ = Penjumlahan untuk seluruh kategori

Interpretasi

- Nilai χ^2 kecil → Nilai yang diamati mendekati nilai yang diharapkan → mendukung H_0
- Nilai χ^2 besar → Nilai yang diamati berbeda dari nilai yang diharapkan → bukti yang menentang H_0

a) Uji Kesesuaian Chi-Square (Chi-Square Goodness-of-Fit Test)

Tujuan: Untuk menentukan apakah distribusi sampel sesuai dengan distribusi teoretis tertentu.

Contoh: Distribusi Rasa Permen

Sebuah perusahaan permen menyatakan bahwa setiap kantong permen kunyah buah berisi empat jenis rasa dengan proporsi yang sama (masing-masing 25%). Anda membeli satu kantong berisi 80 permen dan mencatat jumlah sebagai berikut:

Rasa	Diamati (O)
Ceri	26
Jeruk	23

Lemon	18
Jeruk Nipis	13
Total	80

Langkah 1: Nyatakan Hipotesis

- H_0: Distribusi sesuai dengan proporsi yang diklaim
- H_1: Distribusi tidak sesuai

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Kondisi

- Data bersifat kategorikal ✓
- Jumlah yang diharapkan ≥ 5 untuk semua kategori ✓
- Pengamatan bersifat independen ✓
- Membandingkan distribusi yang diamati dengan yang diharapkan ✓
- Kondisi terpenuhi untuk uji kecocokan χ^2

Langkah 4: Hitung Statistik Uji

Rumus:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Frekuensi Harapan:

- $E = Total\ ukuran\ sampel \times proporsi\ hipotetis$
- E = 80 × 0,25 = 20 (untuk setiap kategori)

Tabel 13.3 Perhitungan Chi-Square

Rasa	**O**	**E**	**(O – E)² / E**
Ceri	26	20	1.80
Jeruk	23	20	0.45
Lemon	18	20	0.20
Jeruk Nipis	13	20	2.45
Total χ^2			**4.90**

$$\chi^2 = 4.90$$

Derajat Kebebasan (*Degrees of Freedom*)

$$df = k - 1 = 4 - 1 = 3$$

Di mana:

k = jumlah kategori

Langkah 5: Tentukan Nilai Kritis ($\chi^2\alpha, df$) dan p-value

- Nilai Kritis: =CHISQ.INV.RT(0.05, 3) Nilai kritis ≈ 7,815
- P-value: =CHISQ.DIST.RT(4,9, 3) p-value ≈ 0,179

Langkah 6: Keputusan

Aturan keputusan:

- Jika χ^2 > nilai kritis → tolak H_0
- Jika χ^2 ≤ nilai kritis → gagal menolak H_0
- Jika p-value ≤ α → tolak H_0
- Jika p-value > α → gagal menolak H_0

Keputusan

- ❖ χ^2 = 4,90 < 7,815 → Gagal menolak H_0
- ❖ p-value > 0,05 → Gagal menolak H_0

Kesimpulan: Tidak terdapat cukup bukti untuk menyimpulkan bahwa distribusi rasa berbeda dari klaim perusahaan.

Fungsi Chi-Square pada Excel

Fungsi untuk P-value dan Nilai Kritis

Tabel 13.4 Fungsi untuk nilai kritis dan p-value Chi-Square

Tugas	Fungsi yang Tepat	Alasan
Hitung nilai-p untuk uji chi-kuadrat	CHISQ.DIST.RT(x, deg_freedom) x = statistik Chi-Square yang dihitung	Uji hipotesis menggunakan ekor kanan.
Hitung nilai kritis untuk uji chi-kuadrat	CHISQ.INV.RT(probability, deg_freedom)	Nilai kritis berasal dari ekor kanan.

x = statistik Chi-Square yang dihitung

Fungsi CHISQ.TEST:

Fungsi CHISQ.TEST(*actual_range, expected_range*) di Excel melakukan uji Chi-Square sisi kanan (*right-tailed*) dengan membandingkan frekuensi yang diamati (*observed*) dan frekuensi yang diharapkan (*expected*). Hal ini sesuai dengan Uji Kecocokan (*Goodness-of-Fit Test*) Chi-Square ketika rentang data mewakili satu variabel kategorikal.

Pada contoh sebelumnya:

=CHISQ.TEST(B2:B5, C2:C5)

- B2:B5 = jumlah yang diamati (*observed counts*)
- C2:C5 = jumlah yang diharapkan (*expected counts*)

	A	B	C	D
1	**Rasa**	**O**	**E**	**$(O-E)^2 / E$**
2	Ceri	26	20	1,8
3	Jeruk	23	20	0,45
4	Lemon	18	20	0,2
5	Limau	13	20	2,45
6	Total χ^2			**4,9**

Gambar 13.1 Data Tte iw Excet

CHISQ.TEST mengembalikan nilai p

b) Uji Chi-Square Independensi

Uji ini mengevaluasi apakah dua variabel kategorikal berasosiasi atau independen dalam satu populasi. Contoh: Jenis Kelamin vs Preferensi Film. Skenario: Seorang peneliti mensurvei 200 individu untuk memeriksa apakah jenis kelamin berhubungan dengan genre film yang disukai.

Tabel 13.5 Jumlah yang diamati

	Aksi	**Romansa**	**Komedi**	**Total**
Laki-laki	55	15	30	100
Perempuan	25	45	30	100
Total	80	60	60	200

Langkah 1: Hipotesis

- H_0: Jenis kelamin dan preferensi genre film bersifat independen (tidak saling bergantung).
- H_1: Jenis kelamin dan preferensi genre film memiliki keterkaitan (asosiasi).

Langkah 2: Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Pemeriksaan Kondisi

- Kedua variabel bersifat kategorikal ✓
- Observasi bersifat independen ✓
- Pengujian untuk melihat keterkaitan (asosiasi) ✓

Kondisi telah terpenuhi untuk melakukan uji independensi Chi-Square (χ^2).

Langkah 4: Statistik Chi-Square dan derajat bebas (df)

Hitung Frekuensi Harapan (*Expected Counts*)

$$E = \frac{(\text{Total Baris})(\text{Total Kolom})}{\text{Total Keseluruhan}}$$

Contoh:

$$E_{\text{Male, Action}} = \frac{100 \times 80}{200} = 40$$

Tabel 13.6 Jumlah yang diharapkan

	Aksi	**Romantis**	**Komedi**
Laki-laki	40	30	30
Perempuan	40	30	30

Perhitungan Chi-Square

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- Pria-Aksi: (55 - 40)² / 40 = 5,625
- Pria-Romansa: (15 - 30)² / 30 = 7,5
- Pria-Komedi: (30 - 30)² / 30 = 0
- Wanita-Aksi: (25 - 40)² / 40 = 5,625
- Wanita-Romansa: (45 - 30)² / 30 = 7,5
- Wanita-Komedi: (30 - 30)² / 30 = 0

$$\chi^2 = 5{,}625 + 7{,}5 + 0 + 5{,}625 + 7{,}5 + 0 = 26{,}25$$

Derajat Kebebasan (df): (jumlah baris - 1) × (jumlah kolom - 1)

$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$$

Langkah 5: Nilai Kritis dan p-value

- Nilai Kritis: =CHISQ.INV.RT(0,05, 2) Nilai kritis ≈ 5,991
- P-value: =CHISQ.DIST.RT(4,9, 2) p-value ≈ 0,0000019

Langkah 6: Keputusan

- χ^2 = 26,25 > 5,991 → Tolak H_0
- p-value < 0,05 → Tolak H_0

Kesimpulan: Terdapat hubungan yang signifikan secara statistik antara jenis kelamin dan preferensi film.

Menggunakan Fungsi CHISQ.TEST pada Excel

Fungsi CHISQ.TEST() juga sangat tepat untuk digunakan dalam kasus ini. Jika jumlah yang diamati berada di B2:D3 dan jumlah yang diharapkan berada di B7:D8:

$$= CHISQ.TEST(B2{:}D3, B2{:}D8)$$

	A	B	C	D	E
1		**Aksi**	**Romantis**	**Komedi**	**Total**
2	**Pria**	55	15	30	100
3	**Wanita**	25	45	30	100
4	**Total**	80	60	60	200
5					
6		**Aksi**	**Romantis**	**Komedi**	
7	**Pria**	40	30	30	
8	**Wanita**	40	30	30	

Gambar 13.2 Jumlah yang diamati dan yang diharapkan (Excet)

Ini menghasilkan nilai p sekitar 0,0000019, yang mengkonfirmasi hasilnya.

13.6 UJI F: APLIKASI PRAKTIS (CONTOH ANOVA)

Pengantar Uji F

Uji F adalah prosedur statistik yang digunakan untuk membandingkan variabilitas antar kelompok. Uji ini didasarkan pada distribusi F, distribusi miring ke kanan yang muncul ketika mengambil rasio dua varians. Dalam praktiknya, uji F membantu menjawab pertanyaan seperti:

- Apakah beberapa kelompok berbeda secara signifikan dalam kinerja rata-rata mereka?
- Apakah variasi antar kelompok lebih besar daripada variasi dalam kelompok?

Dua aplikasi umum uji F adalah:

- **Perbandingan Varians**: Menguji apakah dua populasi memiliki varians yang sama (seringkali merupakan langkah awal dalam pengujian hipotesis).
- **Analisis Varians (ANOVA)**: Menguji apakah tiga atau lebih rata-rata populasi sama.

Pada bagian ini, kita akan fokus pada ANOVA satu arah, aplikasi uji-F yang paling banyak digunakan.

Logika di Balik ANOVA

Misalkan Anda membandingkan efektivitas tiga metode pengajaran: Metode A, Metode B, dan Metode C. Pertanyaan kuncinya adalah:

> *Apakah ketiga metode tersebut sama efektifnya, atau apakah ketiganya menghasilkan hasil yang berbeda?*

Alih-alih melakukan beberapa uji-t (yang meningkatkan risiko kesalahan Tipe I), ANOVA menggunakan satu uji terpadu.

Ide Inti

ANOVA membandingkan dua jenis variasi:

- Variasi antar kelompok: Perbedaan antara rata-rata kelompok
- Variasi dalam kelompok: Variabilitas antar pengamatan dalam setiap kelompok

Statistik F adalah rasio:

$$F = \frac{\text{Mean Square Between}}{\text{Mean Square Within}}$$

Interpretasi

- Nilai F yang besar → Variasi antar-kelompok tergolong besar dibandingkan dengan variasi di dalam kelompok → bukti yang menentang H_0
- Nilai F yang kecil → Perbedaan antar-kelompok sebanding dengan variabilitas di dalam kelompok → bukti tidak cukup untuk menentang H_0

Enam Langkah Pengujian Hipotesis (Contoh ANOVA)

Skenario: Seorang peneliti membandingkan rata-rata nilai ujian mahasiswa yang diajar menggunakan tiga metode pengajaran yang berbeda. Data sampel dikumpulkan dari setiap kelompok, dan ANOVA satu arah (*one-way ANOVA*) digunakan untuk menguji apakah terdapat perbedaan pada nilai rata-rata tersebut.

Langkah 1: Menentukan Hipotesis

- H_0: $\mu_1 = \mu_2 = \mu_3$ (semua rata-rata kelompok adalah sama)
- H_1: Setidaknya satu rata-rata kelompok berbeda

Langkah 2: Menetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Memilih Uji yang Tepat | Memeriksa kondisi

- Satu variabel independen kategorikal (metode pengajaran) ✓
- Tiga kelompok ✓
- Satu variabel dependen kontinu (nilai ujian) ✓
- Observasi yang independen ✓

Kondisi terpenuhi untuk uji ANOVA satu arah

Langkah 4: Menghitung Statistik Uji

Penjelasan:

- ANOVA membagi variabilitas total menjadi dua komponen:
- Jumlah kuadrat antar-kelompok (SS_{between})
- Jumlah kuadrat dalam-kelompok (SS_{within})

Komponen-komponen ini kemudian diubah menjadi kuadrat rata-rata (*mean squares*):

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}, \qquad MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

Kemudian:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

4.1. Data mentah (n = 5 per kelompok, N = 15 total), rerata kelompok, dan rerata keseluruhan

Tabel 13.7

Ob.	Kelompok 1: Tradisional	Kelompok 2: Dibalik (Flipped)	Kelompok 3: Daring
1	79	87	71
2	81	93	72
3	76	88	68
4	80	93	70
5	84	94	74
Jumlah (Σ)	400	455	355
Mean ($\bar{x}_i$)	80,00	91,00	71,00
Rata-rata keseluruhan (grand mean): $\bar{x}$ grand = (400 + 455 + 355) / 15 = 1210 / 15 = **80,67**			

4.2. Jumlah kuadrat antar-kelompok (SS_{between})

Tabel 13.8

$$SS_{\text{between}} = \sum n_i(\bar{x}_i - \bar{x}_{grand})^2$$

Group	n_i	$\bar{x}_i$	$(\bar{x}_i - 80.67)$	$(\bar{x}_i - 80.67)^2$	$n_i \times (\bar{x}_i - 80.67)^2$
Traditional	5	80.00	−0.6667	0.4445	2.2224
Flipped	5	91.00	+10.3333	106.7771	533.8854
Online	5	71.00	−9.6667	93.4451	467.2254
				SS_{between}	1003.3333

4.3. Jumlah kuadrat dalam kelompok (SS_{within})

Tabel 13.9

$$SS_{\text{within}} = \sum(x - \bar{x}_i)^2 \text{ for each group}$$

Obs.	G1: $(x_i-80)^2$	G2: $(x_i-91)^2$	G3: $(x_i-71)^2$
1	$(79-80)^2 = 1$	$(87-91)^2 = 16$	$(71-71)^2 = 0$
2	$(81-80)^2 = 1$	$(93-91)^2 = 4$	$(72-71)^2 = 1$
3	$(76-80)^2 = 16$	$(88-91)^2 = 9$	$(68-71)^2 = 9$
4	$(80-80)^2 = 0$	$(93-91)^2 = 4$	$(70-71)^2 = 1$
5	$(84-80)^2 = 16$	$(94-91)^2 = 9$	$(74-71)^2 = 9$
SS per group	34	42	20

$$SS_{\text{within}} = 34 + 42 + 20 = \mathbf{96}$$

4.4. Derajat bebas, kuadrat rata-rata, & statistik-F

$df_{between}\ (k-1)$	$df_{within}\ (N-k)$	$df_{total}\ (N-1)$
2	12	14

$MS_{between} = SS_{between} / df_{between} = 1003.3333 / 2 = \mathbf{501.66665}$

$MS_{within} = SS_{within} / df_{within} = 96/12 = \mathbf{8}$

$F = MS_{between} / MS_{within} = 501.66665/8 = \mathbf{62.71}$

Langkah 5: Tentukan nilai p atau Nilai Kritis

- F kritis: =F.INV.RT(0.05, 2, 12) = 3.885
- Nilai p: =F.DIST.RT(62.71,2,12) =0.00000044

Tabel 13.10 Ringkasan ANOVA

Sumber	SS	df	MS	F	p-value	F kritis
Antar kelompok	1003,333	2	501.66665	62.71	0,00000044	3,885
Dalam kelompok	96,00	12	8.00	—	—	—
Total	1099,33	14	—	—	—	—

Langkah 6: Keputusan dan Interpretasi

Tabel 13.11

Kriteria	Nilai	Ambang batas	Hasil
p-value vs α	0,00000044	α = 0,05	$p < \alpha$
F vs F kritis	62,71	3,885	F > Fcrit → Tolak H_0

Kesimpulan: Terdapat bukti statistik yang cukup pada tingkat signifikansi 5% untuk menyimpulkan bahwa nilai rata-rata ujian berbeda di antara ketiga metode pengajaran (F(2, 12) = 62,71, p = 0,00000044). Metode Flipped Classroom menghasilkan nilai rata-rata tertinggi (91), diikuti oleh Tradisional (80) dan Online (71).

Melakukan Anova di Excel (Data Analysis Toolpak)

Langkah 1: Mengatur Data

Tabel 13.12 data yang telah diatur di Excel

Catatan	Kelompok 1: Tradisional	Kelompok 2: Dibalik	Kelompok 3: Daring
1	79	87	71
2	81	93	72
3	76	88	68
4	80	93	70
5	84	94	74

Langkah 2: Jalankan ANOVA

- Buka: Data → Data Analysis → ANOVA: Single Factor
- Pilih rentang input (input range)
- Centang "*Labels in First Row*" (jika ada)
- Tetapkan α = 0,05
- Tentukan lokasi output

Langkah 3: Interpretasi Output

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Group 1: Traditional	5	400	80	8.5
Group 2: Flipped	5	455	91	10.5
Group 3: Online	5	355	71	5

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1003.333	2	501.6666667	62.70833	4.435E-07	3.8852938
Within Groups	96	12	8			
Total	1099.333	14				

Gambar 13.3 Laporan ANOVA di Excel

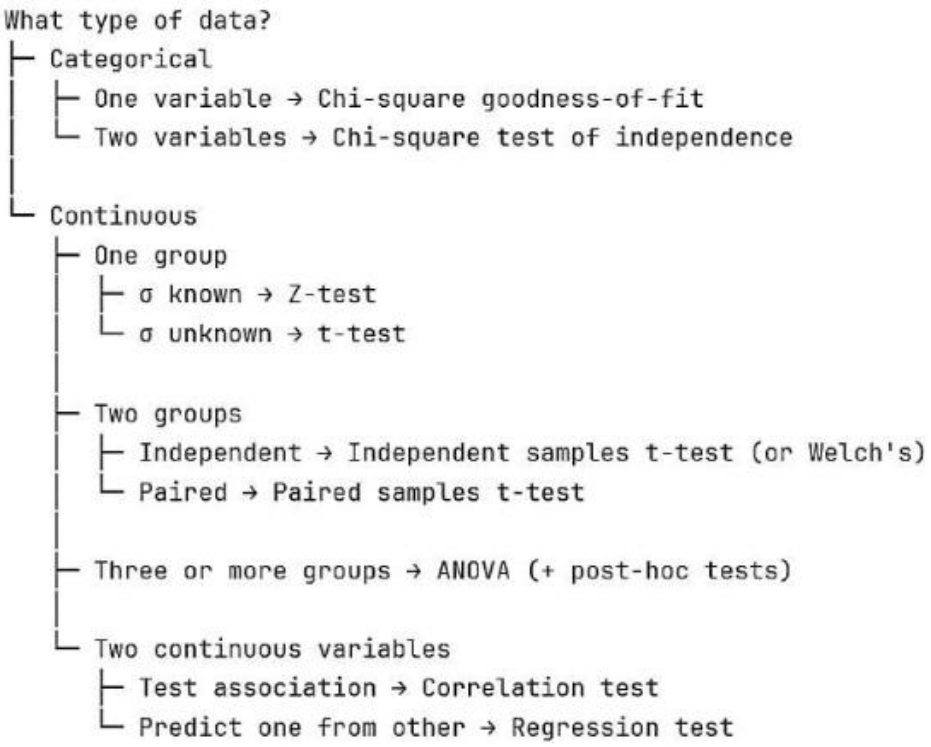
13.7 PANDUAN PENGAMBILAN KEPUTUSAN DALAM PEMILIHAN UJI STATISTIK

Memilih uji statistik yang sesuai merupakan salah satu langkah paling krusial dalam analisis data.

Tabel 13.13 Memilih Uji Statistik yang Sesuai

Situasi	**Uji yang Direkomendasikan**
Rata-rata satu sampel, σ diketahui	Uji Z satu sampel untuk rata-rata
Rata-rata satu sampel, σ tidak diketahui	Uji t satu sampel
Proporsi satu sampel	Uji Z satu sampel untuk proporsi
Rata-rata dua sampel independen (varians sama)	Uji t gabungan (*pooled t-test*)
Rata-rata dua sampel independen (varians tidak sama)	Uji t Welch
Rata-rata dua sampel berpasangan (sebelum–sesudah)	Uji t berpasangan (*paired t-test*)
Proporsi dua sampel	Uji Z dua sampel untuk proporsi
Varians dua sampel	Uji F untuk kesamaan varians
Rata-rata tiga kelompok atau lebih	ANOVA satu arah (uji F)
Satu variabel kategorikal vs ekspektasi	Uji kecocokan (*goodness-of-fit*) Chi-kuadrat
Dua variabel kategorikal	Uji independensi Chi-kuadrat

Panduan Pengambilan Keputusan:



13.8 PENERAPAN UJI STATISTIK DALAM ANALISIS DATA DAN PENGAMBILAN KEPUTUSAN

Dalam bab ini, kita beralih dari teori pengujian hipotesis ke penerapan praktisnya, dengan membekali Anda perangkat yang diperlukan untuk menganalisis data dunia nyata dan menarik kesimpulan yang bermakna. Anda telah mempelajari bagaimana berbagai uji statistik dirancang untuk menjawab jenis pertanyaan pembelajaran tertentu, bergantung pada struktur dan sifat data yang ada.

Kita memulai dengan memperkenalkan uji hipotesis yang paling sering digunakan uji-Z, uji-t, uji Chi-Square (χ^2), dan uji-F (ANOVA) serta menelaah variasi-variasi utamanya. Anda memahami bahwa setiap uji memiliki tujuan yang berbeda, baik untuk membandingkan rata-rata, proporsi, frekuensi, maupun variabilitas antar kelompok. Melalui contoh-contoh langkah demi langkah, Anda telah mengembangkan kemampuan untuk:

- Melakukan uji-Z untuk rata-rata dan proporsi ketika parameter populasi diketahui atau ukuran sampel cukup besar
- Menerapkan uji-t (satu sampel, independen, dan berpasangan) saat bekerja dengan sampel kecil dan varians populasi yang tidak diketahui
- Menggunakan uji Chi-Square untuk menganalisis data kategorikal, termasuk pengujian *goodness-of-fit* (kecocokan model) dan independensi
- Melakukan ANOVA (uji-F) untuk membandingkan rata-rata tiga kelompok atau lebih menggunakan kerangka kerja terpadu

Bab ini juga menekankan pentingnya asumsi, kondisi pengujian, serta interpretasi hasil termasuk penggunaan pendekatan nilai kritis (*critical value*) maupun nilai-*p* (*p-value*). Selain itu, Anda mempelajari cara menerapkan uji-uji tersebut secara efisien menggunakan *Data Analysis ToolPak* dan fungsi-fungsi bawaan dalam Excel, yang menjembatani teori statistik dengan alur kerja analisis data praktis. Terakhir, Anda dibekali dengan kerangka kerja pengambilan keputusan yang jelas untuk memilih uji statistik yang tepat, sehingga Anda dapat menghadapi masalah di dunia nyata dengan penuh keyakinan dan kejelasan.

BAB 14
MENGUJI HUBUNGAN — KORELASI DAN REGRESI

14.1 KORELASI DAN REGRESI: PENGUJIAN SIGNIFIKANSI HUBUNGAN ANTARVARIABEL

Pada bab sebelumnya, kita telah menyusun perangkat yang ampuh untuk membandingkan kelompok-kelompok data. Dengan menggunakan uji-z, uji-t, uji-F, dan uji Chi-Square, kita dapat menjawab pertanyaan seperti: "Apakah Kelompok A berbeda secara signifikan dari Kelompok B?" Dalam bab ini, kita melangkah lebih jauh untuk mengajukan pertanyaan yang lebih mendalam dan bermakna: "Apakah variabel-variabel ini saling berhubungan?"

Sebelumnya dalam buku ini, khususnya pada Bab 7 (Hubungan dalam Data Anda) dan Bab 8 (Pengantar Analisis Regresi), Anda telah diperkenalkan pada korelasi dan regresi sebagai alat untuk mendeskripsikan hubungan antarvariabel. Anda telah mempelajari cara mengukur asosiasi, memvisualisasikan tren, dan membangun model prediksi sederhana. Kini, kita kembali membahas konsep-konsep tersebut namun dari perspektif inferensial yang lebih mendalam. Kita tidak lagi sekadar mendeskripsikan hubungan, melainkan menguji apakah hubungan tersebut signifikan secara statistik. Dalam praktik sains data di dunia nyata, banyak pertanyaan penting yang tidak berfokus pada perbedaan, melainkan pada hubungan:

- Jika pengeluaran untuk iklan meningkat, apakah penjualan juga ikut meningkat?
- Apakah terdapat hubungan antara jumlah jam belajar seorang mahasiswa dengan hasil ujiannya?
- Bisakah kita memprediksi harga sebuah rumah berdasarkan luas bangunannya?

Untuk menjawab pertanyaan-pertanyaan semacam itu, kita mengandalkan dua perangkat utama:

- Korelasi, yang mengukur kekuatan dan arah hubungan antarvariabel.
- Regresi, yang memodelkan hubungan tersebut dan memungkinkan kita untuk melakukan prediksi.

Secara bersama-sama, teknik-teknik ini membawa kita melampaui sekadar deskripsi, menuju pemahaman, penjelasan, dan peramalan (*forecasting*).

Namun, menghitung koefisien korelasi atau menentukan garis regresi hanyalah sebagian dari keseluruhan proses. Seperti dalam setiap analisis statistik, kita harus bertanya: Apakah hubungan yang teramati tersebut signifikan secara statistik, atau mungkinkah hubungan itu terjadi secara kebetulan? Di sinilah pengujian hipotesis kembali memegang peranan sentral. Di akhir bab ini, Anda akan mampu untuk:

- Menguji apakah koefisien korelasi signifikan secara statistik
- Mengevaluasi apakah model regresi memberikan daya prediksi yang bermakna
- Menafsirkan hasil utama seperti nilai-p (*p-value*), koefisien, dan ukuran kecocokan model (*goodness-of-fit*)
- Melakukan analisis korelasi dan regresi menggunakan Excel

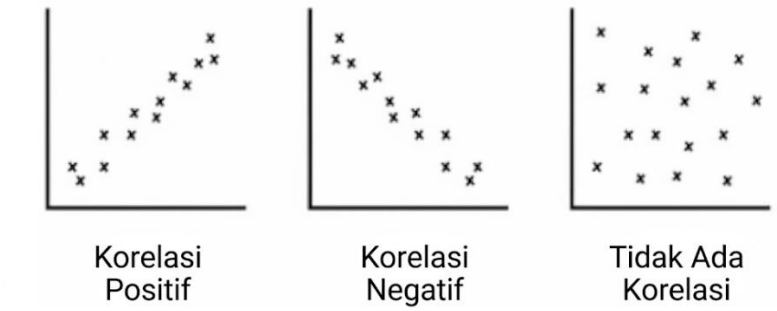
- Menerjemahkan hasil statistik menjadi wawasan praktis yang dapat diterapkan di dunia nyata

14.2 KORELASI, REGRESI, DAN PENGUJIAN HIPOTESIS

Korelasi

Korelasi adalah ukuran numerik tunggal yang menggambarkan kekuatan dan arah hubungan linear antara dua variabel kuantitatif. Ukuran ini disebut sebagai koefisien korelasi, yang dilambangkan dengan r.

- Nilai r berkisar antara −1 hingga +1
- Korelasi positif (r mendekati +1): Saat satu variabel meningkat, variabel lainnya cenderung ikut meningkat.
- Korelasi negatif (r mendekati −1): Saat satu variabel meningkat, variabel lainnya cenderung menurun.
- Tidak ada korelasi linear (r mendekati 0): Tidak terdapat hubungan garis lurus yang bermakna.



Gambar 14.1 Korelasi positif, negatif, dan tidak ada korelasi

Di Excel, korelasi dapat dihitung dengan mudah. Jika data "*Ad Spend*" (Pengeluaran Iklan) berada di rentang A2:A10 dan "*Sales*" (Penjualan) berada di rentang B2:B10, gunakan rumus:

$$= CORREL(A2{:}A10, B2{:}B10)$$

Rumus ini akan menghasilkan koefisien korelasi sampel. Namun, diperolehnya nilai r memunculkan sebuah pertanyaan penting:

> *Apakah korelasi yang teramati ini signifikan secara statistik, atau mungkinkah hal tersebut terjadi secara kebetulan?*

Untuk menjawab hal ini, kita melakukan uji hipotesis terhadap koefisien korelasi populasi, yang dilambangkan dengan (ρ):

Hipotesis untuk Korelasi

- H_0: $\rho = 0$ (tidak ada hubungan linear dalam populasi)

- H_1: $\rho \neq 0$ (terdapat hubungan linear)

Di sini, ρ (rho) merepresentasikan koefisien korelasi populasi, yang membedakannya dari statistik sampel r.

Regresi Linear Sederhana

Meskipun korelasi merangkum suatu hubungan dalam satu angka, korelasi tidak menyediakan mekanisme untuk melakukan prediksi. Untuk itu, kita menggunakan analisis regresi. Regresi linear sederhana memodelkan hubungan antara dua variabel dengan menyesuaikan garis yang paling tepat (*line of best fit*) melalui data tersebut:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

Kemiringan (*slope*), β_1, merupakan aspek sentral dalam interpretasi parameter ini menangkap arah dan besarnya hubungan tersebut.

Uji Hipotesis dalam Regresi

Dalam regresi, pengujian hipotesis berfokus pada apakah kemiringan tersebut bermakna secara statistik.

- H_0: $\beta_1 = 0$ (tidak ada hubungan; garis bersifat datar)
- H_1: $\beta_1 \neq 0$ (terdapat hubungan linear yang signifikan)

Jika kemiringan tidak berbeda secara signifikan dari nol, maka model tersebut tidak memberikan nilai prediktif apa pun.

Di Luar Regresi Linear Sederhana

Dalam regresi linear sederhana, satu prediktor menjelaskan satu hasil (*outcome*). Namun, masalah di dunia nyata sering kali jauh lebih kompleks.

Regresi Linear Berganda (*Multiple Linear Regression*/MLR): Menguji Beberapa Prediktor Sekaligus

Regresi berganda memperluas model dengan menyertakan beberapa prediktor:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta kXk + \varepsilon$$

Hal ini memungkinkan kita untuk memodelkan skenario yang lebih realistis, di mana hasil bergantung pada berbagai faktor.

Contoh: Prediksi Harga Rumah

$$Harga = \beta 0 + \beta 1(Luas\ dalam\ kaki\ persegi) + \beta 2(Jumlah\ Kamar\ Tidur) + \beta 3(Usia\ Bangunan) + \beta 4(Ukuran\ Garasi) + \varepsilon$$

Setiap koefisien merepresentasikan pengaruh parsial dari suatu prediktor, dengan asumsi variabel-variabel lain dianggap konstan.

Pengujian Hipotesis dalam Regresi Berganda

Terdapat dua tingkat pengujian hipotesis yang dilakukan:

1. **Uji Model Secara Keseluruhan (Uji F)**
   - H_0: Semua koefisien $\beta = 0$ (tidak ada prediktor yang menjelaskan Y)
   - Menguji apakah model secara keseluruhan signifikan
2. **Uji Prediktor Secara Individual (Uji t)**
   - H_0: $\beta_i = 0$ (Setiap koefisien kemiringan/slope diuji secara terpisah)
   - Menentukan apakah setiap prediktor memberikan kontribusi yang berarti terhadap model

Regresi Non-Linear

Tidak semua hubungan bersifat linear. Beberapa hubungan menunjukkan kelengkungan (kurvatur), seperti fenomena hasil yang semakin berkurang (*diminishing returns*) atau pertumbuhan eksponensial.

Contoh:

- Kuadratik:

$$Y = \beta 0 + \beta 1X + \beta 2X2 + \varepsilon$$

- Kubik:

$$Y = \beta 0 + \beta 1X + \beta 2X2 + \beta 3X3 + \varepsilon$$

Contoh: Manfaat dari waktu belajar mungkin meningkat pesat pada awalnya, kemudian mendatar; hal ini menunjukkan hubungan yang tidak linear.

Pengujian Hipotesis dalam Model Non-Linear

Di sini, kita menguji apakah suku-suku tambahan (seperti X^2, X^3) secara signifikan meningkatkan kecocokan model dibandingkan dengan model yang lebih sederhana.

Fokus Kita dalam Bab Ini

Dalam bab ini, kita berfokus pada:

- Pengujian korelasi
- Pengujian regresi linear sederhana

Hal ini memungkinkan kita untuk membangun landasan konseptual yang kuat sekaligus menjaga agar analisis tetap mudah dikelola.

Setelah konsep-konsep inti ini dikuasai, pengembangan ke arah regresi berganda maupun regresi non-linear hanyalah masalah penerapan prinsip-prinsip yang sama dalam skala yang lebih luas. Pada bagian selanjutnya, Anda akan mempelajari contoh praktis langkah demi langkah mengenai pengujian hipotesis untuk korelasi dan regresi linear sederhana menggunakan Excel. Contoh-contoh ini akan memandu Anda melalui keseluruhan proses mulai dari merumuskan hipotesis hingga menafsirkan nilai-p (p-value) sehingga membekali Anda dengan keterampilan yang diperlukan untuk menganalisis hubungan dalam data dunia nyata dengan penuh percaya diri.

14.3 PENGUJIAN UNTUK KORELASI DAN REGRESI

Saat menganalisis hubungan antara dua variabel kontinu, sekadar menghitung koefisien korelasi atau membuat garis regresi tidaklah cukup. Kita juga harus menentukan apakah hubungan yang teramati tersebut signifikan secara statistik artinya, hubungan tersebut tidak mungkin terjadi secara kebetulan.

1. Pengujian Signifikansi Koefisien Korelasi

Koefisien korelasi Pearson (r) mengukur kekuatan dan arah hubungan linear antara dua variabel. Untuk menentukan apakah hubungan ini benar-benar ada dalam populasi, kita melakukan pengujian berikut:

Hipotesis

- H_0: $\rho = 0$ (tidak ada korelasi linear dalam populasi)
- H_1: $\rho \neq 0$ (terdapat korelasi linear)

Statistik Uji

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Di mana:

- r = koefisien korelasi sampel
- w = ukuran sampel
- $df = n - 2$

Di bawah hipotesis nol, statistik ini mengikuti distribusi-t dengan derajat kebebasan ($w - 2$).

Asumsi

- Kedua variabel bersifat kontinu
- Hubungan bersifat mendekati linear
- Observasi bersifat independen
- Tidak terdapat pencilan (*outlier*) yang signifikan
- Mendekati normalitas bivariat

Contoh Pengerjaan: Waktu Belajar dan Nilai Ujian

Seorang profesor mengumpulkan data dari 20 mahasiswa dan menghitung korelasi sebesar $r = 0{,}68$. Ujilah apakah hasil ini menunjukkan hubungan positif yang signifikan pada $\alpha = 0{,}05$.

Langkah 1: Nyatakan Hipotesis

- H_0: $\rho = 0$
- H_1: $\rho > 0$ (uji satu arah sisi kanan)

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}05$

Langkah 3: Periksa Asumsi

Semua kondisi terpenuhi dengan cukup baik. ✓

Langkah 4: Hitung Statistik Uji

$$t = \frac{0.68\sqrt{20-2}}{\sqrt{1-0.68^2}} = \frac{0.68\sqrt{18}}{\sqrt{0.5376}} = \frac{0.68 \times 4.243}{0.733} = 3.93$$

$$\mathrm{df} = 18$$

Langkah 5: Menghitung p-value

$$p - value := T.DIST.RT(3.93, 18) \approx 0{,}00049$$

Langkah 6: Keputusan dan Kesimpulan

p-value < α, maka kita menolak H_0.

Kesimpulan: Terdapat bukti kuat mengenai hubungan positif yang signifikan secara statistik antara waktu belajar dan nilai ujian.

2. Pengujian Kemiringan (Slope) Regresi ($\beta_1 \neq 0$)

Dalam regresi linear sederhana, kita memodelkan hubungan antarvariabel sebagai berikut:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

Pertanyaan utamanya adalah apakah X memprediksi Y secara signifikan, yang bergantung pada apakah kemiringan (β_1) berbeda dari nol.

Hipotesis

- $H_0: \beta_1 = 0$ (tidak ada hubungan linear)
- $H_1: \beta_1 \neq 0$ (terdapat hubungan linear yang signifikan)

Statistik Uji

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

Di mana:

- b_1 = kemiringan sampel (*sample slope*)
- SEb_1 = galat baku (*standard error*) dari kemiringan
- $df = n - 2$

Galat Baku Kemiringan

$$SE(b_1) = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Dimana:

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y_i})^2}{n-2}}$$

Wawasan Penting

Pengujian $\beta_1 = 0$ dalam regresi secara matematis setara dengan pengujian $\rho = 0$ dalam korelasi. Kedua pengujian tersebut menghasilkan statistik-t dan nilai-p (p-value) yang sama.

Contoh Pengerjaan: Iklan dan Penjualan

Sebuah perusahaan ingin menentukan apakah pengeluaran untuk iklan (dalam ribuan dolar) dapat memprediksi penjualan bulanan (dalam ribuan rupiah). Data tersebut dapat ditemukan dalam buku kerja bab ini.

Tabel 14.1 Data dari 15 bulan

Bulan	Periklanan (X)	Penjualan (Y)
1	2,5	48
2	3,0	55
3	1,8	42
4	2.2	46
5	3.5	60
6	2.8	52
7	1.5	38
8	3.2	58
9	2.0	44
10	2.7	50
11	3.8	64
12	1.2	35
13	2.9	54
14	3.3	59
15	2.4	47

Langkah 1: Nyatakan Hipotesis

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Langkah 2: Tetapkan Tingkat Signifikansi

$\alpha = 0{,}01$

Langkah 3: Periksa Asumsi

- Linearitas ✓
- Independensi ✓
- Normalitas residual ✓
- Homoskedastisitas ✓

Langkah 4: Hitung Statistik Uji

Lihat buku kerja Excel untuk bab ini guna mendapatkan penjelasan langkah demi langkah yang terperinci.

Rumusnya:

$$t = \frac{b_1 - 0}{SE_{b_1}}$$

Hitung b_1

$$= SLOPE(C2{:}C16, B2{:}B16) = 11.2$$

Hitung $SE(b_1)$

$$SE(b_1) = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sqrt{\frac{7.6846}{13}}}{\sqrt{7.77733}} = \frac{0.7688}{2.7888}$$

$$= 0.2757$$

Hitung t

$$t = \frac{11.2}{0.2757} = 40.6$$

Hitung df

$$df = 13$$

Langkah 5: Hitung p-value dan Nilai Kritis (*Critical Value*)

- p-value: =T.DIST.2T(40.6, 13) ≈ 4.42×10^{-15}
- Nilai kritis: =T.INV.2T(0.01, 13) = ±3.012

Langkah 6: Keputusan dan Kesimpulan

- p-value < α Tolak H_0
- $|t|$ > nilai kritis Tolak H_0

Kesimpulan: Pengeluaran iklan merupakan prediktor yang sangat signifikan terhadap penjualan.

Interpretasi Kemiringan (*Slope*): Untuk setiap tambahan Rp 10.000.000 yang dikeluarkan untuk iklan, penjualan meningkat sekitar Rp 112.000.000.

Implementasi Excel: Analisis Regresi

Data Analysis ToolPak pada Excel menyediakan hasil regresi lengkap dalam satu langkah.

Langkah-langkah

1. Buka Data → Data Analysis → Regression
2. Masukkan Rentang Y (Penjualan)
3. Masukkan Rentang X (Iklan)
4. Centang "*Labels*" jika ada label pada data

5. Tetapkan tingkat kepercayaan (contoh: 99%)
6. Tentukan lokasi output
7. Klik OK

Hasil Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9961
R Square	0.9922
Adjusted R Square	0.9916
Standard Error	0.7688
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	976.0486942	976.0486942	1651.168366	4.37471E-15
Residual	13	7.684639122	0.591126086		
Total	14	983.7333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	21.156	0.740	28.580	4.0448E-13	19.557	22.755	19.557	22.755
Advertising (X)	11.203	0.276	40.635	4.37471E-15	10.607	11.798	10.607	11.798

Gambar 14.2 Laporan analisis regresi

Memeriksa Asumsi Regresi

Sebelum menafsirkan hasil, selalu validasi asumsi-asumsi berikut:

1. Linearitas → Diagram pencar (*scatterplot*) menunjukkan pola linear
2. Independensi → Observasi tidak saling berkaitan
3. Normalitas → Residual berdistribusi mendekati normal
4. Homoskedastisitas → Varians residual bersifat konstan
5. Tidak ada pencilan (*outlier*) yang berpengaruh → Tidak ada titik data dengan pengaruh ekstrem (*leverage points*)

Masalah Diagnostik yang Umum dan Solusinya

Tabel 14.2 Masalah terkait asumsi

Masalah	**Gejala**	**Solusi**
Non-linearitas	Pola residu melengkung	Transformasi variabel, regresi polinomial
Heteroskedastisitas	Residual berbentuk kipas	Transformasi Y, regresi berbobot
Residual tidak berdistribusi normal	Distribusi menceng	Transformasi Y, metode robust
Autokorelasi	Pola seiring berjalannya waktu	Gunakan model deret waktu
Pencilan (*outlier*)	Pengaruh tinggi	Selidiki atau hapus
Multikolinearitas	Korelasi prediktor yang tinggi	Hapus/kurangi prediktor

14.4 PENGUJIAN SIGNIFIKANSI KORELASI DAN REGRESI LINEAR SERTA VALIDASI MODEL

Dalam bab ini, kita beralih dari sekadar mendeskripsikan hubungan antarvariabel menuju pengujian formal untuk menentukan apakah hubungan tersebut signifikan secara statistik. Berdasarkan pembahasan sebelumnya mengenai korelasi dan regresi, kita memperkenalkan pengujian hipotesis sebagai sarana untuk membedakan pola nyata dari variasi acak. Kita mengawalinya dengan meninjau kembali korelasi, yang mengukur kekuatan dan arah hubungan linear antara dua variabel menggunakan koefisien r. Konsep ini kemudian diperluas dengan menguji apakah koefisien korelasi populasi (ρ) berbeda secara signifikan dari nol. Melalui uji-t untuk korelasi, kita mempelajari cara mengevaluasi apakah hubungan yang teramati kemungkinan besar juga ada dalam populasi yang lebih luas.

Selanjutnya, kita membahas regresi linear sederhana, yang memodelkan hubungan serta memungkinkan dilakukannya prediksi. Fokus utamanya adalah pengujian kemiringan (*slope*) regresi (β_1) untuk menentukan apakah variabel independen dapat memprediksi variabel dependen secara bermakna. Kita menunjukkan bahwa pengujian kemiringan dalam regresi secara matematis setara dengan pengujian korelasi, yang menegaskan keterkaitan erat antara kedua pendekatan ini. Melalui contoh praktis, kita menerapkan kerangka kerja pengujian hipotesis secara menyeluruh mulai dari merumuskan hipotesis, menghitung statistik uji, menentukan nilai-p (*p-value*), hingga menarik kesimpulan berdasarkan data. Kita juga mengintegrasikan implementasi menggunakan Excel, yang menunjukkan bagaimana perangkat seperti *Data Analysis ToolPak* dapat menyederhanakan analisis korelasi maupun regresi, sekaligus menyajikan keluaran statistik yang komprehensif.

Terakhir, kita menekankan pentingnya memvalidasi asumsi model, yang mencakup linearitas, independensi, normalitas residual, dan varians konstan. Kita juga membahas berbagai masalah diagnostik yang umum terjadi beserta cara penanganannya, guna memastikan bahwa kesimpulan yang diambil dari model statistik bersifat akurat dan dapat diandalkan. Bab ini menandai transisi penting dari sekadar mengidentifikasi hubungan dalam data menuju upaya memvalidasi dan meyakini hubungan tersebut. Ini merupakan langkah krusial menuju analisis prediktif dan pengambilan keputusan berbasis data di dunia nyata.

BAB 15
MELANJUTKAN PERJALANAN ANDA DALAM SAINS DATA

15.1 STATISTIKA DAN ANALISIS DATA: LANDASAN MENUJU SAINS DATA DAN AI

Selamat karena telah mencapai tonggak pencapaian ini. Anda telah menyelesaikan perjalanan komprehensif melalui dasar-dasar statistika dan analisis data perjalanan yang telah membekali Anda dengan perangkat penting untuk memahami, mempertanyakan, dan mengolah data secara bermakna. Sepanjang buku ini, Anda telah membangun landasan analitis yang kuat. Anda telah mempelajari bagaimana data dikumpulkan dan disusun, bagaimana data diringkas dan divisualisasikan, bagaimana ketidakpastian dimodelkan melalui probabilitas, serta bagaimana kesimpulan ditarik melalui inferensi statistika dan pengujian hipotesis. Lebih penting lagi, Anda telah belajar untuk berpikir kritis mengenai data mempertanyakan asumsi, menafsirkan hasil dengan cermat, dan mengomunikasikan wawasan secara jelas dengan menggunakan Excel sebagai perangkat utama Anda.

Namun, ini bukanlah akhir dari perjalanan Anda. Ini justru merupakan awal dari perjalanan yang jauh lebih besar. Sains data bukanlah disiplin ilmu yang statis. Bidang ini berkembang pesat, dibentuk oleh teknologi baru, sumber data yang kian meluas, serta masalah dunia nyata yang semakin kompleks. Perangkat terus berganti, teknik terus berkembang, dan industri terus bertransformasi tetapi prinsip-prinsip inti yang telah Anda pelajari tetaplah konstan. Prinsip-prinsip tersebut membentuk landasan intelektual yang menjadi dasar bagi semua analisis tingkat lanjut, model pembelajaran mesin (*machine learning*), dan sistem kecerdasan buatan (AI). Saat Anda melangkah lebih jauh setelah menyelesaikan buku ini, Anda tidak memulai dari nol melainkan melanjutkan dari posisi yang kuat. Bab ini hadir sebagai panduan bagi Anda untuk menavigasi langkah selanjutnya, membantu Anda beralih dari pemahaman dasar menuju penerapan praktis di dunia nyata.

15.2 KOMPETENSI DASAR STATISTIKA DAN POLA PIKIR ANALITIS DALAM SAINS DATA

Sebelum melihat ke depan, luangkanlah sejenak untuk mengapresiasi apa yang telah Anda capai. Kini, Anda memiliki seperangkat keterampilan dan pola pikir yang membedakan Anda sebagai seorang profesional data yang sedang berkembang. Anda memahami:

- **Berpikir secara statistik dalam sains data**—cara mendekati masalah dengan logika, sikap skeptis, dan bukti
- **Jenis data dan skala pengukuran**—bagaimana berbagai bentuk data memengaruhi analisis dan interpretasi
- **Statistik deskriptif**—cara meringkas dan mengomunikasikan karakteristik utama data
- **Seni visualisasi data**—cara memvisualisasikan data serta menemukan pola, tren, dan anomali
- **Hubungan dalam data**—termasuk analisis korelasi dan regresi
- **Dasar-dasar probabilitas**—bahasa ketidakpastian yang menjadi landasan sains data

- **Pengambilan sampel dan distribusi sampel**—bagaimana sampel berkaitan dengan populasi yang lebih luas
- **Inferensi statistik**—bagaimana interval kepercayaan dan pengujian hipotesis mendukung pengambilan keputusan
- **Pengujian statistik terapan**—cara memilih dan menggunakan metode yang tepat untuk menjawab pertanyaan di dunia nyata

Ini bukan sekadar konsep akademis. Semuanya merupakan perangkat praktis yang digunakan sehari-hari di berbagai industri mulai dari bisnis dan keuangan hingga layanan kesehatan, kebijakan publik, dan teknologi. Kini, Anda mampu mendekati masalah data dengan cara yang terstruktur, penuh percaya diri, dan memiliki disiplin analitis.

Lebih penting lagi, Anda telah mengembangkan sesuatu yang benar-benar krusial: cara berpikir secara statistik. Pola pikir ini memungkinkan Anda untuk melampaui pengamatan di permukaan dan mendalami data secara lebih jauh. Anda dapat mengenali pola, mengukur ketidakpastian, mempertanyakan kesimpulan, serta mengambil keputusan berdasarkan bukti, bukan sekadar intuisi. Di dunia yang semakin digerakkan oleh data, kemampuan ini tergolong langka dan sangat berharga. Sebagaimana ditekankan di sepanjang buku ini, tujuannya bukan sekadar mengajarkan teknik, melainkan menumbuhkan cara berpikir yang akan terus mendukung perkembangan Anda, bahkan lama setelah Anda menyelesaikan halaman terakhir buku ini.

15.3 PERANGKAT PEMROGRAMAN PYTHON, R, DAN SQL UNTUK ANALISIS DATA

Meskipun Excel merupakan platform yang sangat baik untuk mempelajari statistik dan tetap sangat berguna untuk berbagai analisis di dunia nyata, para praktisi sains data profesional biasanya mengandalkan bahasa pemrograman untuk memperluas kemampuan mereka. Peralihan ini bukan berarti menggantikan Excel, melainkan memperluas apa yang dapat Anda lakukan dengan data. Perangkat pemrograman menawarkan beberapa keunggulan utama:

- **Reproduksibilitas**: Kode menyediakan catatan analisis yang jelas dan bertahap, sehingga pekerjaan Anda menjadi transparan serta mudah untuk diaudit atau diulang
- **Skalabilitas**: Menangani kumpulan data dengan jutaan (atau bahkan miliaran) baris secara efisien
- **Otomatisasi**: Menjalankan analisis yang berulang dengan mudah saat data baru tersedia
- **Metode Lanjutan**: Mengakses teknik statistik modern dan algoritma pembelajaran mesin (*machine learning*)
- **Visualisasi**: Membuat visualisasi berkualitas publikasi yang dapat disesuaikan secara mendalam
- **Kolaborasi**: Bekerja secara lancar dalam tim menggunakan sistem kontrol versi dan alur kerja Bersama

Seiring dengan perkembangan perjalanan Anda di bidang sains data, tiga perangkat akan menjadi sangat penting: Python, R, dan SQL.

Python: Bahasa Serba Guna yang Tangguh

Python telah menjadi bahasa dominan dalam bidang sains data berkat fleksibilitas, kemudahan akses, serta ekosistemnya yang kuat.

Keunggulan utamanya meliputi:

- Pustaka seperti pandas memungkinkan manipulasi data layaknya Excel, namun dalam skala yang jauh lebih besar.
- NumPy dan SciPy menyediakan fungsi numerik dan statistik yang andal.
- scikit-learn mendukung berbagai teknik pembelajaran mesin (*machine learning*).
- Matplotlib dan Seaborn memungkinkan visualisasi data tingkat lanjut.
- Integrasi yang mulus dengan basis data, API, aplikasi web, dan alur kerja otomatisasi.

Mungkin keunggulan terbesar Python adalah cakupannya yang tidak terbatas pada analisis data saja. Sebagai bahasa pemrograman lengkap, Python digunakan dalam pengembangan perangkat lunak, otomatisasi, kecerdasan buatan, hingga aplikasi web. Hal ini menjadikannya investasi jangka panjang yang sangat berharga bagi setiap profesional di bidang data.

R: Bahasa Pemrograman Statistik

R dirancang khusus untuk komputasi statistik dan tetap menjadi pilihan utama di kalangan ahli statistik maupun peneliti.

Keunggulannya meliputi:

- Koleksi paket statistik khusus yang sangat luas.
- Kemampuan visualisasi data yang luar biasa, terutama melalui ggplot2.
- Dukungan komunitas yang kuat, khususnya di lingkungan akademis dan riset.
- Lingkungan kerja yang ramah pengguna, seperti RStudio, yang memudahkan pemula untuk memulai.
- Dukungan bawaan untuk pembelajaran yang dapat direproduksi (*reproducible research*) melalui perangkat seperti R Markdown.

Hampir semua konsep statistik yang telah Anda pelajari dalam buku ini memiliki implementasi langsung di dalam R. Bagi banyak pembelajar, beralih dari rumus Excel ke fungsi-fungsi di R terasa alami dan intuitif.

SQL: Fondasi untuk Bekerja dengan Basis Data

Seiring dengan semakin besarnya ukuran dan kompleksitas kumpulan data, bekerja langsung dengan basis data menjadi suatu keharusan. SQL (*Structured Query Language*) adalah perangkat standar untuk berinteraksi dengan data terstruktur yang tersimpan dalam basis data.

Dengan SQL, Anda dapat:

- Melakukan kueri pada kumpulan data besar secara efisien untuk mengekstrak informasi yang relevan.
- Melakukan agregasi dan penggabungan (*join*) antar tabel.
- Menyaring, membersihkan, dan mengubah data sebelum dianalisis.
- Mengolah data yang ukurannya terlalu besar untuk dimuat ke dalam memori komputer.

SQL tidak menggantikan perangkat statistik, melainkan melengkapinya. Dalam praktiknya, Anda akan sering menggunakan SQL untuk menyiapkan dan mengekstrak data, kemudian melakukan analisis yang lebih mendalam menggunakan Python atau R.

15.4 PEMBELAJARAN PYTHON UNTUK PENGEMBANGAN KOMPETENSI SAINS DATA

Python telah menjadi bahasa pemrograman yang paling banyak diadopsi dalam ilmu data karena kesederhanaan, fleksibilitas, dan ekosistemnya yang luas. Meskipun bahasa seperti R dan Julia tetap berharga, Python menawarkan titik masuk yang paling mudah diakses dan serbaguna bagi sebagian besar pelajar. Jika Anda memilih Python sebagai alat utama Anda, jalur pembelajaran terstruktur sangatlah penting. Seri buku Ilmu Data dengan Python oleh Tina Moré dirancang untuk memandu Anda dari pemula hingga praktisi yang percaya diri melalui pendekatan praktis dan progresif.

Di seluruh seri, Anda akan:

- Pelajari dasar-dasar penulisan kode yang bersih, efisien, dan mudah dibaca
- Kembangkan kemampuan untuk membersihkan, mengubah, dan menyiapkan kumpulan data dunia nyata yang berantakan untuk dianalisis
- Menerapkan metode statistik dan teknik visualisasi untuk mengeksplorasi data, mengungkap pola, dan menghasilkan wawasan
- Membangun dan mengevaluasi model pembelajaran mesin untuk aplikasi dunia nyata

Bersama-sama, sumber daya ini menyediakan jalur lengkap yang berfokus pada praktik—membantu Anda berevolusi dari seseorang yang menulis kode menjadi seseorang yang menggunakan kode untuk menghasilkan wawasan bermakna dan dampak dunia nyata

15.5 PERAN BI DALAM VISUALISASI DATA DAN PENGAMBILAN KEPUTUSAN BISNIS

Alat *Business Intelligence* (BI) dirancang untuk mengubah data mentah menjadi wawasan yang dapat ditindaklanjuti melalui dasbor, laporan, dan visualisasi interaktif. Mereka memainkan peran penting dalam membantu organisasi memantau kinerja, mengidentifikasi tren, dan membuat keputusan berdasarkan data.

Alat BI Umum

Beberapa alat BI yang banyak digunakan di berbagai industri:

- Microsoft Power BI: Terintegrasi dengan mulus dengan Excel, SQL, dan platform cloud; banyak digunakan oleh analis dan pengguna bisnis
- Tableau: Dikenal karena visualisasi yang canggih, fleksibel, dan sangat interaktif
- Looker (Google Looker Studio): Ideal untuk pelaporan berbasis web dan integrasi data
- cloud Qlik Sense: Memungkinkan eksplorasi data yang cepat dan fleksibel melalui analitik asosiatif
- SAP BusinessObjects: Platform BI kelas perusahaan untuk organisasi besar
- IBM Cognos Analytics: Menggabungkan pelaporan, dasbor, dan wawasan yang dibantu AI

Di mana Alat BI Berperan dalam Ilmu Data

Dalam alur kerja ilmu data, alat BI sangat berharga pada tahap-tahap selanjutnya di mana wawasan harus dieksplorasi, dikomunikasikan, dan ditindaklanjuti.

Tabel 15.1 Fungsi Alat BI dalam Alur Kerja Ilmu Data

Tahapan Sains Data (*Data Science Stage*)	**Peran Alat BI (*Role of BI Tools*)**
Pengumpulan Data (*Data Collection*)	Terhubung ke sumber data (basis data, API, lembar kerja/spreadsheet)
Persiapan Data (*Data Preparation*)	Melakukan pembersihan (*cleaning*) dan transformasi data dasar
Analisis Eksploratif (*Exploratory Analysis*)	Memvisualisasikan pola, tren, dan hubungan antar data
Komunikasi Wawasan (*Insight Communication*)	Membuat dasbor dan laporan untuk para pemangku kepentingan (*stakeholders*)

Alat BI menjembatani kesenjangan antara analisis teknis dan pengambilan keputusan bisnis. Sementara alat pemrograman memungkinkan analisis mendalam, alat BI memastikan bahwa wawasan tersebut mudah diakses, dipahami, dan dapat ditindaklanjuti oleh audiens non-teknis.

15.6 STRATEGI UNTUK PEMBELAJARAN SEPANJANG HAYAT

Sains data adalah bidang yang berkembang dengan sangat cepat. Metode-metode baru terus bermunculan, teknik yang ada terus disempurnakan, dan praktik-praktik terbaik pun terus berubah. Agar tetap efektif dan relevan, Anda harus berkomitmen untuk terus belajar. Strategi berikut akan membantu Anda untuk tetap mengikuti perkembangan terkini dan terus berkembang:

1. **Membaca secara Luas dan Kritis**

   Ikuti perkembangan di seluruh bidang ini:

   - Jurnal akademik: Jelajahi pembelajaran inovatif dan metodologi yang baru muncul.
   - Blog dan buletin (*newsletter*): Belajarlah dari para praktisi yang menerapkan sains data dalam konteks dunia nyata.
   - Buku: Dapatkan pemahaman yang mendalam dan terstruktur mengenai topik-topik tertentu.
   - Makalah pembelajaran: Kembangkan kemampuan untuk menafsirkan dan mengevaluasi metode statistik.

   Selalu membacalah secara kritis. Tidak semua analisis yang dipublikasikan itu benar. Belajar mengenali kekurangan, bias, dan keterbatasan dalam karya orang lain akan mempertajam kemampuan berpikir analitis Anda sendiri.

2. **Terlibat dalam Komunitas**

   Sains data berkembang melalui kolaborasi dan pertukaran pengetahuan.

- **Forum daring**: Berpartisipasilah dalam platform seperti Stack Overflow, Cross Validated, dan komunitas sains data lainnya yang relevan.
- **Konferensi dan webinar**: Hadiri berbagai acara atau tonton sesi rekaman untuk mendapatkan informasi terkini dan inspirasi.
- **Jejaring profesional**: Ikuti para praktisi, peneliti, dan pendidik untuk belajar dari wawasan serta pengalaman mereka.

Berinteraksi dengan orang lain akan membuka wawasan Anda terhadap berbagai perspektif, tantangan praktis, dan tren yang sedang berkembang.

3. **Cari Umpan Balik dan Bimbingan (*Mentorship*)**

Proses belajar akan lebih cepat jika dibimbing oleh orang lain.

- **Mentor**: Carilah profesional berpengalaman yang dapat memberikan arahan dan sudut pandang.
- **Belajar bersama rekan (*peer learning*)**: Bergabunglah atau bentuklah kelompok belajar serta kemitraan untuk saling menjaga komitmen (*accountability partnerships*).
- **Pembelajaran terstruktur**: Ikuti kursus yang menawarkan alur pembelajaran terpandu dan umpan balik.
- **Pengembangan profesional**: Investasikan waktu dan sumber daya untuk mengikuti lokakarya, mendapatkan sertifikasi, atau menempuh pendidikan lanjutan jika diperlukan.

Umpan balik yang konstruktif membantu Anda mengenali titik buta (*blind spots*), menyempurnakan pola pikir, dan berkembang dengan lebih efektif.

15.7 PRINSIP DAN ARAH PENGEMBANGAN DIRI DALAM PERJALANAN SAINS DATA

Anda telah sampai di akhir buku ini namun, bisa jadi Anda baru saja memulai perjalanan di bidang sains data. Di sinilah peluang yang sesungguhnya bermula. Keterampilan yang telah Anda kembangkan membuka pintu menuju pembelajaran berkelanjutan, penemuan yang bermakna, dan dampak nyata di dunia. Saat Anda melangkah maju, peganglah prinsip-prinsip berikut:

- **Tetap rendah hati**: Selalu ada hal baru untuk dipelajari. Bahkan para profesional berpengalaman pun bisa melakukan kesalahan. Hadapilah setiap analisis dengan rasa ingin tahu dan keterbukaan.
- **Berpikirlah secara kritis**: Pertanyakan asumsi baik asumsi Anda sendiri maupun orang lain. Validasi hasil analisis dan pertimbangkan penjelasan alternatif.
- **Berkomunikasilah dengan jelas**: Wawasan hanya akan bernilai jika dapat dipahami. Belajarlah untuk menerjemahkan analisis ke dalam pesan yang jelas dan dapat ditindaklanjuti.
- **Bertindaklah secara etis**: Data membawa tanggung jawab. Perhatikan aspek privasi, keadilan, transparansi, serta konsekuensi yang lebih luas dari pekerjaan Anda.
- **Tetaplah memiliki rasa ingin tahu**: Ilmuwan data yang paling efektif didorong oleh keinginan tulus untuk bereksplorasi, bertanya, dan memahami.

Kita hidup di dunia yang digerakkan oleh data, di mana berbagai organisasi semakin bergantung pada bukti untuk mengambil keputusan. Kemampuan untuk menggali makna dari data, mengukur ketidakpastian, dan mengomunikasikan wawasan dengan jelas kini menjadi jauh lebih berharga daripada sebelumnya.

Prinsip-prinsip statistik yang telah Anda pelajari seperti pengujian hipotesis, selang kepercayaan (*confidence intervals*), regresi, dan sebagainya bukanlah keterampilan yang bersifat sementara. Semua itu adalah perangkat abadi yang akan tetap relevan, apa pun teknologi yang Anda gunakan di masa depan. Dalam perjalanan ini, Anda akan menghadapi tantangan, ketidakpastian, dan momen-momen keraguan. Itu semua adalah bagian dari proses. Anda juga akan merasakan kepuasan saat berhasil memecahkan masalah yang kompleks, menemukan pola tersembunyi, serta memberikan kontribusi berarti melalui data.

Jalan ke depan sepenuhnya bergantung pada bagaimana Anda menentukannya. Baik Anda memilih peran di bidang analitik, pembelajaran mesin (*machine learning*), pembelajaran, maupun bidang lainnya, fondasi yang telah Anda bangun akan terus mendukung pertumbuhan Anda. Teruslah belajar. Teruslah berlatih. Teruslah bertanya. Dan yang terpenting, gunakan data untuk menciptakan dampak yang bermakna. Jika perjalanan Anda di dunia sains data baru saja dimulai maka kemungkinannya tidak terbatas.

DAFTAR PUSTAKA

Adams, E., Korn, M., & Laskowski, C. (2021). DIY Analytics: Beyond Excel: Tips and Tools for Mastering the Basics of Statistics and Analytics to Create Your Own Data Projects. *AALL Spectrum*, *26*, 12.

Al Maruf, A., Paul, R., Imam, M. H., & Babar, Z. (2022). A systematic review of the role of SQL and Excel in data-driven business decision-making for aspiring analysts. *American Journal of Scholarly Research and Innovation*, *1*(01), 249-269.

Ali, O. M., Breik, M., Aly, T., Raslan, A. T. N. E. D., & Gheith, M. (2024). Enhancing Data Analysis and Automation: Integrating Python with Microsoft Excel for Non-Programmers. *Journal of Software Engineering and Applications*, *17*(6), 530-540.

Almutairi, B. M. (2025). The effectiveness of statistical learning tasks based on Excel software in developing statistical thinking skills related to the labor market among students of the Applied College. *Educational Process: International Journal (EDUPIJ)*, *14*(1), 1-17.

Aqlan, F., Nwokeji, J. C., & Shamsan, A. (2020, October). Teaching an introductory data analytics course using microsoft access® and excel®. In *2020 IEEE Frontiers in Education Conference (FIE)* (pp. 1-10). IEEE.

Bryant, L. A. (2021). Statistical skills for the workplace: A practical approach to teaching methods with Excel. In *The Palgrave Handbook of Political Research Pedagogy* (pp. 289-300). Cham: Springer International Publishing.

Church, K. S., Riley, J., & Schmidt, P. J. (2022). Has excel become a “golden hammer”: The paradox of data analytics in SME clusters. *Journal of Emerging Technologies in Accounting*, *19*(2), 211-234.

Cobcobo, A. J., & Capua, R. D. G. (2022). Data analysis integration computer imported learning: a Microsoft Excel case with statistics. *EDUCATUM Journal of Science, Mathematics and Technology*, *9*(1), 20-37.

Damjanovic, S., & Katanic, P. (2023). Employing microsoft excel for enhanced mathematical and statistical online pedagogy in economics amidst a pandemic. *Education Science and Management*, *1*(2), 58-72.

Diamant, A. (2024). Introducing prescriptive and predictive analytics to MBA students with Microsoft Excel. *INFORMS Transactions on Education*, *24*(2), 152-174.

Fortino, A. (2024). *Data Analysis for Business Decisions: Case Studies in Excel, R, and ChatGPT*. Walter de Gruyter GmbH & Co KG.

Gasigwa, T., Maniraho, J. F., & Habimana, O. (2022). Effectiveness of excel software integration in teaching and learning statistics on students' performance in selected Kicukiro district upper secondary schools in Rwanda. *Journal of Research Innovation and Implications in Education*, *6*(3), 194-200.

Gonçalves, R. A., & Hornburg, A. M. (2025). Mathematical error in the Microsoft Excel program "MODE" function, its implications and possible corrections. *Revista Aracê, São José dos Pinhais*, *7*(1), 2248-2257.

Gündüz, M., & Asan, K. (2022). GEOstats: an excel-based data analysis program applying basic principles of statistics for geological studies. *Earth Science Informatics*, *15*(1), 705-712.

Hossain, E. (2021). MS Excel in Engineering Data. In *Excel Crash Course for Engineers* (pp. 169-242). Cham: Springer International Publishing.

Hyer, A., Sand, J., Cordova, A., Davis, B., & Neet, A. (2026). Using Microsoft Excel to Teach Foundational Data Analysis Skills in Online Learning: Evolution of a Time-Honored Approach to Presenting Career Relevant Skills Into an Uncertain Future. *New Directions for Adult and Continuing Education*.

Iskakova, M., Toleugaliyeva, S., Karzhaubay, A., & Diyarova, L. (2024). USE OF EXCEL IN TEACHING ELEMENTS OF STATISTICS IN SCHOOL MATHEMATICS COURSE. *Bulletin of Abai KazNPU. Series of Physical and Mathematical sciences*, *88*(4), 99-105.

Kazika, G. M., & Malambo, P. (2025). Effectiveness of Excel-assisted instruction in teaching statistics for learners' acquisition of 21st century skills. *Journal of Research Innovation and Implications in Education*, *9*(1), 254-264.

Kelvin, K., Wahab, W., & Wulandari, M. (2024). Computer resource utilization analysis for microsoft excel and python in data processing. *Engineering, Mathematics and Computer Science Journal (EMACS)*, *6*(2), 137-142.

Kumar, M. D. (2023). A study on importance of microsoft excel data analysis statistical tools in research works. *Journal of Management & Educational Research Innovation*, *1*(3), 25-33.

Lee, J., & Lee, C. F. (2022). *Essentials of Excel VBA, Python, and R: Volume I: Financial Statistics and Portfolio Analysis*. Springer International Publishing.

Mahato, T. K. (2023). Biostatistics: Simple practical way to use Microsoft excel to find mean, median, mode, standard deviation and correlation. *World Journal of Biology Pharmacy and Health Sciences*, *13*(3), 150-155.

Mairing, J. P. (2020). The Effect of Advance Statistics Learning Integrated Minitab and Excel with Teaching Teams. *International Journal of Instruction*, *13*(2), 139-150.

Mustafy, T., & Rahman, M. T. U. (2024). Excel. In *Statistics and Data Analysis for Engineers and Scientists* (pp. 81-134). Singapore: Springer Nature Singapore.

Nath, A. (2021). Analysing Learners Perception about Data Analysis on Microsoft Excel. *Learning Community: An International Journal on Educational and Social Development,*.

Orozova, D., & Hristova, N. (2020). APPLICATION OF EXCEL FOR DATA ANALYTICS IN EDUCATION. In *ICERI2020 Proceedings* (pp. 6346-6350). IATED.

Ozgur, C., Jha, S., & Wallner, M. (2022). R, Python, Excel, SPSS, SAS, and MINITAB in Banking Research. *AIMS International Journal of Management*, *16*(1), 51.

Park, T., Hong, M., & Kondrat, D. (2022). Undergraduate students' attitudes toward statistics: How Excel makes a student's life easier. *Journal of Evidence-Based Social Work*, *19*(3), 263-271.

Sankaran, S., Sankaran, K., & Bui, T. (2023). Student satisfaction with R vs. Excel in Data Mining and Business Analytics: A Herzberg's motivation-hygiene theory perspective. *Decision Sciences Journal of Innovative Education*, *21*(2), 68-82.

Simaremare, S., & Siregar, R. (2024). Effectivity of the Microsoft Excel application on student understanding in statistics courses. *Indonesian Journal of Education and Social Humanities*, *1*(2), 9-19.

Sun, M., Han, R., Jiang, B., Qi, H., Sun, D., Yuan, Y., & Huang, J. (2026). A survey on large language model-based agents for statistics and data science. *The American Statistician*, *80*(3), 331-344.

Weltman, D. (2021). Illustrating the Chi-Square Test Statistic Distribution via Interactive Excel Simulation Application in Introductory Business Statistics. *Business Education Innovation Journal*, *13*(2), 101.

York, R. O. (2022). Using Special Excel Files to Analyze Your Outcome Data. In *Evaluating Human Service Outcomes* (pp. 161-178). Cham: Springer International Publishing.

Cara Berpikir
STATISTIK
dalam mengolah Data dengan Excel

Dr. Ir. Agus Wibowo, M.Kom, M.Si, MM.

BIO DATA PENULIS



Penulis memiliki berbagai disiplin ilmu yang diperoleh dari Universitas Diponegoro (UNDIP) Semarang. dan dari Universitas Kristen Satya Wacana (UKSW) Salatiga. Disiplin ilmu itu antara lain teknik elektro, komputer, manajemen, ilmu sosiologi dan ilmu hukum. Penulis memiliki pengalaman kerja pada industri elektronik dan sertifikasi keahlian dalam bidang Jaringan Internet, Telekomunikasi, Artificial Intelligence, Internet Of Things (IoT), Augmented Reality (AR), Technopreneurship, Internet Marketing dan bidang pengolahan dan analisa data (komputer statistik), Ilmu Perpajakan.

Penulis adalah pendiri dari Universitas Sains dan Teknologi Komputer (Universitas STEKOM) dan juga seorang dosen yang memiliki Jabatan Fungsional Akademik Lektor Kepala (Associate Professor) yang telah menghasilkan puluhan Buku Ajar ber ISBN, HAKI dari beberapa karya cipta dan Hak Paten pada produk IPTEK. Sejak tahun 2023 penulis tercatat sebagai Dosen luar biasa di Fakultas Ekonomi & Bisnis (FEB) Universitas Diponegoro Semarang. Penulis juga terlibat dalam berbagai organisasi profesi dan industri yang terkait dengan dunia usaha dan industri, khususnya dalam pengembangan sumber daya manusia yang unggul untuk memenuhi kebutuhan dunia kerja secara nyata.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id