Statistika Inferential

by Thio Perdana

OUTLINE

Pengenalan Statistika Inferensial

Statistika inferensial adalah cabang statistika yang digunakan untuk membuat kesimpulan atau inferensi tentang populasi berdasarkan sampel yang diambil dari populasi tersebut. Ini melibatkan penggunaan data sampel untuk mencoba memahami dan membuat pernyataan tentang sifat-sifat populasi yang lebih luas. Beberapa konsep dasar statistika inferensial meliputi:

Sampel dan Populasi:

Populasi adalah kumpulan semua individu atau elemen yang relevan untuk studi atau analisis tertentu.

Sampel adalah subset dari populasi yang digunakan untuk analisis. Dalam statistika inferensial, kita mengambil sampel yang representatif dari populasi untuk membuat kesimpulan tentang populasi secara keseluruhan.

Parameter dan Statistik:

Parameter adalah ukuran atau karakteristik yang menggambarkan sifat populasi. Contoh parameter adalah rata-rata populasi, deviasi standar populasi, dan sebagainya.

Statistik adalah ukuran atau karakteristik yang dihitung berdasarkan sampel. Statistik digunakan untuk membuat estimasi tentang parameter populasi.

Inferensi Statistik:

Inferensi statistik melibatkan proses membuat kesimpulan tentang populasi berdasarkan informasi yang ditemukan dalam sampel.

Ini melibatkan penggunaan teknik seperti pengujian hipotesis, interval kepercayaan, regresi, dan analisis varians untuk mengambil kesimpulan yang kuat.

Statistika Deskriptif VS Statistika Inferensial

Tujuan Utama:

Statistika Deskriptif: Tujuannya adalah untuk merangkum, mengorganisir, dan menyajikan data secara informatif. Ini menggambarkan sifat data sampel.

Statistika Inferensial: Tujuannya adalah untuk membuat kesimpulan tentang populasi berdasarkan data sampel yang ada.

Contoh Hasil:

Statistika Deskriptif: Hasilnya termasuk rata-rata, median, deviasi standar, dan grafik yang menggambarkan data sampel.

Statistika Inferensial: Hasilnya mencakup estimasi parameter populasi, pengujian hipotesis, dan interval kepercayaan.

Penggunaan Data:

Statistika Deskriptif: Hanya menggambarkan data yang diamati dalam sampel.

Statistika Inferensial: Menggunakan data sampel untuk membuat kesimpulan tentang populasi yang lebih besar.

Contoh Metode:

Statistika Deskriptif: Menggunakan metode seperti pengukuran pusat dan penyebaran data.

Statistika Inferensial: Menggunakan metode seperti uji hipotesis, analisis regresi, dan metode inferensial lainnya.

Dengan kata lain, statistika deskriptif fokus pada penggambaran data sampel, sedangkan statistika inferensial digunakan untuk membuat kesimpulan yang lebih luas tentang populasi berdasarkan sampel. Keduanya penting dalam ilmu data, dan pemahaman statistika inferensial adalah kunci dalam pengambilan keputusan berdasarkan data dalam berbagai konteks, termasuk data science.

Probabilitas dan Distribusi Probabilitas

Probabilitas adalah konsep penting dalam statistika dan matematika yang membantu kita mengukur tingkat keyakinan terjadinya suatu peristiwa. Di bawah ini adalah konsep dasar probabilitas yang perlu Anda pahami:

Eksperimen: Probabilitas berkaitan dengan eksperimen atau peristiwa yang menghasilkan hasil. Contoh eksperimen termasuk melempar koin, mengambil kartu dari setumpuk kartu, atau memilih bola dari sebuah kotak.

Ruang Sampel (Sample Space): Ruang sampel adalah himpunan semua hasil yang mungkin dari eksperimen tersebut. Ini mencakup semua hasil yang potensial. Misalnya, ruang sampel dalam pelemparan koin adalah {H, T}, yang merupakan singkatan dari "Heads" (muka) dan "Tails" (ekor).

Kejadian (Event): Kejadian adalah subset dari ruang sampel yang mewakili hasil-hasil tertentu yang ingin kita amati atau hitung probabilitasnya. Misalnya, jika kita ingin menghitung probabilitas munculnya "Heads" dalam pelemparan koin, kejadian tersebut adalah {H}.

Probabilitas Kejadian (Probability of an Event): Probabilitas suatu kejadian adalah ukuran tingkat keyakinan bahwa kejadian tersebut akan terjadi. Probabilitas kejadian A biasanya dilambangkan sebagai P(A) dan nilainya berkisar antara 0 hingga 1, di mana 0 berarti kejadian itu pasti tidak terjadi, dan 1 berarti kejadian itu pasti terjadi.

Operasi Probabilitas: Terdapat berbagai operasi probabilitas, seperti:

Gabungan (Union): Probabilitas bahwa setidaknya satu dari dua kejadian terjadi.

Irisan (Intersection): Probabilitas bahwa kedua kejadian terjadi secara bersamaan.

Komplementer (Complement): Probabilitas bahwa kejadian A tidak terjadi.

Asumsi dasar probabilitas: Dalam banyak situasi, kita berasumsi bahwa setiap hasil dalam ruang sampel memiliki probabilitas yang sama. Ini dikenal sebagai asumsi kesetaraan probabilitas atau model probabilitas seragam.

Frekuensi vs. Probabilitas: Probabilitas juga dapat didekati dengan menghitung frekuensi relatif dari kejadian tertentu dalam banyak percobaan yang dilakukan. Semakin banyak percobaan, semakin mendekati probabilitas teoritis.

Hukum Probabilitas:

Hukum Keseluruhan Probabilitas: Ini menyatakan bahwa probabilitas keseluruhan suatu ruang sampel adalah 1. Dengan kata lain, jika kita menjumlahkan probabilitas semua hasil dalam ruang sampel, itu harus sama dengan 1.

Hukum Penjumlahan Probabilitas: Ini berlaku untuk kejadian yang saling eksklusif dan bersifat bersamaan. Probabilitas gabungan dua kejadian yang saling eksklusif adalah jumlah probabilitas masing-masing kejadian.

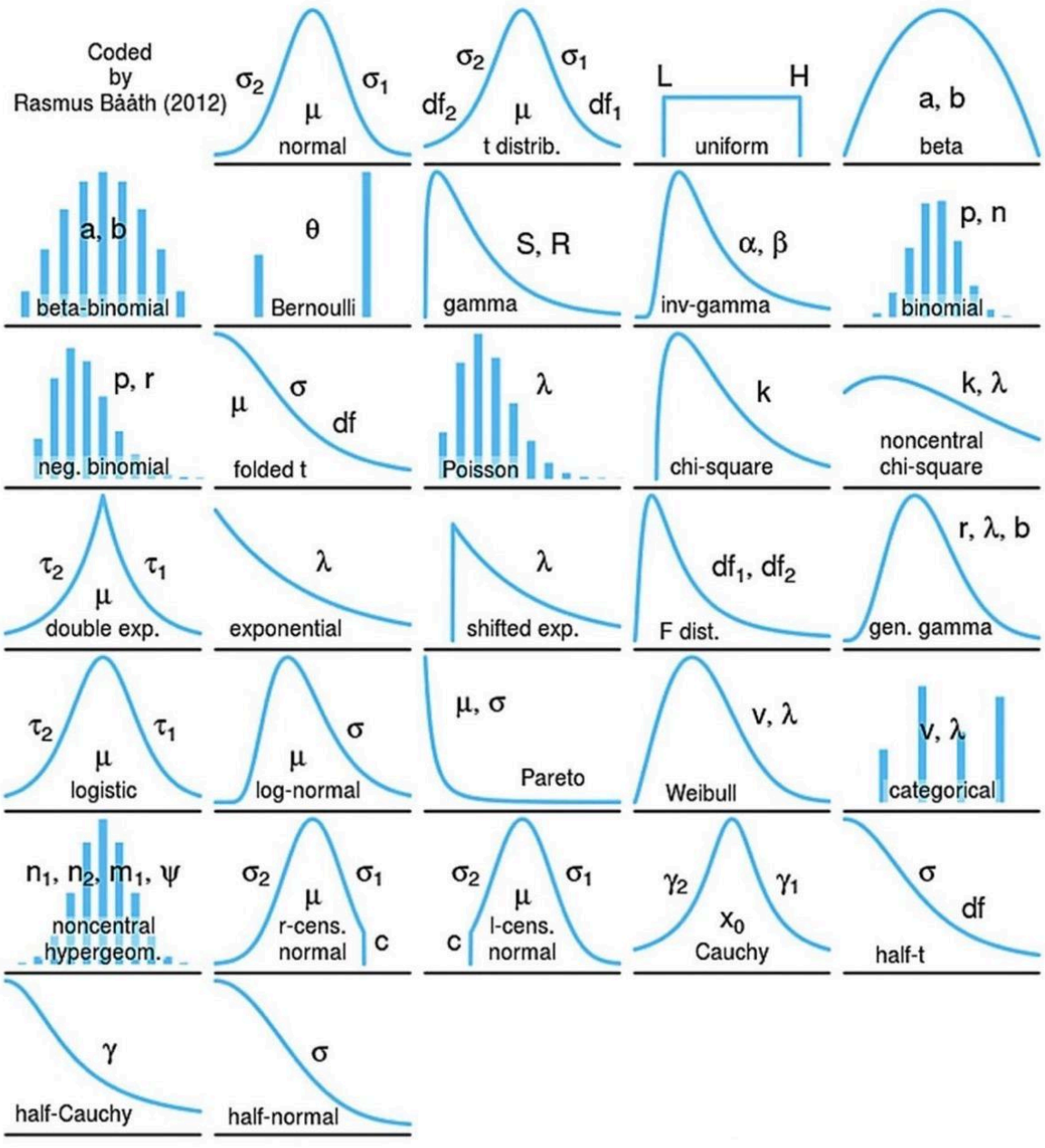
Kemungkinan dan Peluang (Odds): Selain probabilitas, ada juga konsep kemungkinan dan peluang. Kemungkinan adalah rasio probabilitas kejadian terjadi dengan probabilitas kejadian tidak terjadi, sementara peluang adalah rasio probabilitas kejadian terjadi dengan probabilitas kejadian lain terjadi.

Memahami konsep dasar probabilitas adalah langkah awal yang penting dalam statistika dan analisis data. Ini memungkinkan kita untuk membuat prediksi, mengambil keputusan yang lebih baik, dan memahami peluang dalam berbagai konteks. Probabilitas sangat berguna dalam ilmu data, pemodelan statistik, dan pengambilan keputusan.

Distribusi Probabilitas

Distribusi probabilitas adalah konsep dalam statistika yang digunakan untuk menggambarkan bagaimana probabilitas terdistribusi di seluruh kemungkinan hasil atau nilai dari suatu variabel acak. Ini adalah cara untuk menyatakan seberapa mungkin berbagai hasil atau nilai dapat terjadi dalam suatu situasi tertentu.

Probability Distributions



sumber : https://www.kaggle.com/discussions/general/366492

Dalam konteks distribusi probabilitas, terdapat dua jenis distribusi utama: distribusi probabilitas diskrit dan distribusi probabilitas kontinu.

Distribusi Probabilitas Diskrit: Distribusi probabilitas diskrit digunakan ketika variabel acak hanya dapat mengambil sejumlah nilai tertentu atau terhitung. Contoh distribusi probabilitas diskrit termasuk distribusi binomial (digunakan untuk menghitung probabilitas sukses dan gagal dalam serangkaian uji coba), distribusi Poisson (digunakan untuk menggambarkan jumlah peristiwa dalam interval waktu tertentu), distribusi eksponensial diskrit, dan lain-lain.

Distribusi Probabilitas Kontinu: Distribusi probabilitas kontinu digunakan ketika variabel acak dapat mengambil nilai dalam rentang kontinu. Salah satu distribusi probabilitas kontinu yang paling terkenal adalah distribusi normal (dikenal sebagai distribusi Gaussian). Distribusi normal digunakan untuk menggambarkan banyak fenomena di alam, seperti tinggi manusia, suhu, dan banyak lainnya. Selain itu, ada distribusi probabilitas kontinu lainnya seperti distribusi t-Student, distribusi chi-kuadrat, dan distribusi eksponensial kontinu.

Tujuan utama dari distribusi probabilitas adalah untuk memodelkan dan memahami tingkat probabilitas berbagai hasil atau nilai yang mungkin terjadi dalam situasi tertentu. Ini memungkinkan kita untuk membuat prediksi, melakukan analisis statistik, dan membuat keputusan berdasarkan pengetahuan tentang cara probabilitas terdistribusi.

Dalam praktiknya, kita menggunakan fungsi probabilitas untuk menghitung probabilitas tertentu, seperti probabilitas bahwa suatu variabel acak akan jatuh dalam rentang tertentu. Pemahaman distribusi probabilitas sangat penting dalam statistika dan analisis data, karena membantu kita membuat model dan mengambil keputusan berdasarkan pengetahuan tentang sebaran probabilitas dalam situasi tertentu.

Distribusi Probabilitas Diskrit

Distribusi probabilitas diskrit adalah jenis distribusi probabilitas yang digunakan untuk menggambarkan sebaran probabilitas dari variabel acak diskrit. Variabel acak diskrit adalah variabel yang hanya dapat mengambil nilai-nilai tertentu atau terhitung, bukan dalam rentang kontinu. Dalam hal ini, kita akan membahas konsep distribusi probabilitas diskrit dan dua distribusi diskrit yang paling umum: distribusi binomial dan distribusi Poisson.

Konsep Dasar Distribusi Probabilitas Diskrit:

1. **Variabel Acak Diskrit:** Variabel acak diskrit adalah variabel yang dapat mengambil sejumlah nilai yang terbatas atau terhitung. Contohnya termasuk jumlah mata dadu yang mungkin muncul dalam satu lemparan, jumlah orang yang tiba di lokasi tertentu dalam interval waktu tertentu, atau jumlah kejadian tertentu dalam suatu periode.
2. **Hasil atau Nilai yang Mungkin:** Dalam distribusi probabilitas diskrit, kita menentukan semua hasil atau nilai yang mungkin dari variabel acak diskrit tersebut. Ini membentuk ruang sampel (sample space).
3. **Probabilitas Setiap Hasil:** Kemudian, kita menetapkan probabilitas untuk masing-masing hasil atau nilai dalam ruang sampel. Ini adalah probabilitas bahwa variabel acak akan mengambil nilai tertentu.
4. **Fungsi Probabilitas:** Fungsi probabilitas adalah aturan yang mengaitkan hasil atau nilai tertentu dengan probabilitas munculnya hasil tersebut. Ini memungkinkan kita untuk menghitung probabilitas berbagai skenario.

Distribusi Binomial

Distribusi binomial adalah salah satu distribusi probabilitas diskrit yang paling umum digunakan. Ini digunakan ketika kita memiliki dua hasil yang mungkin dalam setiap percobaan, biasanya disebut sebagai "sukses" dan "gagal," dan ketika percobaan-percobaan ini bersifat independen. Distribusi binomial memiliki beberapa karakteristik penting:

1. **Parameter:**

- **n:** Jumlah total percobaan atau uji coba.
- **p:** Probabilitas sukses dalam setiap uji coba.

2. **Fungsi Probabilitas:**
   - Fungsi probabilitas binomial memberikan probabilitas kumulatif bahwa sejumlah k percobaan berhasil dalam n percobaan.
3. **Notasi:**
   - Fungsi probabilitas binomial ditulis sebagai P(X = k) = (n choose k) * p^k * (1-p)^(n-k), di mana "n choose k" merupakan simbol kombinasi.
4. **Contoh:**
   - Misalkan Anda melempar koin (bernilai 1 jika muncul "muka" dan 0 jika "ekor") 10 kali, dan Anda ingin menghitung probabilitas mendapatkan 5 "muka" dari 10 percobaan dengan probabilitas "muka" sekitar 0,5. Ini adalah contoh distribusi binomial.

Distribusi Poisson

Distribusi Poisson digunakan untuk menggambarkan jumlah peristiwa yang terjadi dalam interval waktu atau ruang tertentu, ketika peristiwa-peristiwa tersebut terjadi dengan tingkat frekuensi yang konstan dalam interval tersebut. Beberapa karakteristik distribusi Poisson adalah:

1. **Parameter:**
   - **λ (lambda):** Ini adalah tingkat frekuensi peristiwa dalam interval tertentu.
2. **Fungsi Probabilitas:**
   - Fungsi probabilitas Poisson memberikan probabilitas bahwa sejumlah k peristiwa terjadi dalam interval yang ditentukan.
3. **Notasi:**
   - Fungsi probabilitas Poisson ditulis sebagai P(X = k) = (e^(-λ) * λ^k) / k!, di mana e adalah bilangan Euler (sekitar 2.71828).
4. **Contoh:**
   - Misalnya, jika kita ingin menghitung probabilitas bahwa ada 3 pesawat yang lewat melintasi langit pada suatu interval waktu tertentu, dengan tingkat frekuensi rata-rata 2 pesawat per jam, maka kita dapat menggunakan distribusi Poisson dengan λ = 2.

Distribusi binomial dan Poisson adalah dua distribusi probabilitas diskrit yang sering digunakan dalam berbagai aplikasi, termasuk ilmu data, manajemen rantai pasokan, analisis risiko, dan banyak lagi. Memahami parameter dan konsep dasar dari kedua distribusi ini sangat penting dalam menganalisis data yang melibatkan perhitungan diskrit.

Distribusi Probabilitas Kontinu

Distribusi probabilitas kontinu adalah jenis distribusi probabilitas yang digunakan untuk menggambarkan sebaran probabilitas dari variabel acak kontinu. Variabel acak kontinu adalah variabel yang dapat mengambil nilai dalam rentang kontinu, tidak hanya dalam sejumlah nilai tertentu seperti yang terjadi dalam variabel acak diskrit. Dalam hal ini, kita akan membahas konsep distribusi probabilitas kontinu dan fokus pada distribusi normal sebagai contoh utama.

Konsep Dasar Distribusi Probabilitas Kontinu:

1. **Variabel Acak Kontinu:** Variabel acak kontinu adalah variabel yang dapat mengambil nilai dalam rentang yang terus menerus. Ini bisa berupa nilai numerik seperti tinggi manusia, suhu, berat badan, atau banyak variabel lain yang tidak terbatas pada nilai-nilai tertentu.
2. **Hasil atau Nilai yang Mungkin:** Dalam distribusi probabilitas kontinu, kita mendefinisikan semua nilai yang mungkin dari variabel acak dalam rentang yang terus menerus. Ini membentuk rentang nilai yang mungkin, dan probabilitas diukur sebagai sebaran probabilitas di dalam rentang ini.
3. **Fungsi Probabilitas (Probability Density Function):** Distribusi probabilitas kontinu dijelaskan oleh fungsi probabilitas yang dikenal sebagai "Probability Density Function" (PDF). Fungsi ini memberikan probabilitas relatif dari suatu nilai atau rentang nilai tertentu.

Distribusi Normal (Gaussian):

Distribusi normal (atau disebut juga distribusi Gaussian) adalah salah satu distribusi probabilitas kontinu yang paling umum digunakan dan penting dalam statistika. Beberapa karakteristik distribusi normal adalah:

- **Parameter:** Distribusi normal memiliki dua parameter utama:
    - **Mean (μ):** Ini adalah nilai rata-rata dari distribusi dan menentukan pusatnya.
    - **Standard Deviation (σ):** Ini mengukur sebaran data atau "spread" dari distribusi.
- **Fungsi Probabilitas (PDF):** Fungsi probabilitas distribusi normal memiliki bentuk lonceng (bell-shaped) dan terkenal dengan kurva berbentuk lonceng simetris.

- **Notasi:** Distribusi normal sering dinotasikan sebagai N(μ, σ^2), di mana μ adalah nilai rata-rata dan σ^2 adalah varians (kuadrat dari deviasi standar).
- **Sifat-Sifat Distribusi Normal:** Distribusi normal memiliki beberapa sifat penting, termasuk simetri, rata-rata, dan median yang sama, serta sebagian besar data berada dalam satu, dua, atau tiga deviasi standar dari rata-rata.

Contoh Distribusi Probabilitas Kontinu:

Misalkan Anda ingin mengukur suhu tubuh manusia dalam populasi. Anda dapat menggunakan distribusi normal untuk menggambarkan distribusi suhu tubuh manusia, dengan rata-rata suhu tubuh manusia sebagai μ dan sebaran suhu tubuh sebagai σ. Dengan informasi ini, Anda dapat menghitung probabilitas suatu individu memiliki suhu tertentu.

Distribusi probabilitas kontinu sangat penting dalam berbagai aplikasi statistik dan ilmu data, karena banyak fenomena di alam mengikuti distribusi normal atau distribusi probabilitas kontinu lainnya. Pemahaman konsep dasar dan parameter dari distribusi probabilitas kontinu adalah langkah penting dalam menganalisis data kontinu.

Pengujian Hipotesis

Pengujian hipotesis adalah prosedur statistik yang digunakan untuk mengambil keputusan berdasarkan data yang kita miliki terkait dengan suatu masalah atau pertanyaan penelitian. Ini adalah cara untuk menguji apakah perbedaan atau hubungan yang diamati antara variabel-variabel dalam sampel kita adalah hasil dari faktor sebenarnya atau hanya kebetulan.

Pengertian Hipotesis Nol dan Hipotesis Alternatif:

- **Hipotesis Nol (H0):** Ini adalah pernyataan awal yang menyatakan tidak ada perbedaan, efek, atau hubungan yang signifikan dalam populasi. Hipotesis nol merupakan dasar untuk pengujian, dan kita berusaha untuk menguji apakah kita dapat menolak hipotesis nol dengan data yang kita kumpulkan.
- **Hipotesis Alternatif (H1 atau Ha):** Ini adalah pernyataan yang mengindikasikan bahwa terdapat perbedaan, efek, atau hubungan yang signifikan dalam populasi. Hipotesis alternatif merupakan apa yang kita ingin menunjukkan dengan pengujian.

Langkah-Langkah Pengujian Hipotesis:

1. **Tentukan Hipotesis:**
    - Tentukan hipotesis nol (H0) dan hipotesis alternatif (H1).
2. **Pilih Tingkat Signifikansi (α):**
    - Tingkat signifikansi adalah batas yang menentukan seberapa kecil probabilitas kesalahan tipe I yang Anda bersedia terima. Biasanya, α diatur pada tingkat 0,05 (5%).
3. **Kumpulkan Data dan Hitung Statistik Uji:**
    - Kumpulkan data yang relevan.
    - Hitung statistik uji berdasarkan data Anda. Statistik uji berbeda-beda tergantung pada jenis pengujian yang Anda lakukan (uji Z, uji t, uji chi-kuadrat, dll.).
4. **Hitung Nilai-p (p-value):**
    - Nilai-p adalah probabilitas mengamati data yang setidaknya ekstrem seperti yang Anda peroleh jika hipotesis nol benar. Semakin kecil nilai-p, semakin kuat bukti menolak hipotesis nol.
5. **Tentukan Keputusan:**
    - Jika nilai-p lebih kecil dari tingkat signifikansi (α), Anda dapat menolak hipotesis nol dan menerima hipotesis alternatif.
    - Jika nilai-p lebih besar dari α, Anda gagal menolak hipotesis nol.

Uji Z, Uji t, Uji Chi-Kuadrat:

- **Uji Z:** Digunakan ketika kita memiliki data dengan jumlah observasi yang besar dan kita ingin menguji perbedaan rata-rata atau proposi dalam sampel.
- **Uji t:** Digunakan ketika kita memiliki data dengan jumlah observasi yang kecil dan ingin menguji perbedaan rata-rata antara dua sampel atau populasi.
- **Uji Chi-Kuadrat:** Digunakan untuk menguji hubungan antara variabel-variabel kategoris atau untuk menguji kesesuaian antara distribusi data observasi dan distribusi yang diharapkan.

Error Tipe I dan Tipe II:

- **Error Tipe I (α):** Error Tipe I terjadi ketika kita menolak hipotesis nol padahal sebenarnya hipotesis nol adalah benar. Tingkat signifikansi (α) adalah probabilitas error Tipe I.
- **Error Tipe II (β):** Error Tipe II terjadi ketika kita gagal menolak hipotesis nol padahal sebenarnya hipotesis alternatif adalah benar. Tingkat β adalah probabilitas error Tipe II.

Contoh Soal:

Sebuah perusahaan produsen lampu neon mengklaim bahwa umur rata-rata lampu neon yang mereka hasilkan adalah 10.000 jam. Anda adalah seorang peneliti yang ingin memeriksa apakah klaim perusahaan tersebut benar. Anda mengambil sampel 100 lampu neon dan menemukan bahwa umur rata-rata lampu neon dalam sampel Anda adalah 9.800 jam dengan deviasi standar sampel sebesar 200 jam. Apakah Anda dapat mengatakan dengan tingkat signifikansi 0.05 bahwa klaim perusahaan tersebut tidak benar?

Langkah Penyelesaian:

1. Tentukan hipotesis nol (H0) dan hipotesis alternatif (H1).
    - H0: Umur rata-rata lampu neon sama dengan klaim perusahaan (μ = 10,000 jam).
    - H1: Umur rata-rata lampu neon tidak sama dengan klaim perusahaan (μ ≠ 10,000 jam).
2. Hitung nilai uji Z dengan rumus:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

di mana:

- ¯X¯ adalah rata-rata sampel.
- μ adalah nilai klaim perusahaan (10,000 jam).
- σ adalah deviasi standar populasi (tidak diketahui, sehingga kita mengganti dengan deviasi standar sampel, yaitu 200 jam).
- n adalah ukuran sampel (100 lampu neon).

3. Tentukan tingkat signifikansi (α), misalnya α=0.05.
4. Temukan zona kritis dengan mengacu pada distribusi Z (Z-distribution) atau gunakan tabel Z. Pada tingkat signifikansi ﷺ=0.05α=0.05, zona kritis akan berada pada kedua ekor distribusi Z.
5. Hitung nilai uji Z:
   Z = (9800−10000)/(200/(100**0.5)) = −200/20=−10
6. Bandingkan nilai uji Z dengan zona kritis. Jika nilai uji Z berada dalam zona kritis, tolak H0; jika tidak, gagalkan tolak H0.
   Dalam contoh ini, dengan Z = -10, kita akan jelas berada dalam zona kritis. Oleh karena itu, kita akan menolak H0.
7. Buat kesimpulan: Kami dapat menyimpulkan bahwa ada cukup bukti statistik yang mendukung klaim bahwa umur rata-rata lampu neon yang mereka hasilkan tidak sama dengan 10,000 jam dengan tingkat signifikansi 0.05.

Ini adalah contoh uji Z untuk satu sampel, yang digunakan untuk menguji perbedaan antara rata-rata sampel dan klaim populasi.

Memahami konsep pengujian hipotesis dan cara menguji hipotesis adalah keterampilan penting dalam statistika dan analisis data. Ini memungkinkan kita untuk membuat kesimpulan berdasarkan bukti statistik dan membuat keputusan yang berdasarkan data.

Interval Kepercayaan

Interval Kepercayaan (Confidence Interval) adalah konsep statistik yang digunakan untuk mengukur sejauh mana kita yakin tentang perkiraan suatu parameter dalam statistika, seperti rata-rata atau proporsi dalam populasi. Ini memberikan rentang nilai di mana kita berkeyakinan bahwa parameter sebenarnya berada, dengan tingkat kepercayaan tertentu.Interval kepercayaan adalah rentang nilai yang kita perkirakan mengandung parameter populasi dengan tingkat kepercayaan tertentu. Misalnya, jika kita

menghitung interval kepercayaan 95% untuk rata-rata suatu populasi, kita mengatakan bahwa kita yakin 95% bahwa nilai rata-rata populasi berada dalam interval tersebut.

Cara Menghitung Interval Kepercayaan

Proses menghitung interval kepercayaan bergantung pada jenis data dan parameter yang ingin diestimasi. Berikut adalah langkah-langkah umum yang digunakan:

Pilih Tingkat Kepercayaan (Confidence Level): Anda perlu memilih tingkat kepercayaan yang sesuai, misalnya, 95% atau 99%. Ini mengukur seberapa yakin Anda ingin menjadi terkait dengan interval kepercayaan Anda.

Kumpulkan Data:

Dalam kasus rata-rata, Anda perlu mengumpulkan sampel data yang mewakili populasi yang ingin Anda estimasi.

Hitung Estimasi Parameter: Hitung estimasi parameter yang ingin Anda ketahui, seperti rata-rata atau proporsi, dari sampel Anda.

Hitung Standar Error:

Standar error adalah pengukuran ketidakpastian dalam estimasi Anda. Ini bergantung pada jenis data dan parameter yang Anda estimasi.

Gunakan Distribusi Statistik: Anda akan menggunakan distribusi statistik tertentu (misalnya, distribusi normal atau distribusi t) dengan tingkat kepercayaan yang Anda pilih untuk menentukan z-score atau t-score kritis yang sesuai.

tabel untuk z-score : Universitas Muhammadiyah Malang

Hitung Interval Kepercayaan:

Hitung interval kepercayaan menggunakan rumus yang sesuai:

Untuk rata-rata, interval kepercayaan 95% dapat dihitung sebagai estimasi ± (z-score atau t-score kritis) × (standar error).

Untuk proporsi, interval kepercayaan 95% dapat dihitung sebagai estimasi ± (z-score kritis) × √[(estimasi × (1 - estimasi)) / ukuran sampel].

Interpretasi Interval Kepercayaan

- Misalkan Anda menghitung interval kepercayaan 95% untuk rata-rata suatu populasi (misalnya, rata-rata tinggi manusia).

- Interval kepercayaan tersebut mungkin adalah (160 cm, 170 cm).

- Interpretasinya adalah bahwa dengan tingkat kepercayaan 95%, kita yakin bahwa rata-rata tinggi populasi berada di antara 160 cm dan 170 cm.

- Jika Anda menghitung interval kepercayaan lebih sempit (misalnya, 165 cm hingga 167 cm), tingkat kepercayaan akan lebih tinggi (misalnya, 99%), tetapi rentang nilainya lebih sempit, sehingga estimasi lebih akurat tetapi lebih konservatif.

Contoh Soal

Sebuah penelitian dilakukan untuk mengestimasi rata-rata waktu tidur per malam bagi sekelompok siswa SMA. Sampel terdiri dari 50 siswa, dan rata-rata waktu tidur dalam sampel adalah 7,5 jam dengan deviasi standar 0,8 jam. Hitung interval kepercayaan 95% untuk rata-rata waktu tidur per malam di seluruh populasi siswa SMA.

Langkah Penyelesaian

1. **Tentukan Tingkat Kepercayaan (Confidence Level):** Tingkat kepercayaan yang diminta adalah 95%, sehingga tingkat signifikansi (α) adalah 5% (0,05) karena 100% - 95% = 5%.
2. **Hitung Standar Error (SE):**
    - Standar error (SE) dihitung dengan rumus: SE = (deviasi standar / √ukuran sampel).
    - SE = (0,8 / √50) ≈ 0,113.
3. **Tentukan Z-Score Kritis:**
    - Menggunakan tabel distribusi normal atau kalkulator statistik, temukan Z-score kritis yang sesuai untuk α/2 (0,025 karena 0,05/2) pada tingkat kepercayaan 95%. Z-score kritis sekitar ±1,96.
4. **Hitung Interval Kepercayaan:**
    - Interval kepercayaan 95% dapat dihitung sebagai berikut: (7,5 - 1,96 * 0,113, 7,5 + 1,96 * 0,113).
    - Interval kepercayaan adalah sekitar (7,28, 7,72) jam.

Jadi, dengan tingkat kepercayaan 95%, kita yakin bahwa rata-rata waktu tidur per malam bagi seluruh populasi siswa SMA berada dalam rentang 7,28 jam hingga 7,72 jam.

Interval kepercayaan adalah alat penting dalam statistika yang membantu kita mengukur ketidakpastian dalam estimasi dan memberikan informasi tentang sejauh mana kita bisa yakin tentang perkiraan

parameter. Semakin tinggi tingkat kepercayaan yang Anda pilih, semakin besar interval kepercayaan, dan sebaliknya. Ini membantu dalam pengambilan keputusan berdasarkan data yang dikumpulkan.

Regresi dan Korelasi

Regresi

Analisis Regresi Sederhana

Analisis regresi sederhana adalah teknik statistik yang digunakan untuk memahami hubungan antara dua variabel, di mana satu variabel adalah variabel independen (biasanya disebut sebagai "X") dan yang lain adalah variabel dependen (biasanya disebut sebagai "Y"). Tujuannya adalah untuk memprediksi atau menjelaskan perubahan dalam variabel dependen (Y) sebagai fungsi dari perubahan dalam variabel independen (X). Regresi sederhana menghasilkan persamaan garis regresi yang mencerminkan hubungan antara kedua variabel.

Contoh: Anda ingin memahami bagaimana hubungan antara jumlah jam belajar (X) dan nilai ujian (Y) dalam siswa. Dengan analisis regresi sederhana, Anda dapat memodelkan bagaimana perubahan jumlah jam belajar (X) memengaruhi nilai ujian (Y).

Analisis Regresi Berganda

Analisis regresi berganda adalah pengembangan dari regresi sederhana yang digunakan ketika Anda ingin memahami hubungan antara satu variabel dependen (Y) dan dua atau lebih variabel independen (X1, X2, X3, dst.). Analisis ini menghasilkan persamaan regresi yang memungkinkan Anda untuk

memprediksi variabel dependen berdasarkan beberapa variabel independen. Ini membantu dalam menjelaskan kontribusi masing-masing variabel independen terhadap variabel dependen.

Contoh Soal:

Seorang peneliti ingin memahami faktor-faktor yang memengaruhi nilai rumah di suatu kawasan. Dia mengumpulkan data dari 100 rumah yang termasuk luas tanah (variabel X1), jumlah kamar (variabel X2), usia rumah (variabel X3), dan nilai rumah (variabel Y). Berikut adalah data yang dikumpulkan:

Rumah 1: Luas Tanah = 1600 sqft, Jumlah Kamar = 3, Usia Rumah = 12 tahun, Nilai Rumah = $210,000

Rumah 2: Luas Tanah = 1900 sqft, Jumlah Kamar = 4, Usia Rumah = 8 tahun, Nilai Rumah = $245,000

Rumah 3: Luas Tanah = 1750 sqft, Jumlah Kamar = 3, Usia Rumah = 10 tahun, Nilai Rumah = $200,000

Rumah 4: Luas Tanah = 2050 sqft, Jumlah Kamar = 5, Usia Rumah = 6 tahun, Nilai Rumah = $260,000

Rumah 5: Luas Tanah = 1500 sqft, Jumlah Kamar = 3, Usia Rumah = 11 tahun, Nilai Rumah = $195,000

Rumah 6: Luas Tanah = 2200 sqft, Jumlah Kamar = 4, Usia Rumah = 7 tahun, Nilai Rumah = $270,000

Rumah 7: Luas Tanah = 1800 sqft, Jumlah Kamar = 3, Usia Rumah = 9 tahun, Nilai Rumah = $215,000

Rumah 8: Luas Tanah = 1950 sqft, Jumlah Kamar = 4, Usia Rumah = 5 tahun, Nilai Rumah = $250,000

Rumah 9: Luas Tanah = 1700 sqft, Jumlah Kamar = 3, Usia Rumah = 13 tahun, Nilai Rumah = $190,000

Rumah 10: Luas Tanah = 2100 sqft, Jumlah Kamar = 5, Usia Rumah = 8 tahun, Nilai Rumah = $255,000

Peneliti ingin melakukan analisis regresi berganda untuk memahami hubungan antara variabel luas tanah (X1), jumlah kamar (X2), dan usia rumah (X3) terhadap nilai rumah (Y).

Pertanyaan

1. Buatlah model regresi berganda untuk memprediksi nilai rumah berdasarkan luas tanah, jumlah kamar, dan usia rumah.
2. Hitung koefisien regresi untuk masing-masing variabel independen.
3. Interpretasikan hasil analisis regresi untuk menjelaskan hubungan antara variabel luas tanah, jumlah kamar, dan usia rumah terhadap nilai rumah.
4. Jika rumah memiliki luas tanah 1800 sqft, 4 kamar, dan usia rumah 12 tahun, berapa nilai perkiraan rumah tersebut?

Jawaban

import numpy as np

```
# Data

luas_tanah = np.array([1600, 1900, 1750, 2050, 1500, 2200, 1800, 1950, 1700, 2100])

jumlah_kamar = np.array([3, 4, 3, 5, 3, 4, 3, 4, 3, 5])

usia_rumah = np.array([12, 8, 10, 6, 11, 7, 9, 5, 13, 8])

nilai_rumah = np.array([210000, 245000, 200000, 260000, 195000, 270000, 215000, 250000, 190000, 255000])



# Membuat model regresi berganda

X = np.column_stack((luas_tanah, jumlah_kamar, usia_rumah))

X = np.column_stack((np.ones(len(luas_tanah)), X))  # Menambahkan kolom konstan (intercept)

Y = nilai_rumah



# Menghitung koefisien regresi dengan metode OLS (Least Squares)

beta = np.linalg.lstsq(X, Y, rcond=None)[0]
```

```
# Menampilkan koefisien regresi

print("Koefisien Regresi (beta):")

print("Intercept (beta0):", beta[0])

print("Koefisien Luas Tanah (beta1):", beta[1])

print("Koefisien Jumlah Kamar (beta2):", beta[2])

print("Koefisien Usia Rumah (beta3):", beta[3])


# Prediksi nilai rumah untuk rumah dengan luas tanah 1800 sqft, 4 kamar, dan usia rumah 12 tahun

rumah_prediksi = np.array([1, 1800, 4, 12])

nilai_prediksi = np.dot(rumah_prediksi, beta)

print("Prediksi Nilai Rumah:", nilai_prediksi)
```

Output

```
Koefisien Regresi (beta):

Intercept (beta0): 112582.1808544048

Koefisien Luas Tanah (beta1): 63.196040748139396

Koefisien Jumlah Kamar (beta2): 8670.409867983042

Koefisien Usia Rumah (beta3): -3695.6576352514244

Prediksi Nilai Rumah: 216668.8020499708
```

Korelasi

Korelasi adalah konsep dalam statistik yang digunakan untuk mengukur hubungan atau hubungan antara dua atau lebih variabel. Ini membantu kita memahami apakah ada korelasi atau sejauh mana dua variabel bergerak bersama-sama. Korelasi tidak menyiratkan kausalitas (sebab-akibat), tetapi hanya menunjukkan hubungan statistik antara variabel tersebut.

Ada beberapa jenis koefisien korelasi yang digunakan dalam analisis statistik, tetapi yang paling umum adalah **koefisien korelasi Pearson** (sering disebut sebagai "korelasi Pearson" atau "Pearson's r"). Mari kita bahas konsep ini secara lebih terperinci:

Korelasi Pearson

Korelasi Pearson adalah salah satu metode paling umum yang digunakan untuk mengukur hubungan antara dua variabel yang memiliki skala pengukuran interval atau rasio. Koefisien korelasi Pearson (r) berkisar antara -1 dan 1. Nilai positif menunjukkan hubungan positif, nilai negatif menunjukkan hubungan negatif, dan nilai mendekati 0 menunjukkan hubungan lemah.

Jika r = 1, ini menunjukkan hubungan sempurna positif, artinya ketika satu variabel naik, yang lain juga naik dengan tingkat yang sama.

Jika r = -1, ini menunjukkan hubungan sempurna negatif, artinya ketika satu variabel naik, yang lain turun dengan tingkat yang sama.

Jika r ≈ 0, ini menunjukkan hubungan lemah atau tidak ada hubungan.

Langkah-langkah dalam Menghitung Korelasi Pearson

1. **Persiapan Data:** Pertama, Anda harus memiliki data yang mengandung pasangan nilai dari kedua variabel yang ingin Anda korelasikan.
2. **Hitung Mean (Rata-Rata) dan Standar Deviasi:** Hitung rata-rata dan standar deviasi dari kedua variabel.
3. **Hitung Nilai Pearson (r):** Gunakan rumus korelasi Pearson untuk menghitung nilai r. Rumusnya adalah:
   $r = \Sigma((X - \bar{X})(Y - \bar{Y})) / [\sqrt{\Sigma(X - \bar{X})^2 * \Sigma(Y - \bar{Y})^2}]$
   Di mana:
   - X dan Y adalah nilai individu dari kedua variabel.
   - $\bar{X}$ dan $\bar{Y}$ adalah rata-rata dari masing-masing variabel.
4. **Interpretasi Hasil:** Setelah menghitung nilai r, Anda dapat menginterpretasikan hasilnya sesuai dengan rentang -1 hingga 1.

Contoh Soal

Seorang peneliti ingin menentukan apakah ada hubungan korelasi antara jumlah jam belajar per minggu dengan nilai ujian siswa. Peneliti mengumpulkan data dari 20 siswa yang mencakup jumlah jam belajar per minggu (X) dan skor ujian (Y). Berikut adalah data yang dikumpulkan:

Siswa 1: Jam Belajar = 10 jam, Skor Ujian = 80

Siswa 2: Jam Belajar = 12 jam, Skor Ujian = 85

Siswa 3: Jam Belajar = 8 jam, Skor Ujian = 75

Siswa 4: Jam Belajar = 15 jam, Skor Ujian = 92

Siswa 5: Jam Belajar = 9 jam, Skor Ujian = 78

Siswa 6: Jam Belajar = 11 jam, Skor Ujian = 88

Siswa 7: Jam Belajar = 7 jam, Skor Ujian = 70

Siswa 8: Jam Belajar = 14 jam, Skor Ujian = 90

Siswa 9: Jam Belajar = 6 jam, Skor Ujian = 72

Siswa 10: Jam Belajar = 13 jam, Skor Ujian = 86

Siswa 11: Jam Belajar = 9 jam, Skor Ujian = 78

Siswa 12: Jam Belajar = 10 jam, Skor Ujian = 80

Siswa 13: Jam Belajar = 8 jam, Skor Ujian = 75

Siswa 14: Jam Belajar = 16 jam, Skor Ujian = 94

Siswa 15: Jam Belajar = 12 jam, Skor Ujian = 85

Siswa 16: Jam Belajar = 14 jam, Skor Ujian = 90

Siswa 17: Jam Belajar = 7 jam, Skor Ujian = 70

Siswa 18: Jam Belajar = 11 jam, Skor Ujian = 88

Siswa 19: Jam Belajar = 5 jam, Skor Ujian = 68

Siswa 20: Jam Belajar = 13 jam, Skor Ujian = 86

Apakah ada hubungan korelasi antara jumlah jam belajar per minggu dan skor ujian siswa?

Jawaban

```
X = [10, 12, 8, 15, 9, 11, 7, 14, 6, 13, 9, 10, 8, 16, 12, 14, 7, 11, 5, 13]

Y = [80, 85, 75, 92, 78, 88, 70, 90, 72, 86, 78, 80, 75, 94, 85, 90, 70, 88, 68, 86]


n = len(X)  # Jumlah data


mean_X = sum(X) / n

mean_Y = sum(Y) / n


XY_sum = sum([X[i] * Y[i] for i in range(n)])


X_squared_sum = sum([X[i] ** 2 for i in range(n)])

Y_squared_sum = sum([Y[i] ** 2 for i in range(n)])
```

```
numerator = n * XY_sum - sum(X) * sum(Y)

denominator = ((n * X_squared_sum - sum(X) ** 2) * (n * Y_squared_sum - sum(Y) ** 2)) ** 0.5


r = numerator / denominator


df = n - 2


t = r * ((df / (1 - r**2)) ** 0.5)


alpha = 0.05

t_critical = 2.101


if abs(t) > t_critical:

    print("Ada hubungan korelasi yang signifikan antara jumlah jam belajar dan skor ujian.")

else:

    print("Tidak ada hubungan korelasi yang signifikan antara jumlah jam belajar dan skor ujian.")
```

Output

```
Ada hubungan korelasi yang signifikan antara jumlah jam belajar dan skor ujian.
```

Hal yang Perlu Diingat:

Korelasi tidak menyiratkan kausalitas. Bahkan jika dua variabel berkorelasi kuat, ini tidak berarti bahwa satu variabel menyebabkan yang lain.

Korelasi hanya mengukur hubungan linear antara variabel. Beberapa hubungan non-linear mungkin tidak terdeteksi dengan korelasi Pearson.

Ketidakcocokan dalam distribusi data atau keberadaan outlier dapat memengaruhi hasil korelasi.

Korelasi adalah alat yang berguna untuk memahami hubungan antara variabel dalam data, tetapi perlu diingat bahwa korelasi sendiri tidak memberikan penjelasan sebab-akibat atau menentukan hubungan yang lebih kompleks. Itu adalah alat awal yang dapat digunakan untuk menyelidiki dan memahami data Anda.

Analisis regresi dan korelasi adalah alat penting dalam statistika yang membantu kita memahami hubungan antara variabel dalam data. Mereka digunakan dalam berbagai disiplin ilmu untuk membuat prediksi, menjelaskan hubungan, dan membuat keputusan berdasarkan data.

Analisis Varians (ANOVA)

Analisis Varians (ANOVA) adalah teknik statistik yang digunakan untuk membandingkan rata-rata dari tiga atau lebih kelompok atau perlakuan yang berbeda. Ini digunakan untuk menentukan apakah terdapat perbedaan yang signifikan antara rata-rata kelompok tersebut. ANOVA mengukur variasi antara kelompok (variabilitas antar kelompok) dan variasi dalam kelompok (variabilitas dalam kelompok) untuk menentukan apakah variasi antara kelompok lebih besar dari variasi dalam kelompok.

Jenis-jenis ANOVA

ANOVA Satu Arah (One-Way ANOVA):

Digunakan ketika ada satu variabel dependen dan satu variabel independen (faktor) yang memiliki tiga atau lebih kelompok atau tingkatan.

Contoh: Membandingkan rata-rata hasil tes antara tiga kelompok siswa yang berbeda (A, B, C).

ANOVA Dua Arah (Two-Way ANOVA):

Digunakan ketika ada dua variabel independen (faktor) yang mempengaruhi variabel dependen.

Dua faktor tersebut dapat bersifat independen (tidak saling bergantung) atau bersifat interaksi (salah satu faktor memengaruhi pengaruh faktor lainnya).

Contoh: Membandingkan rata-rata hasil tes antara tiga kelompok siswa (A, B, C) yang mungkin juga terpengaruh oleh jenis pelatihan (X, Y).

Interpretasi Hasil ANOVA

Hasil ANOVA menghasilkan nilai F-statistic dan p-value. Interpretasi hasil ANOVA melibatkan beberapa langkah:

Pernyataan Nol (H0) dan Hipotesis Alternatif (H1):

H0: Tidak ada perbedaan yang signifikan antara kelompok.

H1: Terdapat perbedaan yang signifikan antara kelompok.

F-statistic:

F-statistic adalah statistik uji yang mengukur perbandingan antara variabilitas antar kelompok dan variabilitas dalam kelompok.

Semakin besar nilai F-statistic, semakin besar kemungkinan terdapat perbedaan signifikan antara kelompok.

P-value:

P-value adalah probabilitas mendapatkan hasil seperti yang diamati jika tidak ada perbedaan antara kelompok.

Jika p-value kecil (biasanya kurang dari tingkat signifikansi yang telah ditentukan, misalnya 0.05), maka kita menolak H0 dan menyimpulkan bahwa terdapat perbedaan yang signifikan antara kelompok.

Post Hoc Tests (Tes Lanjutan):

Jika ANOVA menunjukkan adanya perbedaan signifikan, tes lanjutan (seperti uji t atau uji Bonferroni) dapat digunakan untuk menentukan kelompok mana yang berbeda satu sama lain.

ANOVA adalah alat statistik yang penting untuk memahami perbedaan antara kelompok dan digunakan dalam berbagai disiplin ilmu, termasuk ilmu sosial, ilmu alam, dan eksperimen laboratorium. Analisis ini membantu peneliti menentukan apakah perbedaan antara kelompok adalah hasil kebetulan atau ada faktor yang signifikan yang memengaruhi perbedaan tersebut.