Basic Statistics

14 Mar 2024

Agenda

- 1. Introduction to Statistics
- 2. Statistic data types
- 3. Descriptive Statistics
- 4. Correlation and Causality
- 5. Probability and Distribution

INTRODUCTION

Statistik

- Statistik dapat diartikan sebagai cabang dari matematika yang berfokus pada pengumpulan, analisis, interpretasi, dan presentasi data.
- Statistik digunakan untuk menggambarkan karakteristik dari kumpulan data dan untuk membuat inferensi tentang populasi berdasarkan sampel data.

Statistik

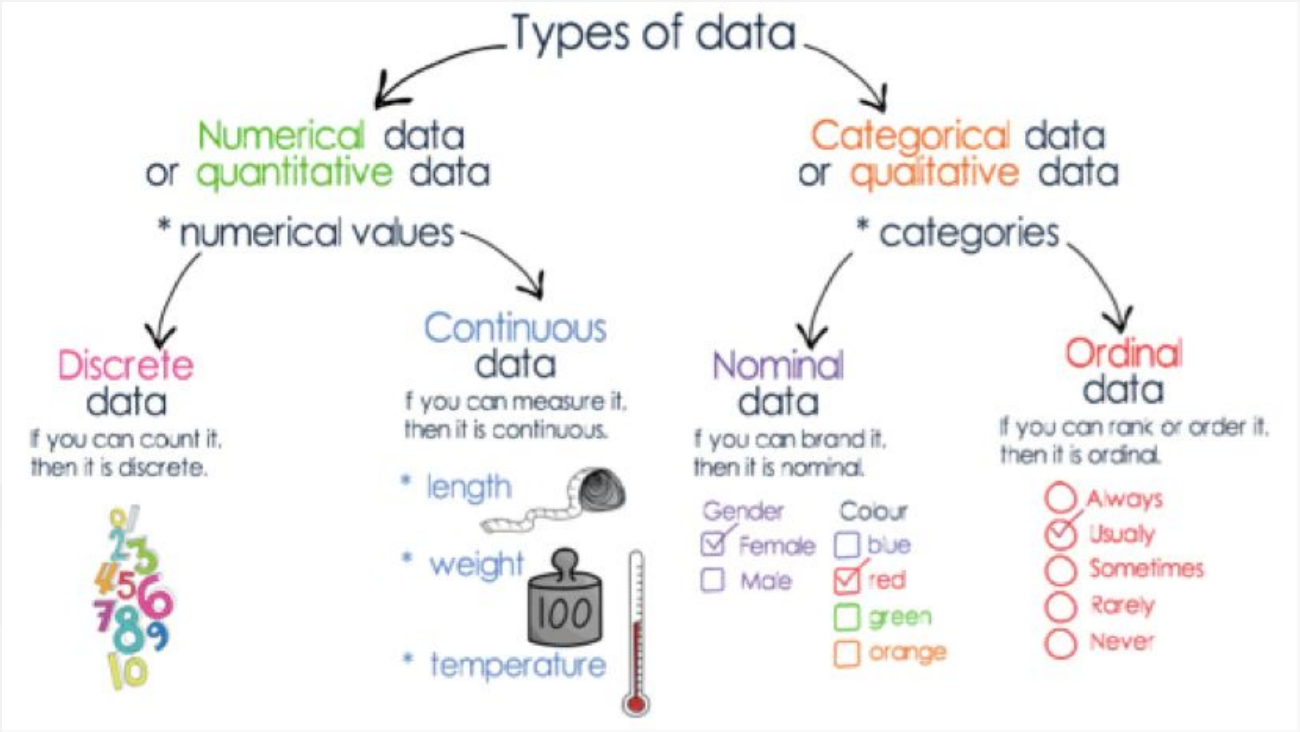
- Statistik dapat diartikan sebagai cabang dari matematika yang berfokus pada pengumpulan, analisis, interpretasi, dan presentasi data.
- Statistik digunakan untuk menggambarkan karakteristik dari kumpulan data dan untuk membuat inferensi tentang populasi berdasarkan sampel data.
- Statistik adalah fondasi penting dalam Data Science, memungkinkan data scientist untuk melakukan analisis data yang mendalam, model prediktif, dan interpretasi hasil dengan cara yang ilmiah dan terukur.

Tipe Data Statistik

Tipe Data Statistik

- Dalam analisis data, memahami jenis data adalah langkah awal yang penting karena menentukan teknik statistik yang akan digunakan dan bagaimana data tersebut harus diolah.
- Ada dua kategori utama data yang dikenal dalam statistik: data kategorikal dan data numerik.

Tipe Data Statistik



Categorical / Qualitative

- Data kategorikal, juga dikenal sebagai data kualitatif, mengklasifikasikan individu atau item ke dalam kelompok atau kategori yang berbeda.
    - **Nominal**: Data nominal merupakan jenis data kategorikal **tanpa urutan atau hierarki**. Contohnya termasuk jenis kelamin, warna mobil, atau negara asal.
    - **Ordinal**: Data ordinal adalah data kategorikal dengan **urutan atau peringkat**. Contohnya termasuk tingkat pendidikan, skala kepuasan, atau tingkat keparahan penyakit.

Numeric / Quantitative

- Data numerik, atau data kuantitatif, adalah data yang dapat dihitung dan biasanya dinyatakan dalam angka.
    - **Diskrit/Discrete**: Data diskrit adalah data numerik yang nilainya terbatas dan dapat dihitung. Ini biasanya merupakan hasil dari penghitungan, seperti jumlah anak dalam keluarga atau jumlah mobil dalam parkiran.
    - **Kontinu/Continue**: Data kontinu adalah data numerik yang nilainya dapat berubah-ubah dan tidak terbatas, biasanya merupakan hasil dari pengukuran. Contohnya termasuk berat badan, tinggi, atau suhu.

Continuous

- Dalam statistik, data kontinu dibagi menjadi dua sub-kategori: data interval dan data rasio.
- **Interval** : selisih (interval) antara dua nilai memiliki arti, tetapi **tidak memiliki titik nol mutlak** atau inheren. Ini berarti bahwa nilai nol dalam skala interval tidak menunjukkan ketiadaan nilai tersebut, melainkan sebuah titik acak dalam skala, contoh:
    - Suhu dalam derajat Celsius atau Fahrenheit, di mana 0 derajat tidak menunjukkan ketiadaan suhu tetapi merupakan nilai dalam skala.
    - Tahun, di mana nol tidak berarti "tidak ada tahun" tetapi hanya sebuah titik dalam sistem penanggalan.

Continuous

- **Ratio** : data kontinu yang memiliki semua sifat data interval dengan **tambahan titik nol mutlak**. Titik nol menunjukkan ketiadaan nilai yang diukur. Keberadaan titik nol ini memungkinkan pengukuran absolut dan memungkinkan semua operasi matematika, termasuk perkalian dan pembagian, contoh:
  - Berat, di mana 0 menunjukkan ketiadaan berat.
  - Pendapatan, di mana 0 menunjukkan ketiadaan pendapatan.
  - Jarak, di mana 0 menunjukkan tidak ada jarak.

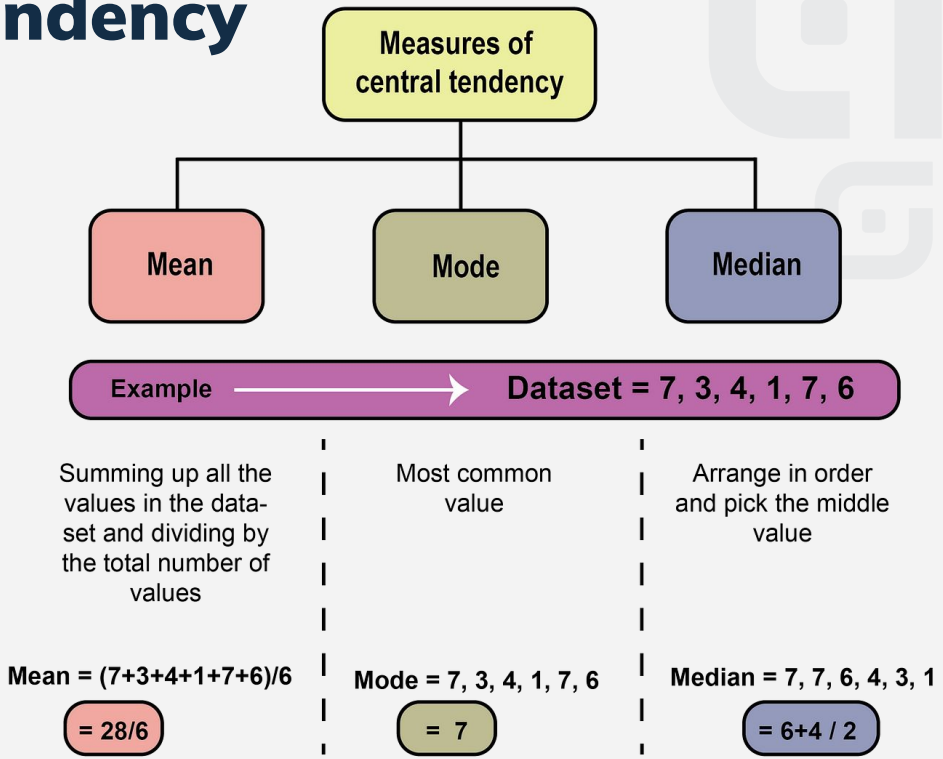
Statistik Deskriptif

Statistik Deskriptif

- Statistik Deskriptif adalah cabang statistik yang berfokus pada pengumpulan, penyajian, dan karakterisasi set data melalui ringkasan numerik dan visualisasi.
- Tujuannya adalah untuk memberikan deskripsi atau ringkasan yang jelas dari data yang diamati, yang memungkinkan kita untuk memahami secara cepat dan efektif struktur dan aspek penting dari dataset tersebut.

Measures of Central Tendency

- Perhitungan nilai yang mencoba mendeskripsikan pusat kumpulan data, dimana data cenderung berkumpul.



Measures of Central Tendency : Mean

- Mean (Rata-rata): Total dari semua nilai data dibagi dengan jumlah nilai. **Sensitif terhadap nilai ekstrem.**

```python
mean = sum(data) / len(data)
```

```
#"Mean:Computed column-wise
meanData = dataFrame.mean()
```

```
#"Mean:Computed row-wise:
meanData = dataFrame.mean(axis=1)
```

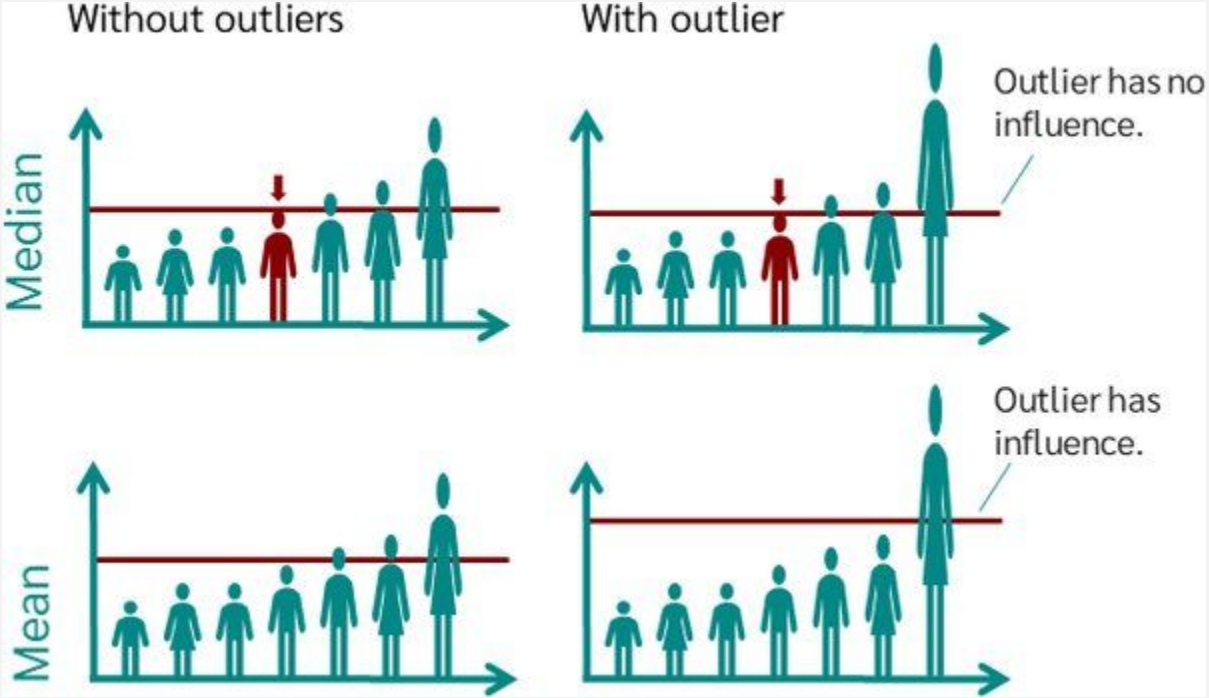
Measures of Central Tendency : Median

- Nilai tengah dataset yang telah diurutkan. **Lebih tidak sensitif terhadap nilai ekstrem** dibandingkan mean.

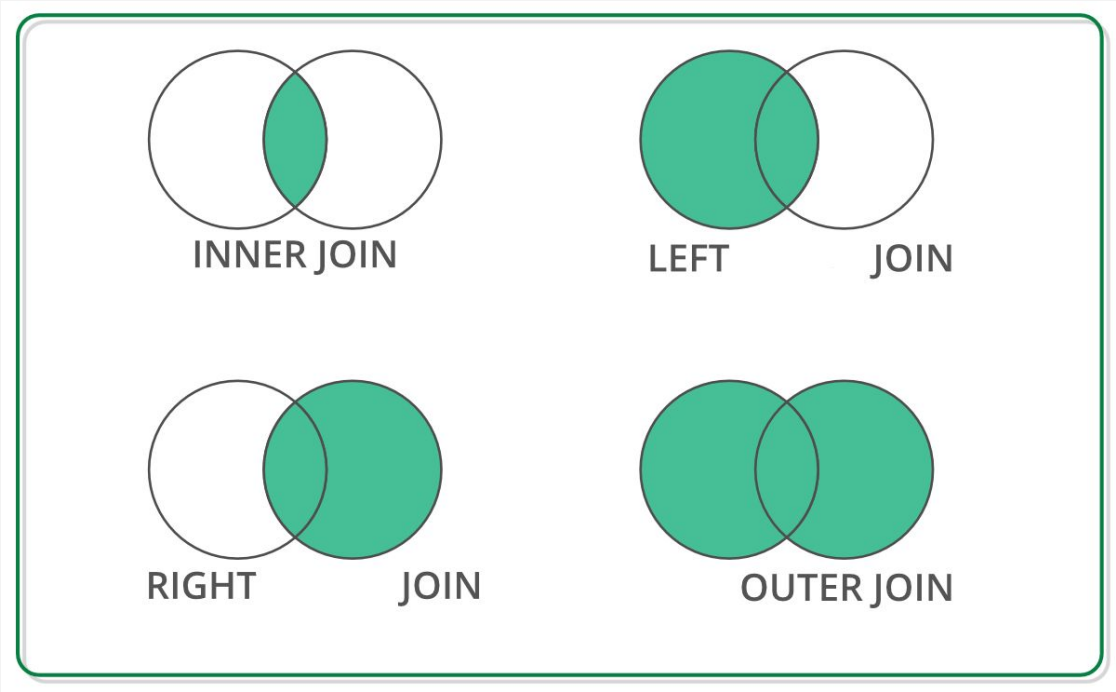
```
# Median:Computed column-wise:
medianData = dataFrame.median()
```

```
# Median:Computed row-wise:
medianData = dataFrame.median(axis=1)
```

Measures of Central Tendency : Mean VS Median



COMBINING DATAFRAME : Merge



Measures of Central Tendency : Mode / Modus

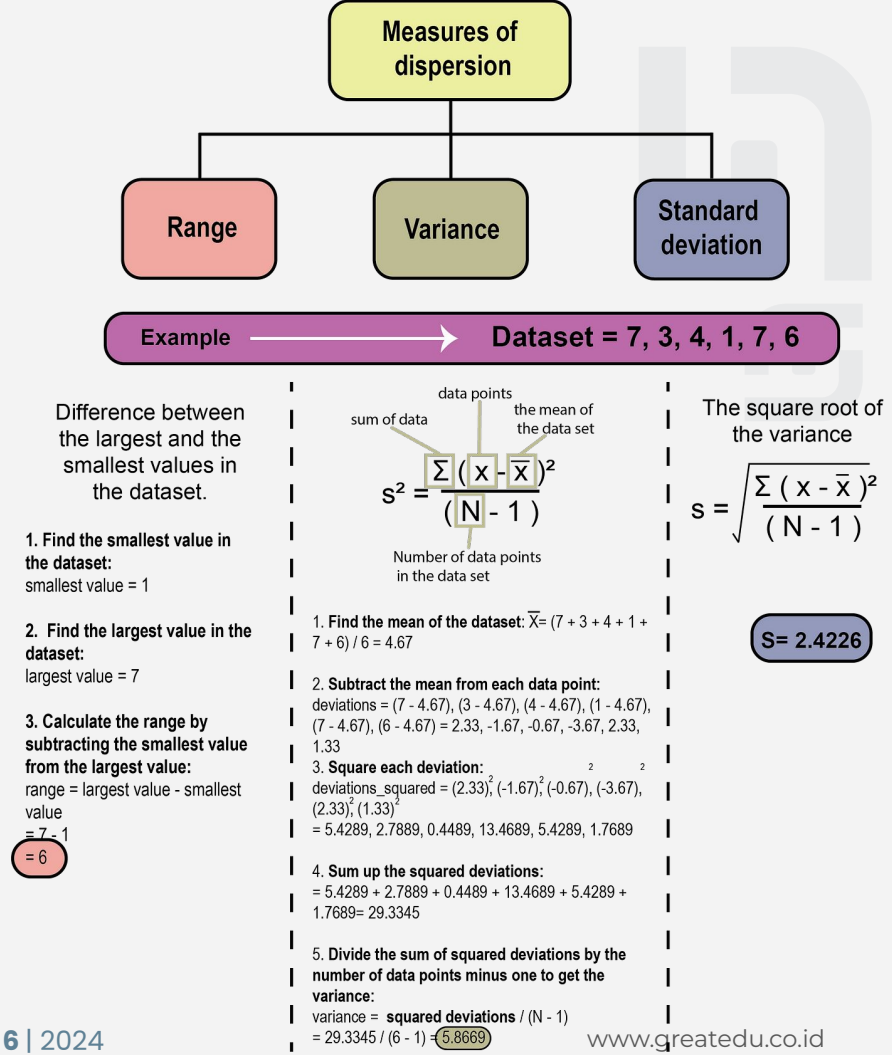
- Modus: Nilai yang paling sering muncul dalam dataset.
- Bisa lebih dari satu modus dalam dataset.
- Dapat digunakan untuk data categorical

```
#Mode:Computed column-wise:
modeData = dataFrame.mode()
```

```
#Mode:Computed row-wise:
modeData = dataFrame.mode(axis=1)
```

Measures of Dispersion

- Ukuran variabilitas (spread) menggambarkan seberapa jauh sebuah kumpulan data menyebar.
- Menggunakan: Range, Variance, Standard Deviation, dan Quartiles



Measures of Dispersion : Quartiles / Kuartil

- Kuartil: Memecah data menjadi empat bagian yang sama. Kuartil pertama (Q1) adalah median dari separuh pertama data, sedangkan kuartil ketiga (Q3) adalah median dari separuh kedua data.
- **Deviasi Kuartil (Quartile Deviation):**

$$\text{Deviasi Kuartil} = \frac{(Q3-Q1)}{2}$$

Interpretasi: Nilai yang lebih kecil menunjukkan bahwa data lebih terkonsentrasi di sekitar median, sedangkan nilai yang lebih besar menunjukkan penyebaran data yang lebih luas.

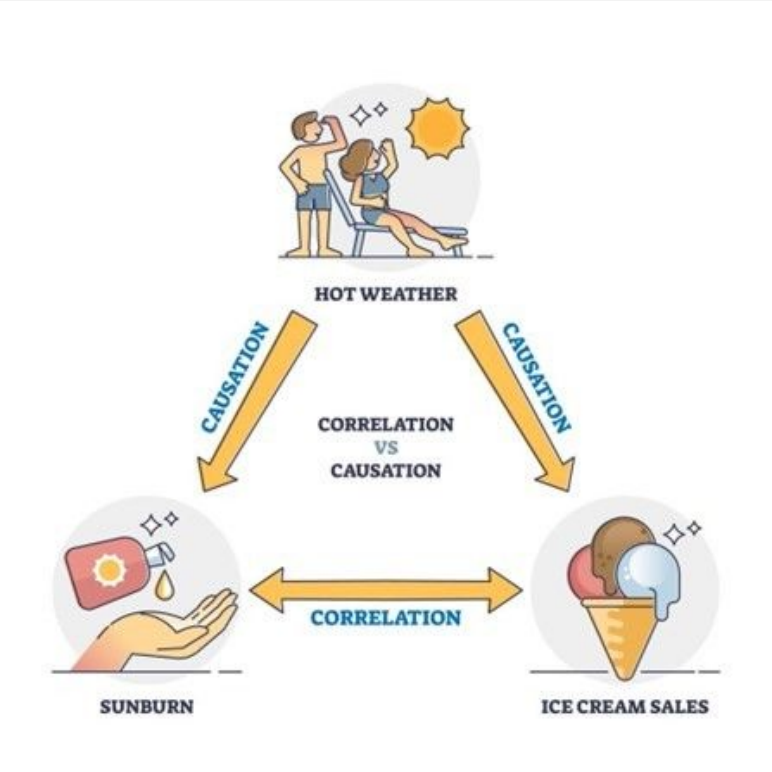
- **Rentang Kuartil (Interquartile Range, IQR):**

$$\text{IQR} = Q3 - Q1$$

Interpretasi: IQR memberikan gambaran tentang variabilitas dalam dataset dengan mengabaikan outlier. Nilai IQR yang rendah menunjukkan homogenitas yang lebih besar di antara nilai-nilai, sedangkan nilai yang tinggi menunjukkan variabilitas yang lebih besar.

Correlation & Causation

CORRELATION & CAUSATION



CORRELATION

- Korelasi mengacu pada hubungan statistik antara dua variabel atau lebih di mana perubahan pada satu variabel terkait dengan perubahan pada variabel lain.
- Korelasi dapat positif, negatif, atau nol:
  - Positif: Jika satu variabel meningkat, variabel lain juga meningkat.
  - Negatif: Jika satu variabel meningkat, variabel lain menurun.
  - Nol: Tidak ada hubungan linier antara dua variabel.
-

CORRELATION

- Korelasi diukur dengan koefisien korelasi:
  - Pearson Correlation Coefficient : data interval dan rasio
  - Spearman's Rank Correlation Coefficient : data ordinal
- Nilai koefisien berkisar dari -1 hingga 1, di mana 1 menunjukkan korelasi positif sempurna, -1 menunjukkan korelasi negatif sempurna, dan 0 menunjukkan tidak ada korelasi.
- Scatter plot adalah alat visualisasi yang sering digunakan untuk memperlihatkan hubungan korelasi antara dua variabel.

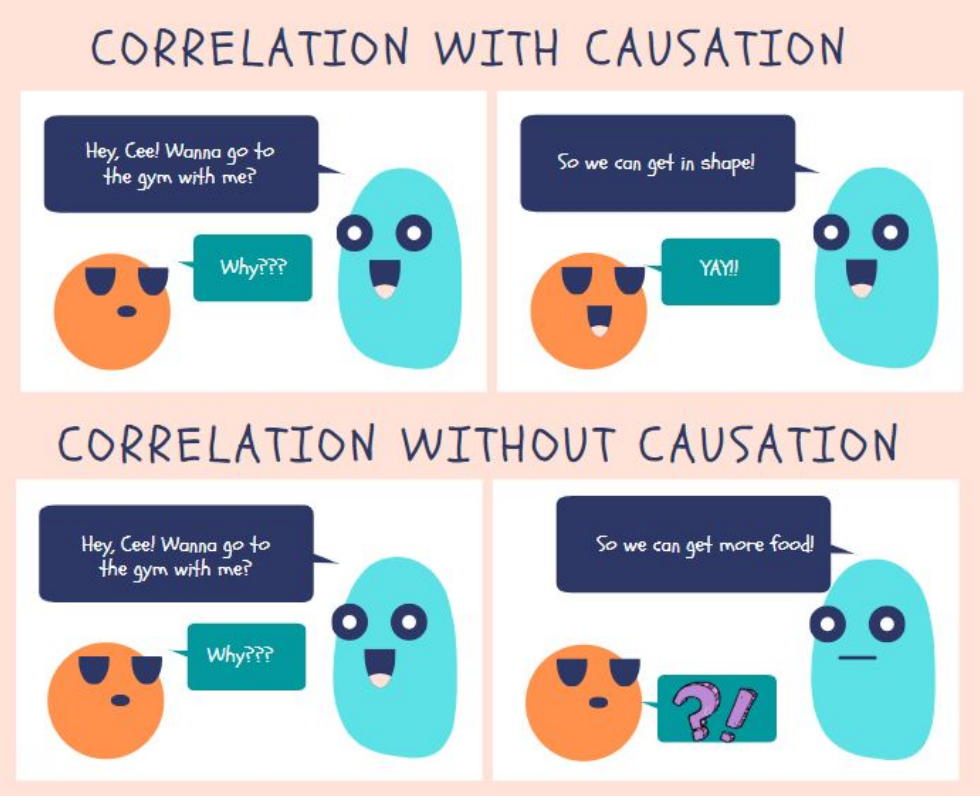
CAUSATION

- Kausalitas menunjukkan hubungan sebab-akibat antara dua variabel atau lebih, di mana perubahan pada satu variabel menyebabkan perubahan pada variabel lain.
- Menentukan kausalitas lebih kompleks dibandingkan mengidentifikasi korelasi. **Eksperimen yang dikendalikan**, di mana peneliti memanipulasi satu variabel (variabel independen) dan mengamati efek pada variabel lain (variabel dependen), sering digunakan untuk menetapkan hubungan kausal.

CAUSATION

- Permasalahan dalam Menentukan Kausalitas:
  - **Variabel Pencemar (Confounding Variables)**: Variabel yang mempengaruhi variabel dependen tetapi tidak diperhitungkan dalam analisis, yang bisa menyebabkan kesimpulan yang salah mengenai hubungan sebab-akibat.
  - **Kesalahan Atribusi Kausal**: Kesalahan dalam menyimpulkan bahwa hubungan sebab-akibat ada hanya berdasarkan korelasi.

CORRELATION VS CAUSATION

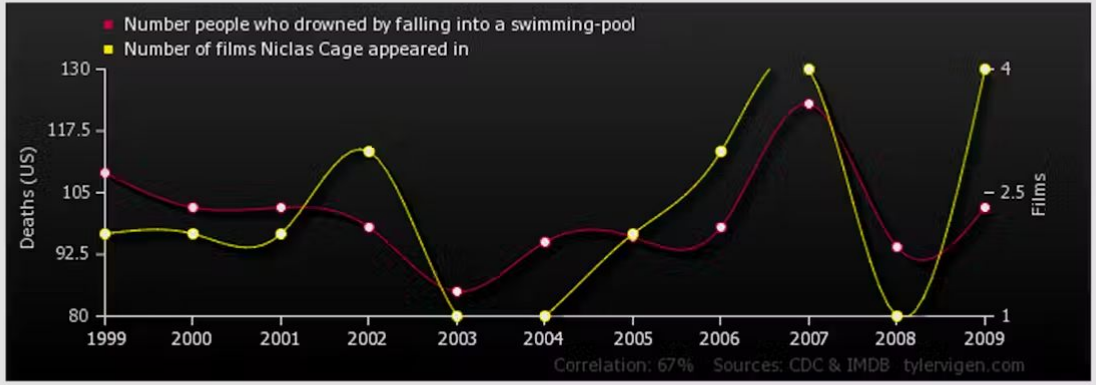


CORRELATION VS CAUSATION

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool *Deaths (US) (CDC)*	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in *Films (IMDB)*	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

Probability & Distribution

Probability / Peluang

- Peluang dalam statistik adalah konsep fundamental yang mengukur seberapa besar kemungkinan terjadinya suatu peristiwa.
- Peluang dapat berkisar dari 0 hingga 1, di mana 0 berarti peristiwa tersebut tidak mungkin terjadi, dan 1 berarti peristiwa tersebut pasti akan terjadi.
- Peluang membantu dalam membuat keputusan berdasarkan ketidakpastian dan digunakan secara luas dalam berbagai bidang seperti matematika, statistik, keuangan, permainan, dan ilmu pengetahuan.

Probability / Peluang

- Peluang dalam statistik adalah konsep fundamental yang mengukur seberapa besar kemungkinan terjadinya suatu peristiwa.
- Peluang dapat berkisar dari 0 hingga 1, di mana 0 berarti peristiwa tersebut tidak mungkin terjadi, dan 1 berarti peristiwa tersebut pasti akan terjadi.
- Peluang membantu dalam membuat keputusan berdasarkan ketidakpastian dan digunakan secara luas dalam berbagai bidang seperti matematika, statistik, keuangan, permainan, dan ilmu pengetahuan.

Konsep Dasar Peluang

- **Ruang Sampel (Sample Space, S)**: Kumpulan semua hasil yang mungkin dari suatu percobaan. Misalnya, dalam pelemparan sebuah koin, ruang sampel adalah {Kepala, Ekor}.
- **Peristiwa (Event, E):** Kumpulan hasil tertentu dari ruang sampel. Sebuah peristiwa bisa terdiri dari satu atau lebih hasil. Misalnya, mendapatkan angka genap pada pelemparan dadu adalah peristiwa yang mencakup hasil {2, 4, 6}.
- **Peluang Suatu Peristiwa (P(E)):** Mengukur seberapa sering kita mengharapkan suatu peristiwa terjadi jika kita mengulangi percobaan dalam banyak kali. Diukur dengan rumus:

$$P(E) = \frac{\text{Jumlah kasus yang menguntungkan}}{\text{Total kasus yang mungkin}}$$

Aturan Peluang

- **Aturan Penjumlahan**: Jika dua peristiwa, A dan B, adalah saling eksklusif (tidak bisa terjadi bersamaan), maka peluang terjadinya A atau B adalah jumlah peluang masing-masing peristiwa.

$$P(A \cup B) = P(A) + P(B)$$

- **Aturan Perkalian**: Jika dua peristiwa, A dan B, adalah independen, maka peluang terjadinya A dan B bersamaan adalah produk dari peluang masing-masing peristiwa.

$$P(A \cap B) = P(A) \times P(B)$$

Contoh Aturan Penjumlahan

- Misalkan Anda memiliki dek kartu standar berisi 52 kartu (13 dari setiap jenis: hati, keriting, wajik, dan sekop). Anda ingin mengetahui peluang mengambil kartu yang berjenis hati atau kartu raja.

1. **Menghitung Peluang Mengambil Kartu Hati**:

   Ada 13 kartu hati dalam dek.

   $$P(\text{Hati}) = \frac{13}{52}$$
2. **Menghitung Peluang Mengambil Kartu Raja**:

   Ada 4 kartu raja dalam dek (satu dari setiap jenis).

   $$P(\text{Raja}) = \frac{4}{52}$$
3. **Menghitung Peluang Mengambil Kartu Hati atau Raja**:

   Ada 1 kartu raja yang juga berjenis hati, yang harus kita kurangi dari total karena sudah kita hitung dua kali.

   $$P(\text{Hati} \cup \text{Raja}) = P(\text{Hati}) + P(\text{Raja}) - P(\text{Hati} \cap \text{Raja})$$
   $$P(\text{Hati} \cup \text{Raja}) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

Contoh Aturan Perkalian

- Bayangkan Anda melempar dua dadu yang adil secara bersamaan. Anda ingin mengetahui peluang kedua dadu menunjukkan angka 6.

1. **Menghitung Peluang Dadu Pertama Menunjukkan Angka 6**:
   Setiap dadu memiliki 6 sisi, jadi peluang untuk mendapatkan angka 6 pada satu lemparan adalah:
   $$P(6 \text{ pada dadu } 1) = \frac{1}{6}$$
2. **Menghitung Peluang Dadu Kedua Menunjukkan Angka 6**:
   Sama seperti dadu pertama, peluangnya adalah:
   $$P(6 \text{ pada dadu } 2) = \frac{1}{6}$$
3. **Menghitung Peluang Kedua Dadu Menunjukkan Angka 6**:
   Karena peristiwa ini independen (hasil pada dadu pertama tidak mempengaruhi hasil pada dadu kedua), kita dapat menggunakan aturan perkalian.
   $$P(6 \text{ pada kedua dadu}) = P(6 \text{ pada dadu } 1) \times P(6 \text{ pada dadu } 2)$$
   $$P(6 \text{ pada kedua dadu}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Peluang Kondisional

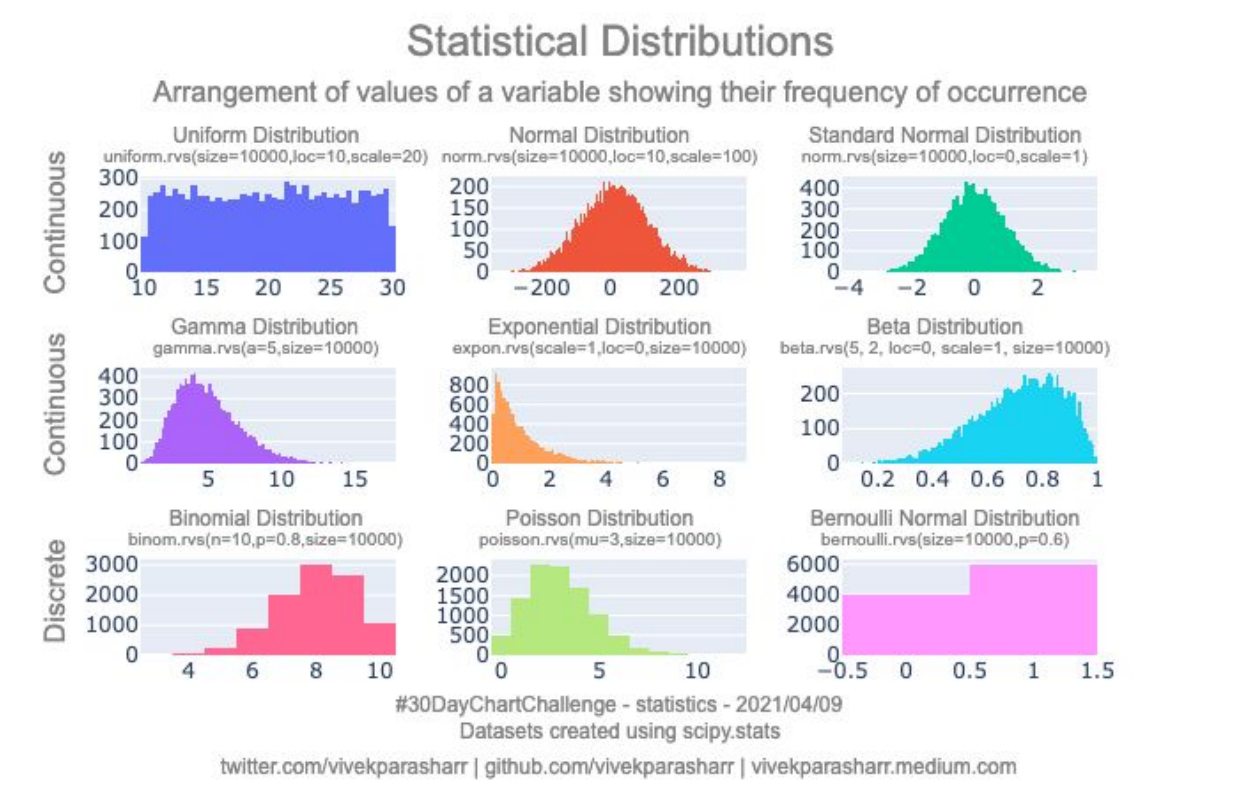
- Peluang kondisional mengukur peluang suatu peristiwa terjadi dengan asumsi bahwa peristiwa lain telah terjadi. Ini dinyatakan sebagai P(A|B), yang dibaca "peluang A, dengan asumsi B terjadi"

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Distribusi

- Distribusi dalam statistik merupakan konsep kunci yang menjelaskan frekuensi kemunculan nilai dalam kumpulan data.
- Konsep ini memungkinkan kita untuk memahami pola data, termasuk bagaimana data cenderung menyebar atau berkumpul, nilai mana yang sering muncul, dan kemungkinan outlier.

Distribusi


Statistical Distributions
Arrangement of values of a variable showing their frequency of occurrence
Continuous
Uniform Distribution
uniform.rvs(size=10000,loc=10,scale=20)
Normal Distribution
norm.rvs(size=10000,loc=10,scale=100)
Standard Normal Distribution
norm.rvs(size=10000,loc=0,scale=1)
Continuous
Gamma Distribution
gamma.rvs(a=5,size=10000)
Exponential Distribution
expon.rvs(scale=1,loc=0,size=10000)
Beta Distribution
beta.rvs(5, 2, loc=0, scale=1, size=10000)
Discrete
Binomial Distribution
binom.rvs(n=10,p=0.8,size=10000)
Poisson Distribution
poisson.rvs(mu=3,size=10000)
Bernoulli Normal Distribution
bernoulli.rvs(size=10000,p=0.6)
#30DayChartChallenge - statistics - 2021/04/09
Datasets created using scipy.stats
twitter.com/vivekparasharr | github.com/vivekparasharr | vivekparasharr.medium.com

Distribusi Bernoulli

- Distribusi Bernoulli adalah distribusi probabilitas diskrit yang sangat dasar dalam teori probabilitas dan statistik.
- Distribusi ini menggambarkan eksperimen di mana hanya ada dua kemungkinan hasil: sukses (biasanya dinotasikan dengan 1) dan gagal (dinotasikan dengan 0).
- Probabilitas keberhasilan (p) tetap sama setiap kali eksperimen dilakukan, dan probabilitas kegagalan adalah 1-p.

$$P(X = x) = p^x(1-p)^{1-x}$$

di mana:

- X adalah variabel acak yang merepresentasikan hasil (0 atau 1),
- p adalah probabilitas keberhasilan,
- $1 - p$ adalah probabilitas kegagalan,
- x adalah hasil yang diamati (0 atau 1).

Distribusi Bernoulli - Contoh kasus

Sebuah dadu dilempar 1 kali. Jika diasumsikan sukses ketika muncul mata dadu yang habis dibagi 3. Tentukan peluang sukses dari hasil lemparan sebuah dadu.

- **T = {1,2,3,4,5,6}**
- **Misal A adalah kejadian munculnya mata dadu yang habis dibagi 3.**
- **A = {3,6}**
- **P(A) = 2/6 = 0.333**

Distribusi Binomial

- **Karakteristik**: Distribusi diskrit yang menggambarkan jumlah keberhasilan dalam n percobaan independen, dengan setiap percobaan memiliki dua kemungkinan hasil (sukses atau gagal) dan probabilitas keberhasilan yang sama pada setiap percobaan.
- **Pentingnya**: Cocok untuk memodelkan peristiwa dengan hasil Ya atau Tidak, seperti lemparan koin atau keberhasilan produk.
- **Contoh**: Jumlah kepala dalam 10 lemparan koin atau jumlah produk cacat dalam sampel.

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

Distribusi Binomial - Contoh kasus

Jika sebuah koin dilempar 5 kali, gunakan distribusi binomial untuk menemukan peluang untuk mendapatkan tepat 2 kepala

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Dimana $n = 5, k = 2,$ dan $p = 0.5.$

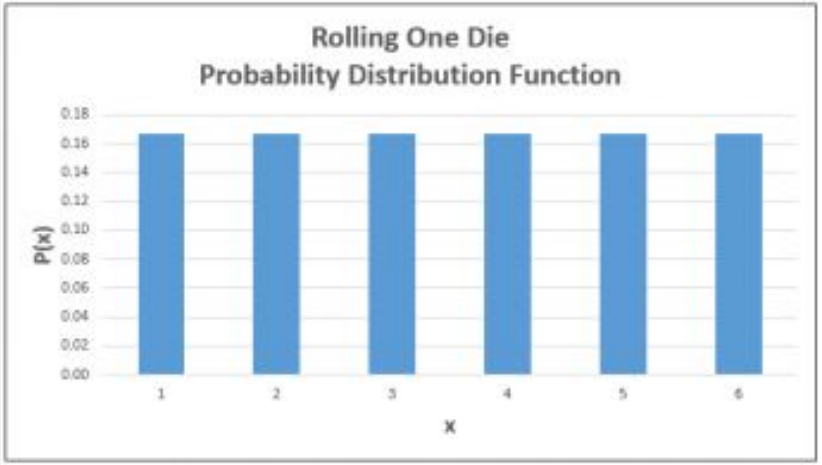
$$P(X = 2) = \binom{5}{2} (0.5)^2 (0.5)^{5-2} = \frac{5!}{2!(5-2)!} \cdot 0.5^2 \cdot 0.5^3$$

$$P(X = 2) = 10 \cdot 0.5^2 \cdot 0.5^3 = 10 \cdot 0.25 \cdot 0.125 = 10 \cdot 0.03125 = 0.3125$$

Jadi, peluang mendapatkan tepat 2 kepala adalah 0.3125 atau 31.25%.

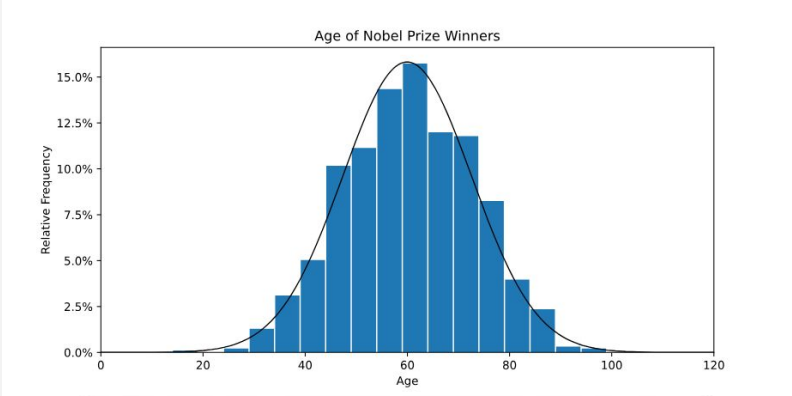
Distribusi Uniform

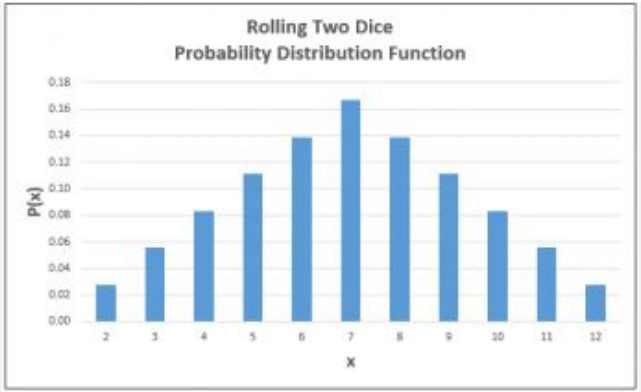
- **Karakteristik**: Semua nilai memiliki peluang yang sama untuk terjadi dalam rentang tertentu. Distribusi ini bisa diskrit atau kontinu.
- **Pentingnya**: Berguna ketika tidak ada cukup informasi untuk mengasumsikan distribusi lain, atau untuk simulasi acak.
- **Contoh**: Pemilihan nomor acak antara 1 dan 10, di mana setiap nomor memiliki peluang yang sama untuk dipilih.



Distribusi Normal (Gaussian)

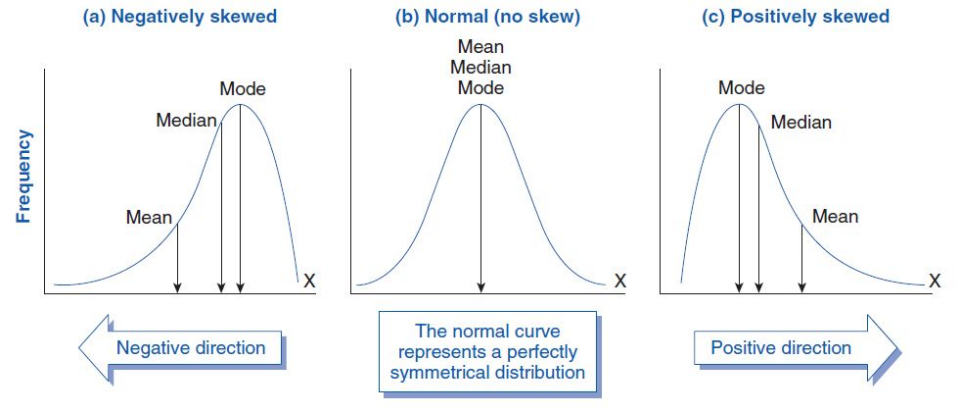
- **Karakteristik**: Distribusi simetris di mana **nilai terkonsentrasi di sekitar mean** dan menyebar secara seragam ke kedua arah (kiri dan kanan). Kurva distribusi ini berbentuk **lonceng**.
- **Pentingnya**: Distribusi normal adalah dasar dari banyak prosedur statistik karena banyak fenomena alam dan proses sosial yang cenderung mengikuti pola distribusi ini.
- **Properti**: Mean, median, dan modus dari distribusi normal adalah sama. Standar deviasi menentukan lebar kurva.
- **Contoh**: Tinggi badan orang dewasa, tekanan darah, dan skor tes standar.

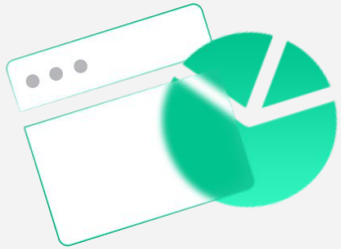




Distribusi Miring (Skewed Distribution)

- **Karakteristik**: Distribusi yang tidak simetris di mana mayoritas data berkumpul di satu sisi mean, menyebabkan ekor yang panjang di sisi lainnya. Distribusi miring dapat dibagi menjadi dua jenis: miring ke kanan (**positif**) dan miring ke kiri (**negatif**), contoh:
    - **Penghasilan**: Pendapatan dalam populasi sering kali memiliki distribusi miring ke kanan, di mana sebagian besar orang memiliki pendapatan di bawah rata-rata, tetapi ada beberapa orang dengan pendapatan sangat tinggi yang menarik rata-rata ke arah mereka.
    - **Usia Kematian**: Dalam beberapa populasi, distribusi usia kematian mungkin miring ke kiri, di mana sebagian besar kematian terjadi pada usia tua, tetapi ada ekor panjang ke arah usia muda karena penyakit atau kecelakaan.





Terima Kasih