LEARNING PROGRESS REVIEW

OMICRON

DATA SCIENCE BOOTCAMP

WEEK 8

DigitalSkola

Omicron

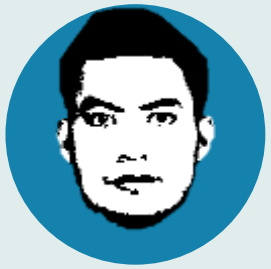
Team members

ANUGRAH YAZID GHANI

https://www.linkedin.com/in/anugrah-yazid-7253bb221/

EDO MOHAMMAD HADAD GIBRAN

https://www.linkedin.com/in/edo-gibran-38505a142 /



FAJAR ACHMAD

https://www.linkedin.com/in/fajar-achmad-755945111 /

MUHAMMAD FIKRI FADILA

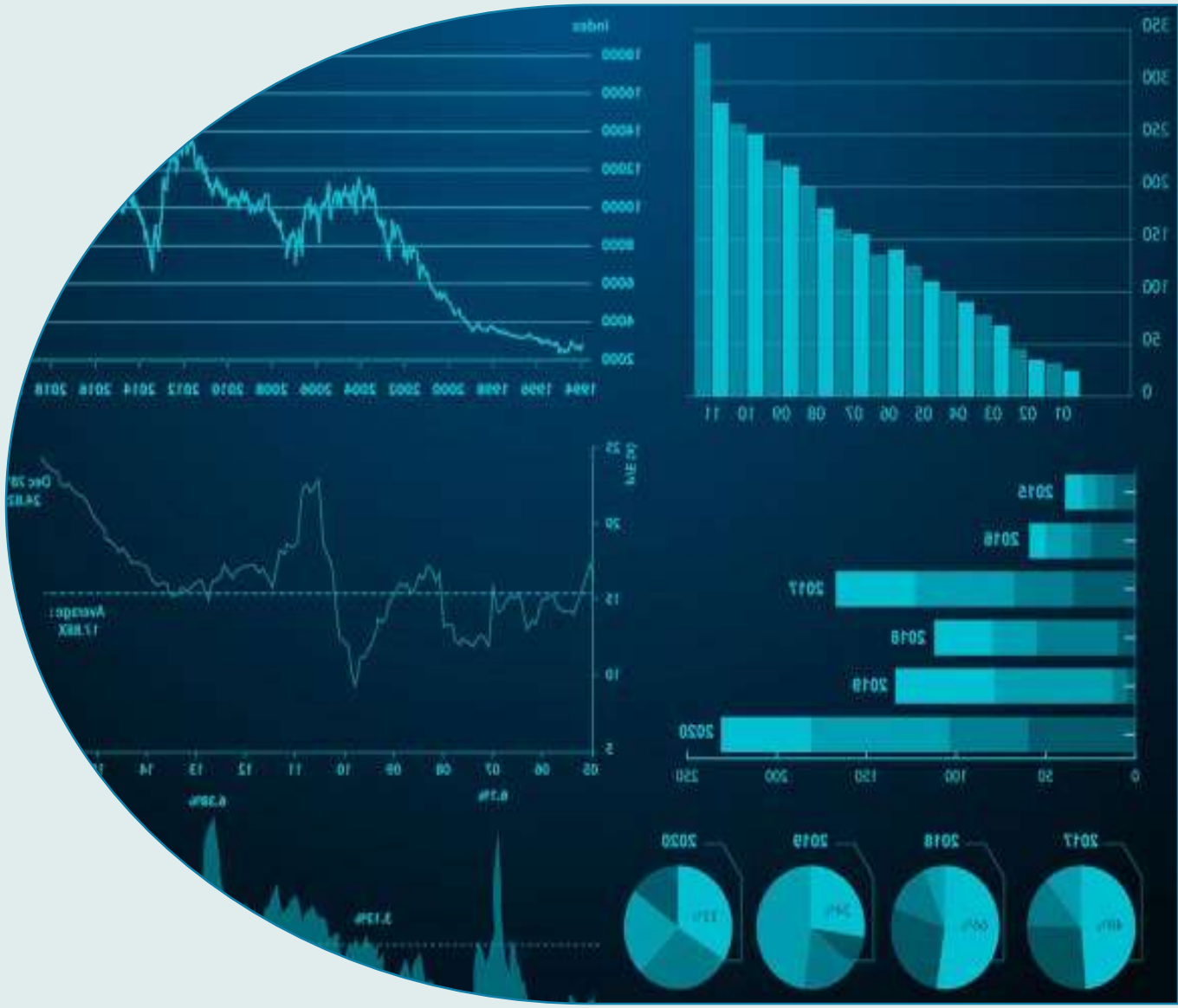
https://www.linkedin.com/in/muhammad-fikri-fadila-a551161a6 /

Table of
Content

DigitalSkola

01

BASIC STATISTICS



Mengapa Statistik Penting?

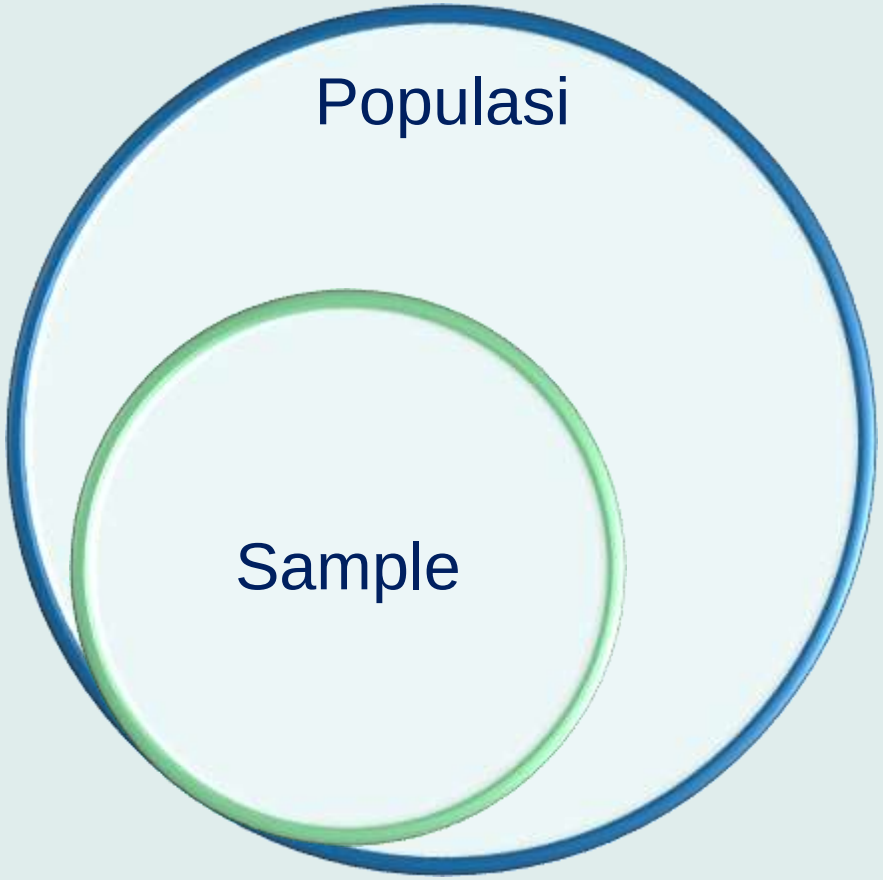
- **Data is everywhere**
- Statistik membantu kita dalam menginterpretasi data dalam jumlah besar dan membuat rangkuman informasi yang bermakna.
- Di era informasi, data dibutuhkan untuk melakukan *data-driven decision making*, dan pengambilan keputusan ini bisa terjadi dengan sangat cepat dalam dunia bisnis.
- Pemahaman terhadap statistik membantu kita dalam menilai suatu informasi apakah benar atau tidak.

Populasi

Populasi merepresentasikan keseluruhan elemen yang menjadi objek obsrervasi.

Sampel

Sample adalah bagian dari populasi



Parameter

Statistics untuk keseluruhan populasi.

Statistik

- Pengukuran *properties* dari *sample*. (Contoh : Modus, Median, *Mean*, Standar Deviasi, dan *Variance*).
- Angka *statistics* digunakan untuk memperkirakan nilai dari suatu parameter yang terkadang tidak diketahui nilainya.
- Studi dalam *statistics* berkaitan erat dengan studi mengenai *sample* dari populasi.

Tipe Data

CATEGORICAL

- **Nominal** → Tipe data yg direpresentasikan sebagai kumpulan *event* atau objek dalam sebuah kategori yang bersifat diskrit. Contoh : Kumpulan merk mobil
- **Ordinal** → Tipe data yang mirip dengan nominal namun dapat diurutkan. Contoh: Tingkat kepuasan pelanggan: Tidak Puas, Puas, Sangat Puas.

NUMERIK

- **Interval** → Tipe data numerik yg tidak memiliki nilai *true zero* (nilai bisa dibawah nol/negative). Contoh: Suhu Udara.
- **Ratio** → Tipe data yang mirip dengan interval namun memiliki angka *true zero.* Contoh: Usia, Pendapatan.

Domain Statistics

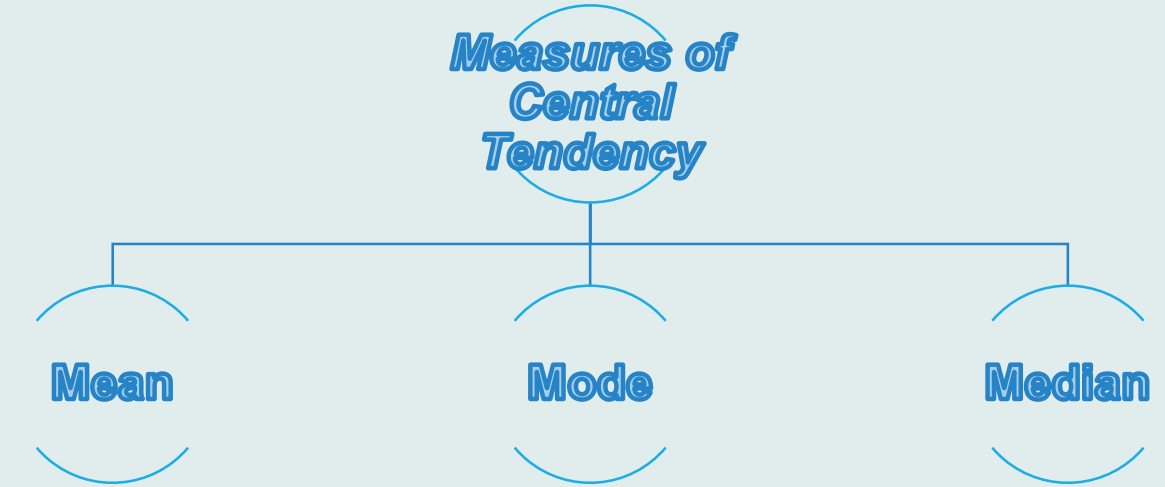
Descriptive Statistics

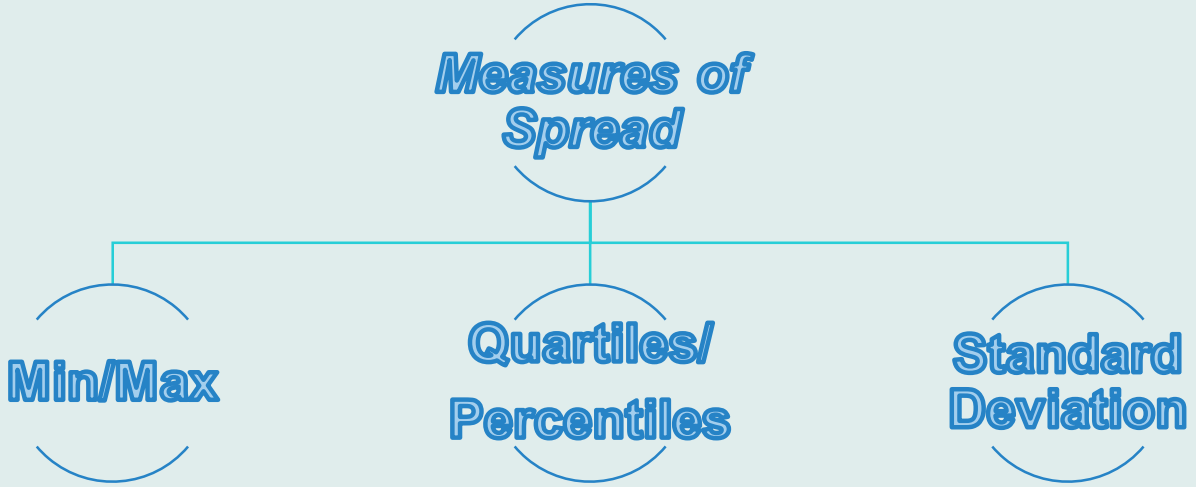
Metode statistika yang digunakan untuk mendeskripsikan dan menampilkan rangkuman data dalam bentuk visual (gambar).

Inferential Statistics

Metode statistika yang digunakan untuk melakukan generalisasi terhadap suatu *sample* pada sebuah populasi.

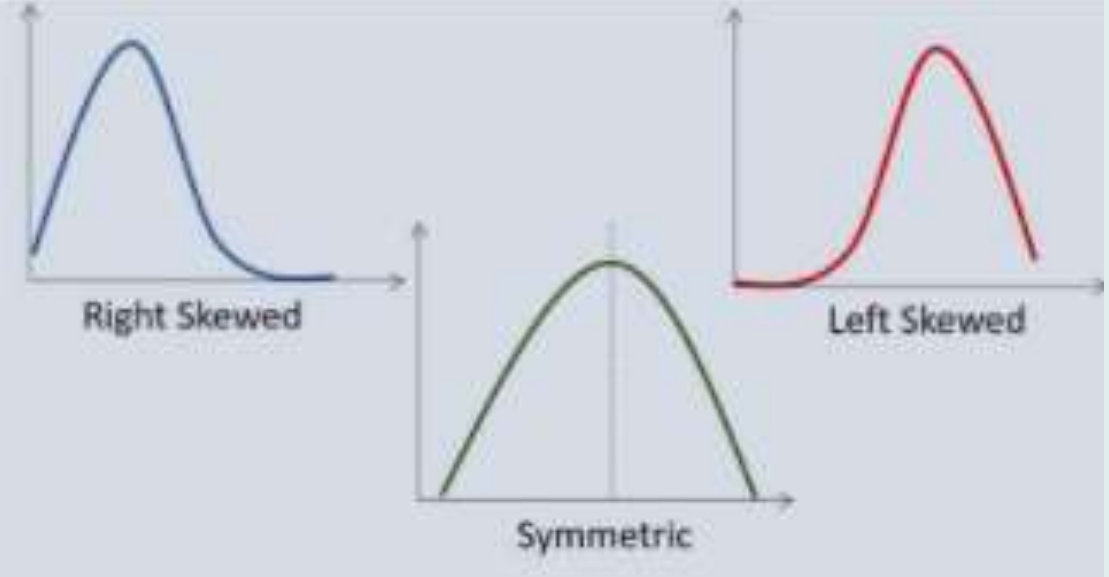
Descriptive Statistics





DigitalSkola

Measures of Central Tendency



Mean

- *Metric* yang paling popular digunakan dalam *descriptive statistics*.

- Rumus :

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

  Dimana :

  $\bar{x}$: *mean*

  x : elemen data *sample*

  n : banyaknya elemen data *sample*

- *Mean* sensitif terhadap data yang *skewed* (cenderung tidak terdistribusi normal).
- Semakin *skewed* maka *mean* bisa kehilangan kemampuan untuk memberikan gambaran nilai tengah suatu data.

Measures of Central Tendency

DigitalSkola

Median

- *Median* dapat membantu untuk menyelesaikan isu representasi data nilai tengah dengan *mean* apabila ada *outlier.*
- Median diukur dengan tahapan :
  1. Mengurutkan elemen numerik dari terkecil ke terbesar.
  2. Menentukan banyaknya elemen (n).
  3. Apabila n ganjil, maka median :

$$Median = \left(\frac{n+1}{2}\right)$$

  4. Apabila n genap, maka median :

$$Median = \frac{\left(\frac{n}{2}\right)+\left(\frac{n}{2}+1\right)}{2}$$

Measures of Central Tendency

Mode

- *Mode* adalah elemen yang memiliki **frekuensi terbanyak** dalam suatu data numerik.

- Apabila *mean* dan median tidak bias digunakan dalam data yang berbentuk *categorical*, lain hal dengan *mode* yang bisa digunakan untuk data *categorical*.

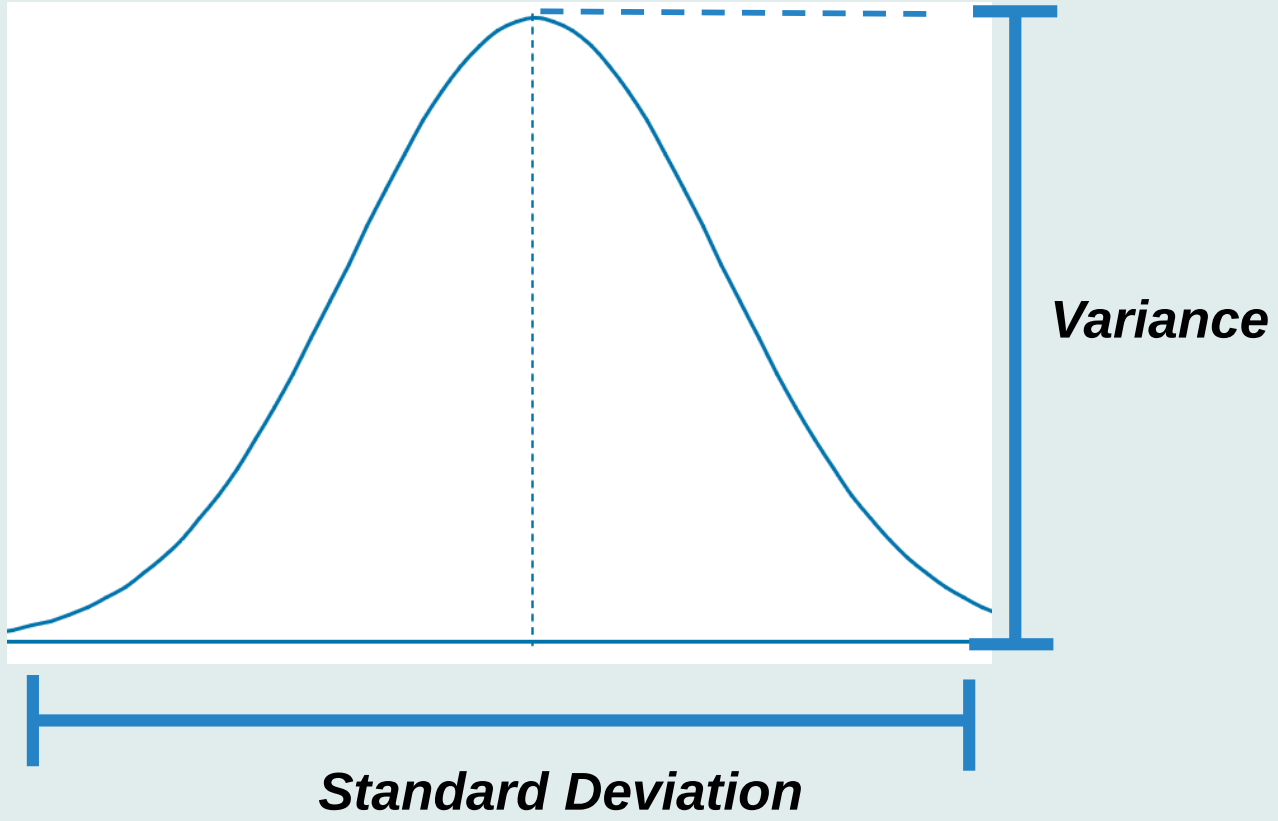
DigitalSkola

Measure of Spread

Quartiles	*Percentiles*	*Deciles*
• Membagi elemen data numerik menjadi 4 bagian sama besar.	• Membagi elemen data numerik menjadi 100 bagian sama besar.	• Membagi elemen data numerik menjadi 10 bagian sama besar.

Measure of Spread

- *Standard deviation/Variance* digunakan untuk mengetahui sebaran dari suatu data.
- *Standard Deviation* merepresentasikan selebar apa distribusi data.
- *Variance* merepresentasikan selandai apa distribusi data.



Rumus:

$$\textbf{variance} = \sigma^2 = \frac{\sum(x_x - \mu)^2}{n}$$

$$\textbf{Standard Deviation} = \sigma = \sqrt{\frac{\sum(x_x - \mu)^2}{n}}$$

Dimana :

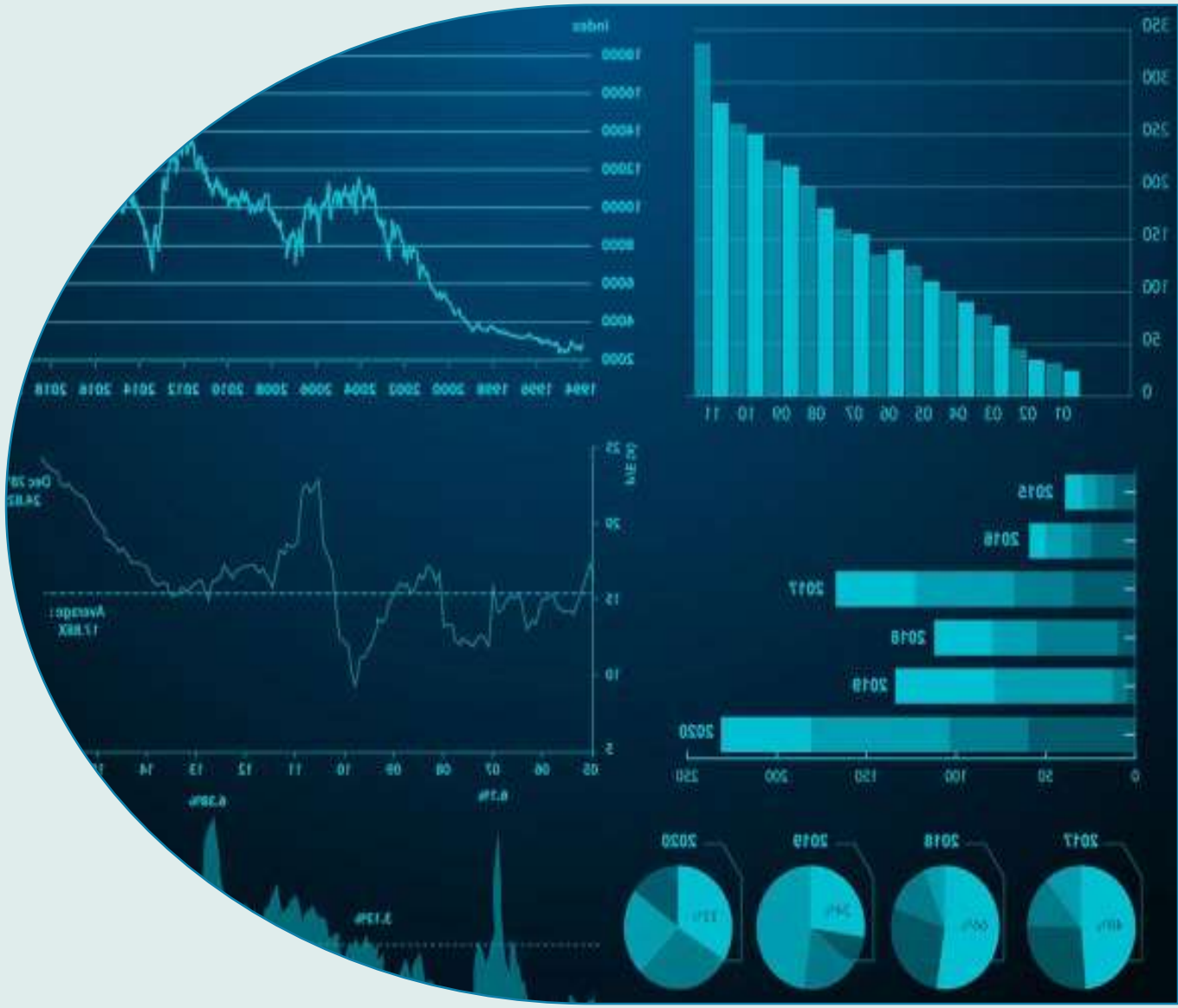
μ : *mean*

n : banyaknya elemen data *sample*

x : elemen data *sample*

02

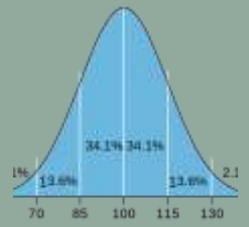
INTERMEDIATE STATISTICS



DigitalSkola

Content

Probabilitas



Distribusi dan Skewness

Korelasi dan sebab akibat

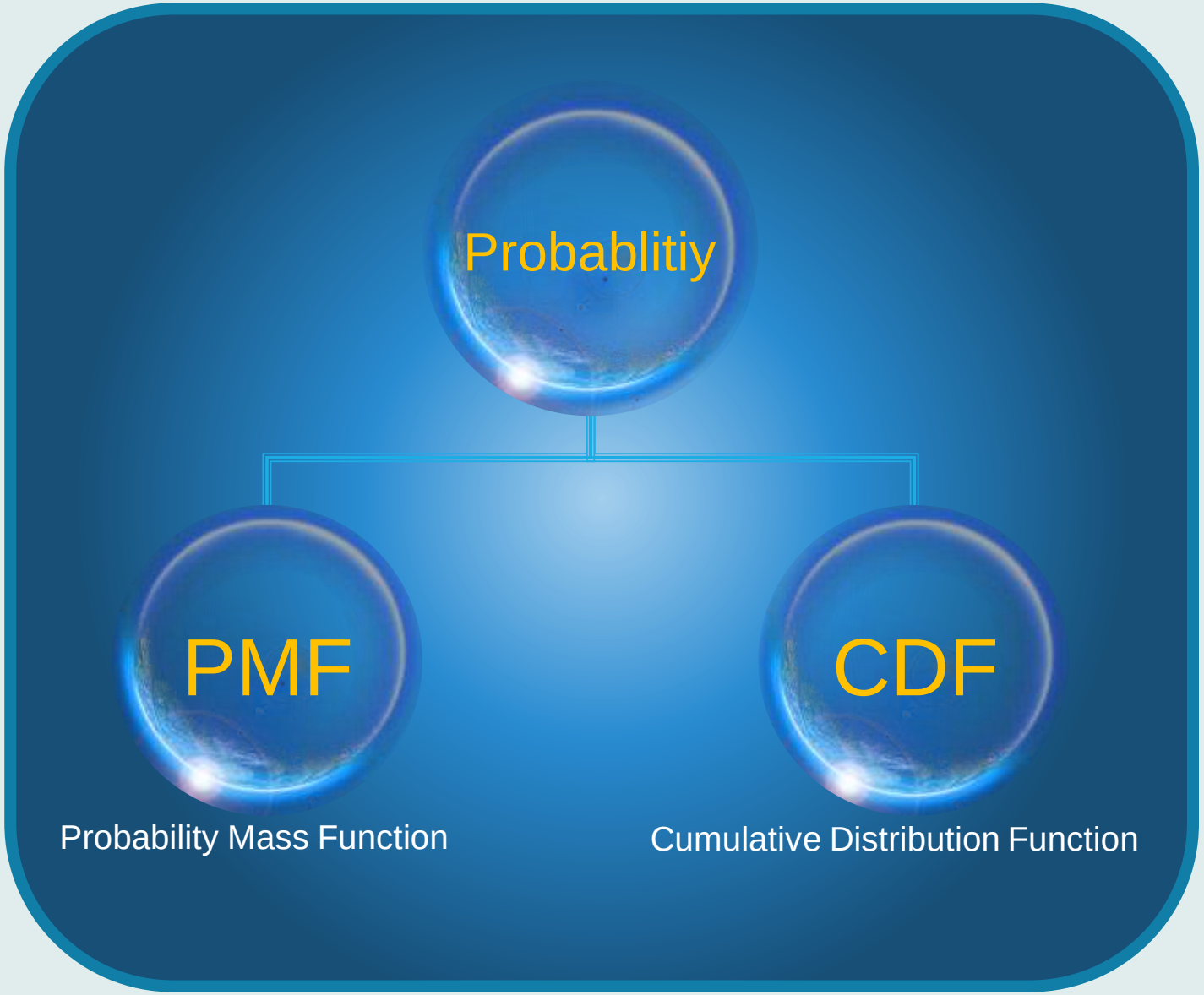
Plot Statistik

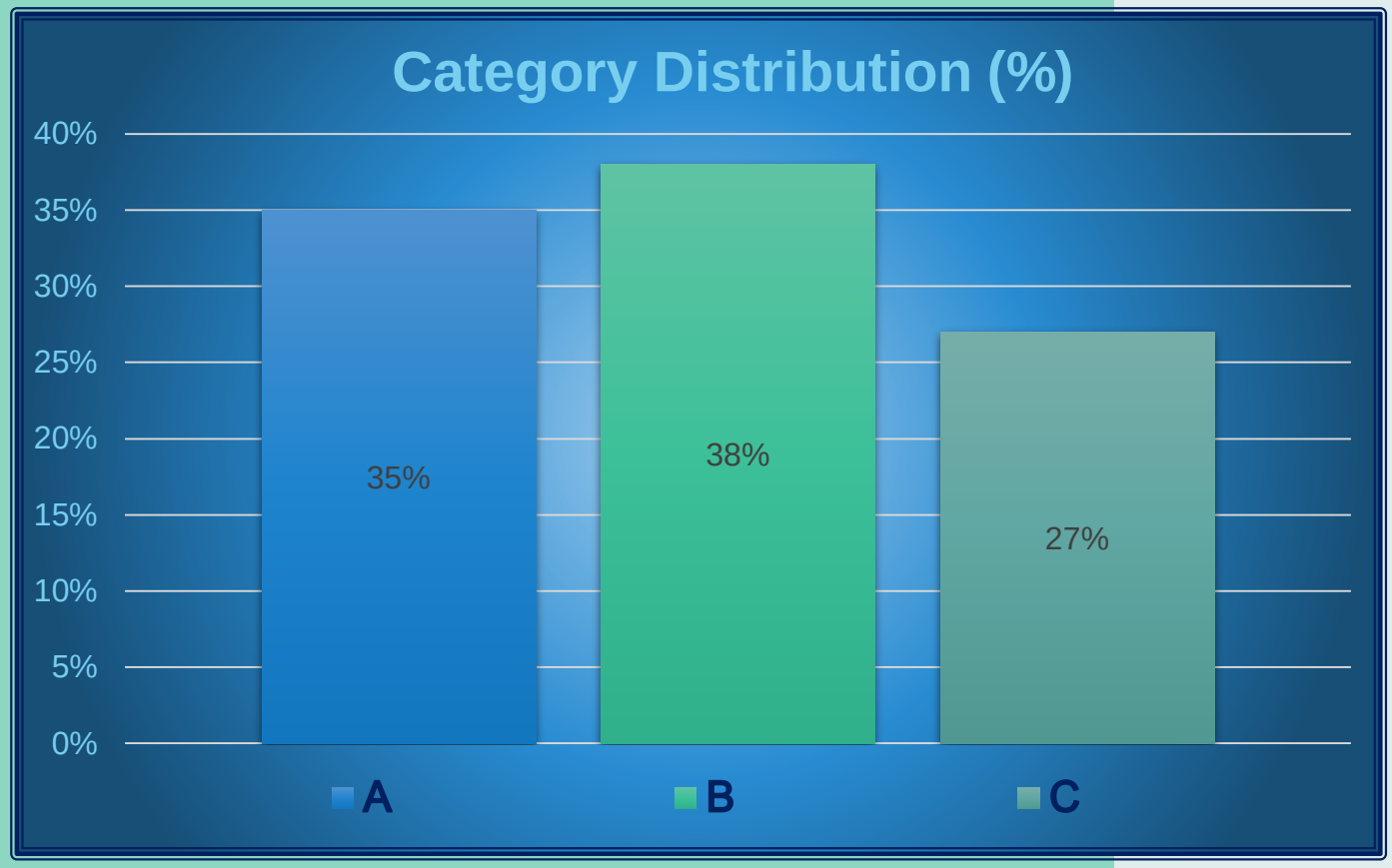
Probabilitas

Mengukur seberapa besar kemungkinan suatu peristiwa terjadi. Skala pengukuran dari 0 – 1, dimana:

Skala 0 : menunjukkan peristiwa tidak pernah terjadi

Skala 1 : menunjukkan peristiwa selalu terjadi



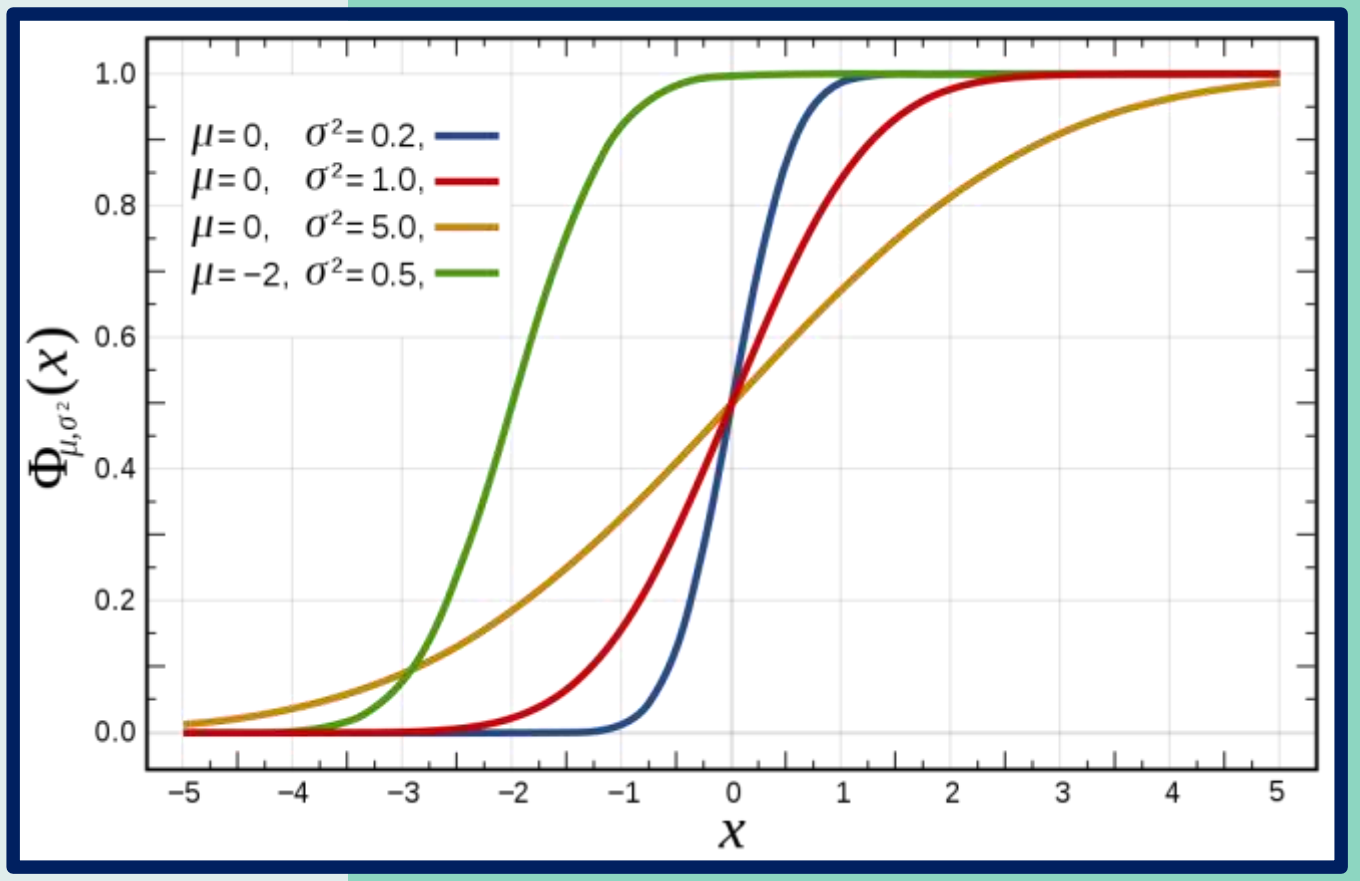


Probability Mass Function

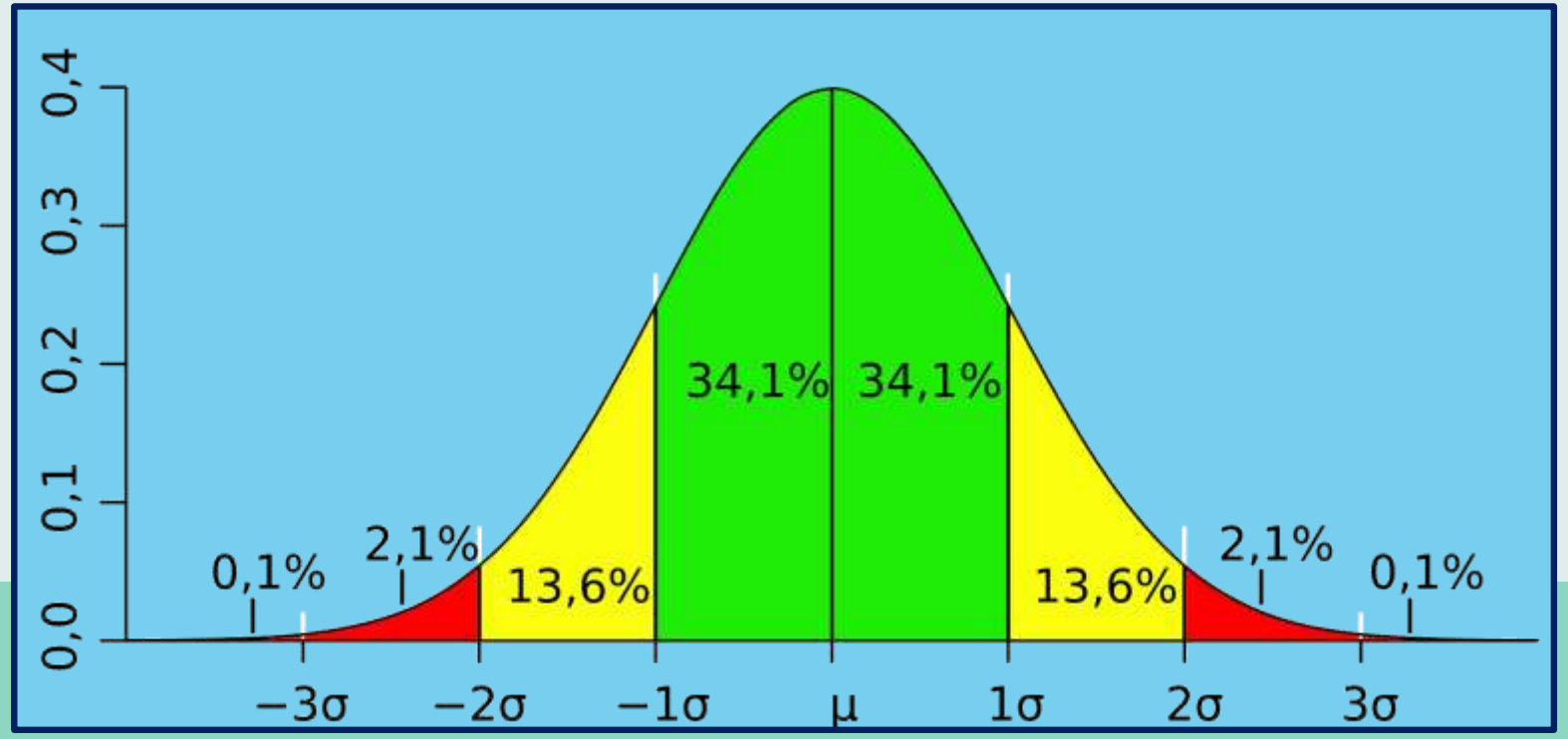
- Pembagian antara sebaran data dengan jumlah total data.
- Frekuensi yang dinyatakan sebagai sebuah pecahan dari suatu sampel.
- Disebut juga normalisasi.

Cumulative Distribution Function

Memiliki fungsi yang sama dengan PMF, namun lebih efektif untuk data dengan lebih banyak varian, dimana akan terjadi peningkatan ***random noise*** pada PMF.



DigitalSkola



Distribusi

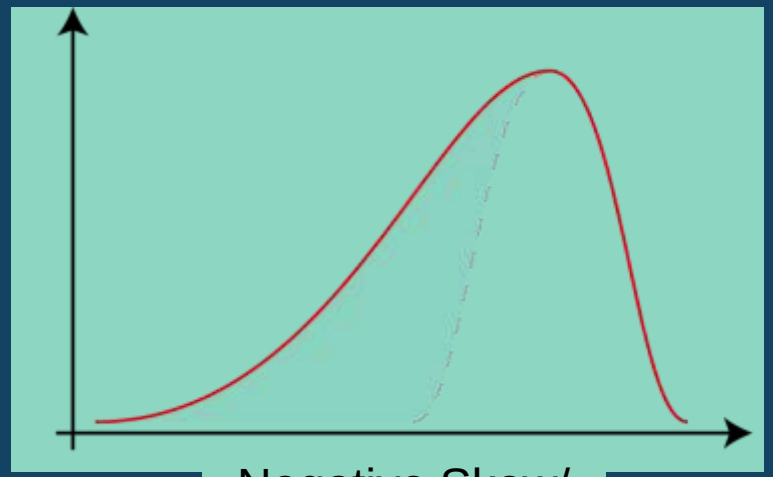
Merupakan distribusi probabilitas kontinu yg dicirikian oleh kurva berbentuk lonceng simetris (*bell-shaped-curved*).

Karakteristik Distribusi Normal

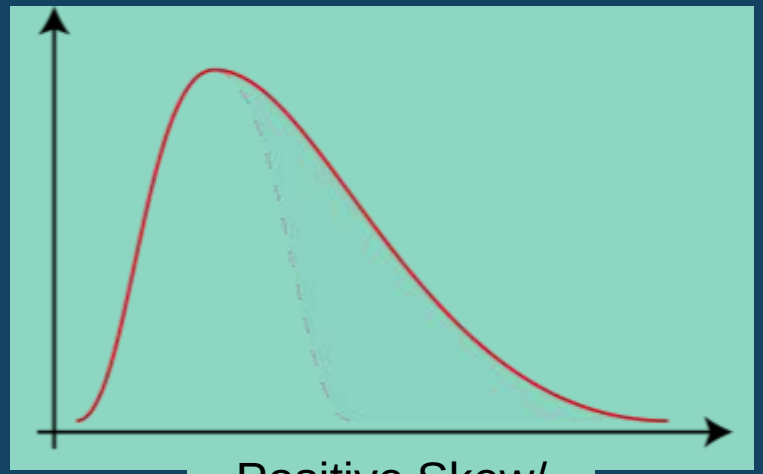
- Simetris jika dibagi 2 dari pusatnya.
- Nilai rata-rata dan median hampir sama.
- Nilai rata-rata sebagai pusatnya dan standar deviasi adalah penyebarannya.

Skewness

Ketika terdapat varian data yg menyimpang dari nilai rata-rata sehingga menyebabkan kurva distribusi tidak simetris.



Negative Skew/
Left Skewness



Positive Skew/
Right Skewness

Negative Skew/ Left Skewness

Ketika nilai mean/rata-rata condong ke kiri dan berada di sisi kiri nilai median.

Positive Skew/ Right Skewness

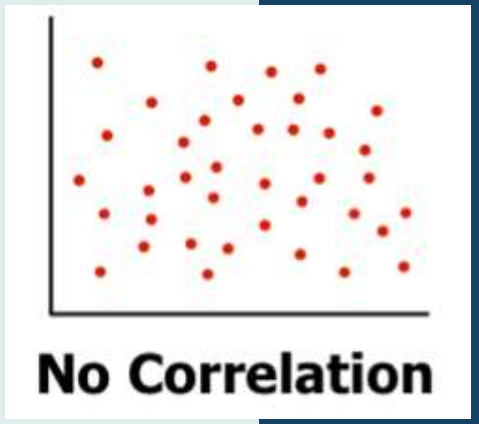
Ketika nilai mean/rata-rata condong ke kanan dan berada di sisi kanan nilai median.

DigitalSkola

Korelasi

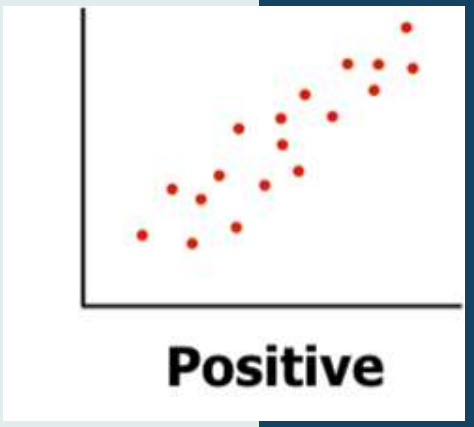
Hubungan statistic antara dua variable secara positif ataupun negative, baik memiliki hubungan yg lemah ataupun kuat yg ditunjukkan dari nilai korelasinya, yaitu **semakin mendekati angka 1 nilai korelasi semakin kuat**.

DigitalSkola



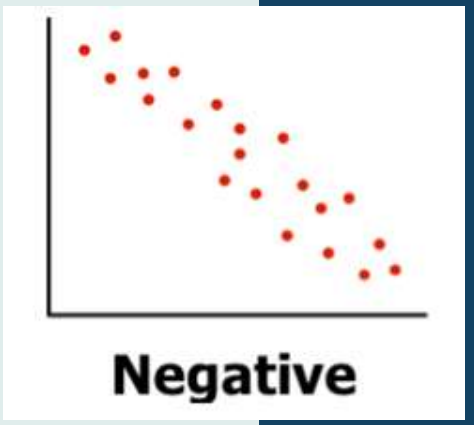
Neutral Correlation

Tidak ada hubungan diantara kedua variable.



Positive Correlation

Kedua variable menunjukan angka yg saling mendukung ke arah yg sama.

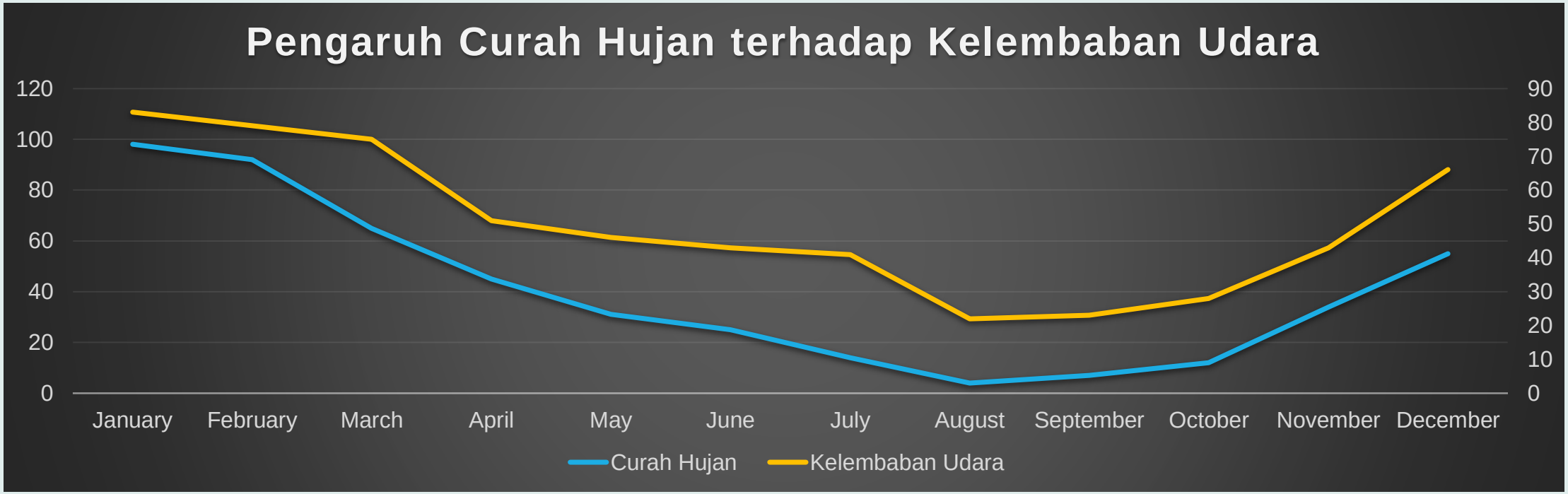


Negative Correlation

Nilai kedua variable menunjukan hubungan yg bertolak belakang.

Sebab Akibat

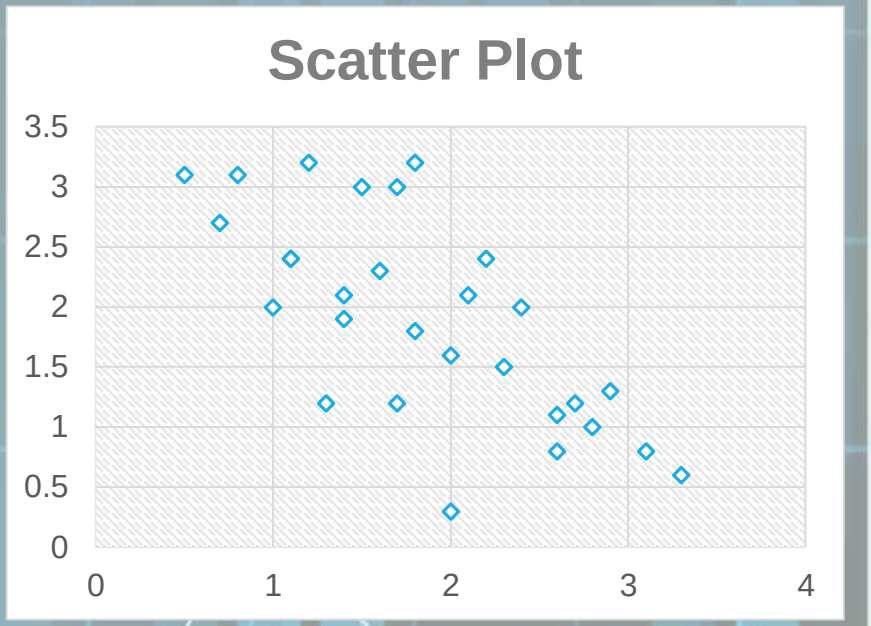
Sebuah peristiwa atau proses yg berkontribusi pada kejadian yg lain.

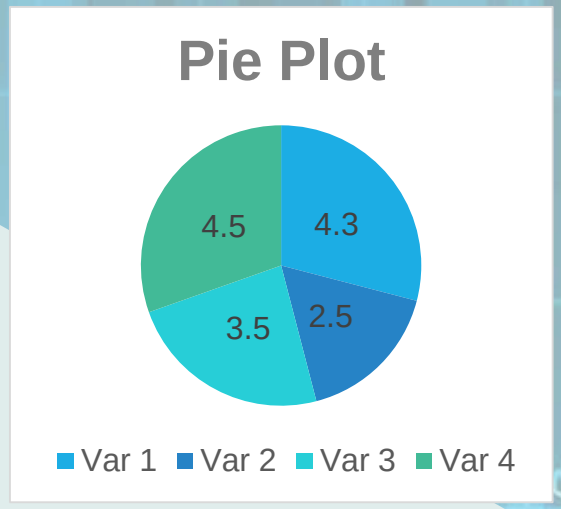


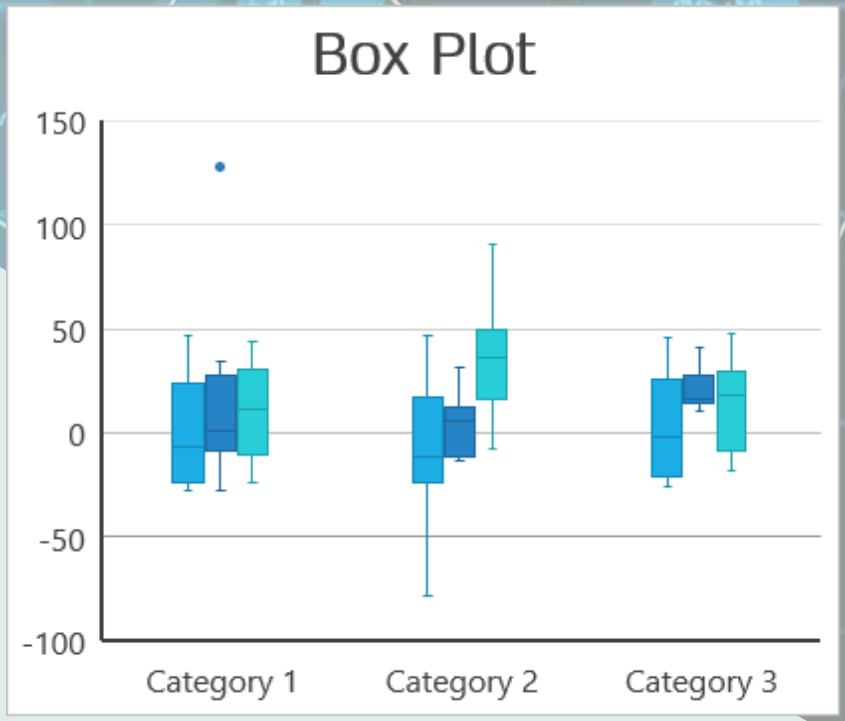
DigitalSkola

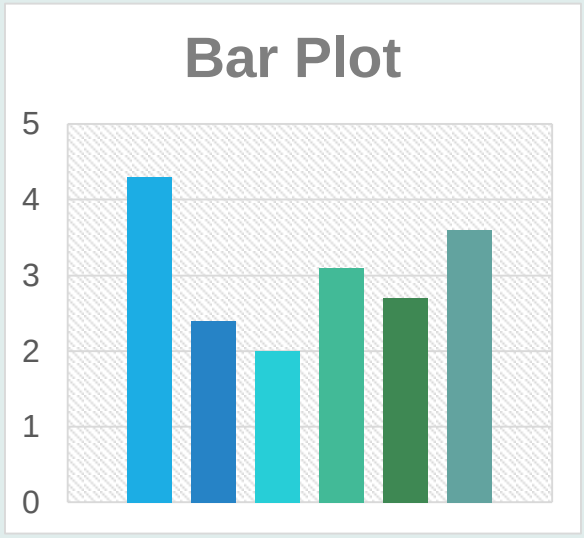
Statistical Plot

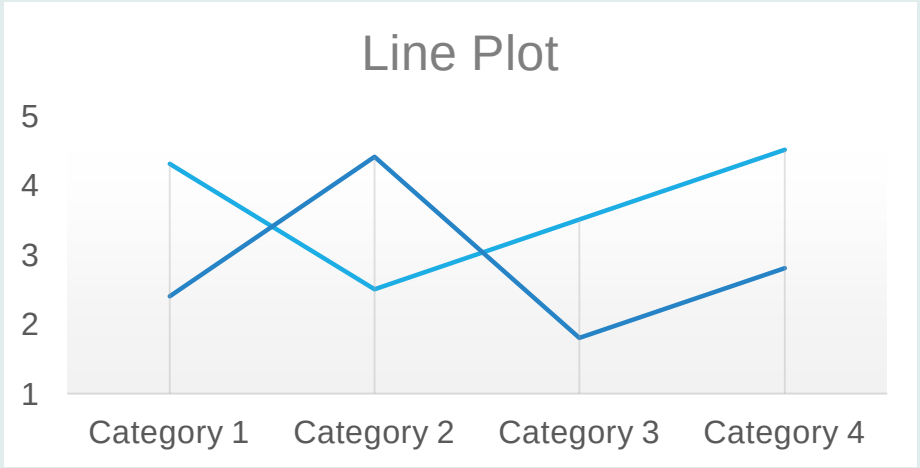
Suatu proses visualisasi statistic yg digunakan untuk mempermudah penyajian data.





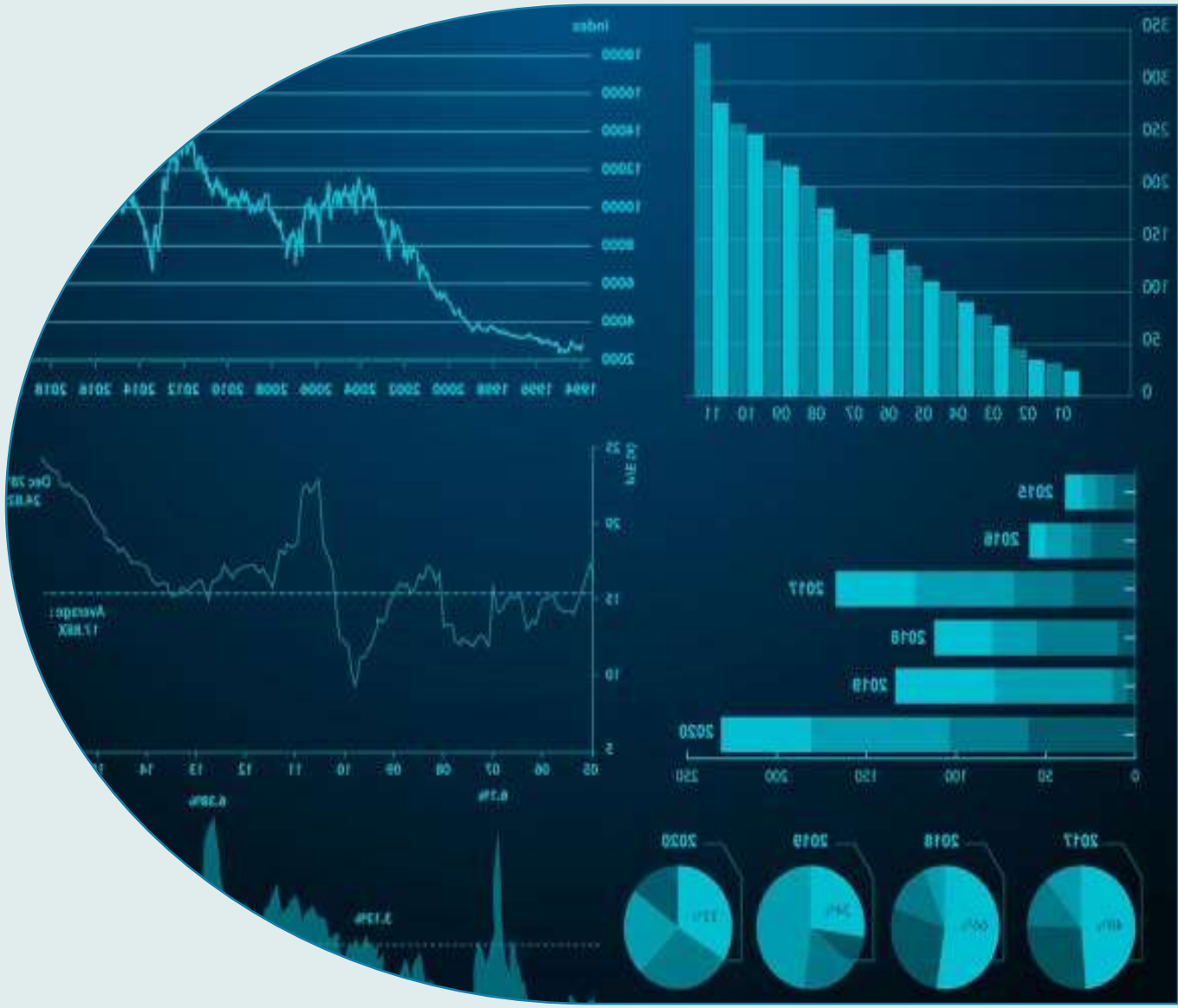






03

ADVANCED STATISTICS



DigitalSkola

Content

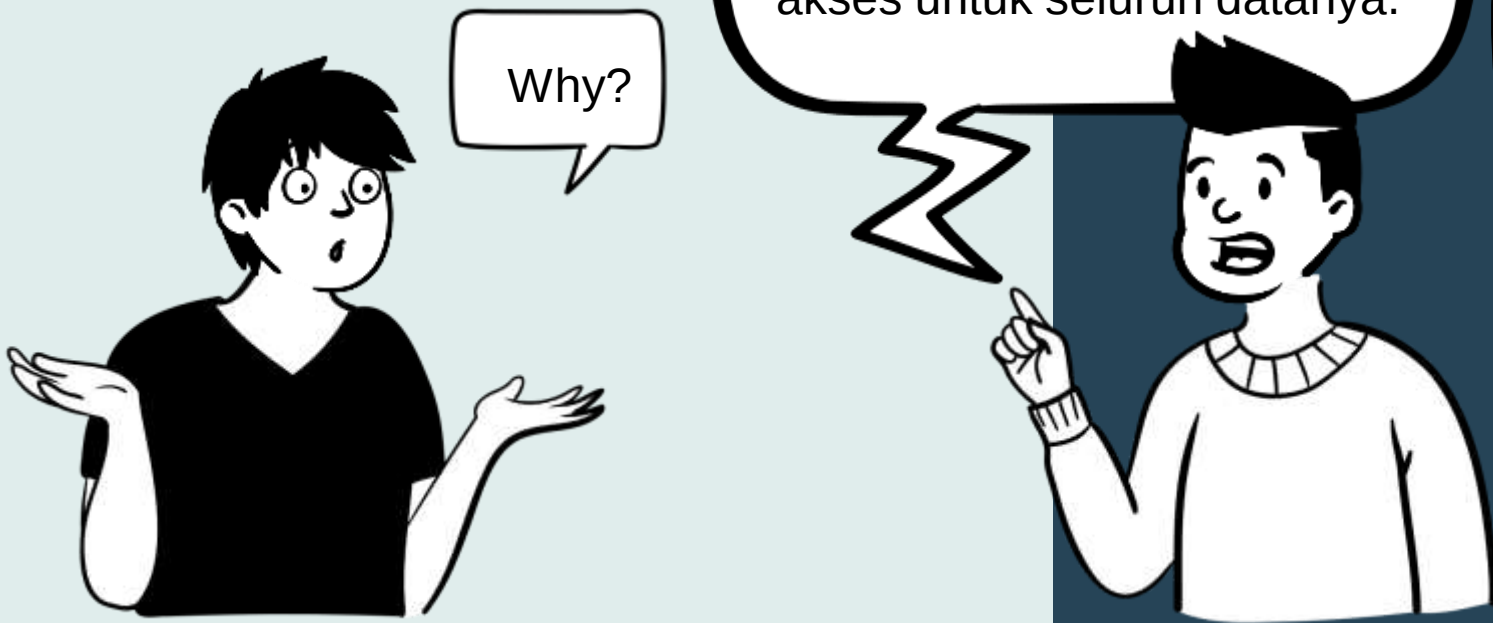
Sampling

Hypothesis Testing

AB Testing

Sampling

Mengacu pada metode statistik untuk memilih pengamatan dengan tujuan memperkirakan parameter populasi.



Mengumpulkan seluruh data akan sangat sulit, mahal dan memakan waktu yang banyak.

Observasi lanjutan dapat dilakukan jika sampling belum terpenuhi.

Pengembangan data di kemudian hari untuk analisis/penelitian lain.

DigitalSkola

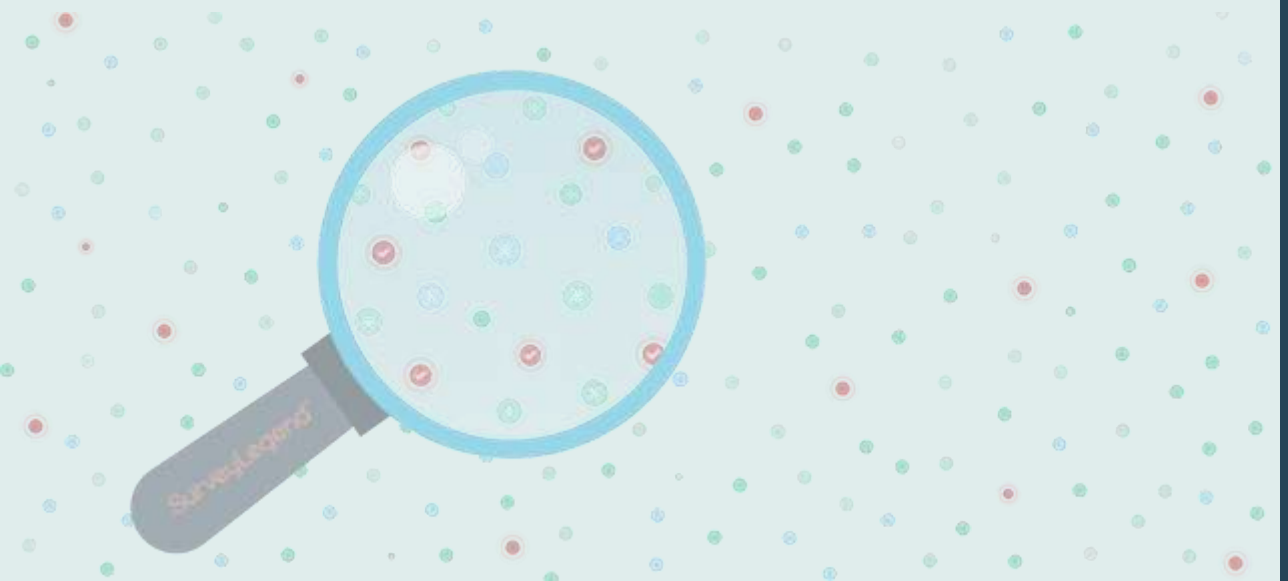
Aspek pertimbangan dalam mengumpulkan data :

- **Tujuan sampel** → Bagian dari populasi yang ingin anda perkirakan.
- **Population** → Ruang lingkup dari mana pengamatan anda dimulai
- **Kriteria Seleksi** → Metodologi yang digunakan untuk mengambil spesifik informasi dari observasi
- **Ukuran sampel** → Banyaknya pengamatan yang akan dijadikan sampel

DigitalSkola

Teknik Sampling

DigitalSkola



Pengambilan sampel statistik adalah bidang studi yang luas, tapi dalam pembelajaran "*applied machine learning*", terdapat tiga jenis Teknik Sampling :

1. ***Simple Random Sampling*** : Sampel yang diambil dengan probabilitas seragam dari populasi
2. ***Systematic Sampling*** : Sampel yang diambil menggunakan pola yang ditentukan sebelumnya dengan bantuan interval
3. ***Stratified Sampling*** : Sampel yang diambil dengan kategori yang ditentukan

Hypothesis Testing

Digunakan untuk melakukan praduga / prediksi / hipotesa / dugaan dari populasi dan sampel data yang ada.

DigitalSkola

Tujuan Hypothesis Testing

Menjawab pertanyaan hipotesis null

Melanjutkan pengujian ke level berikutnya

Istilah dalam statistik

Ho (Null Hypothesis)

Dugaan awal yg diujikan

Ha (Alternative Hypothesis)

Alternative dari dugaan awal

Confidence Level

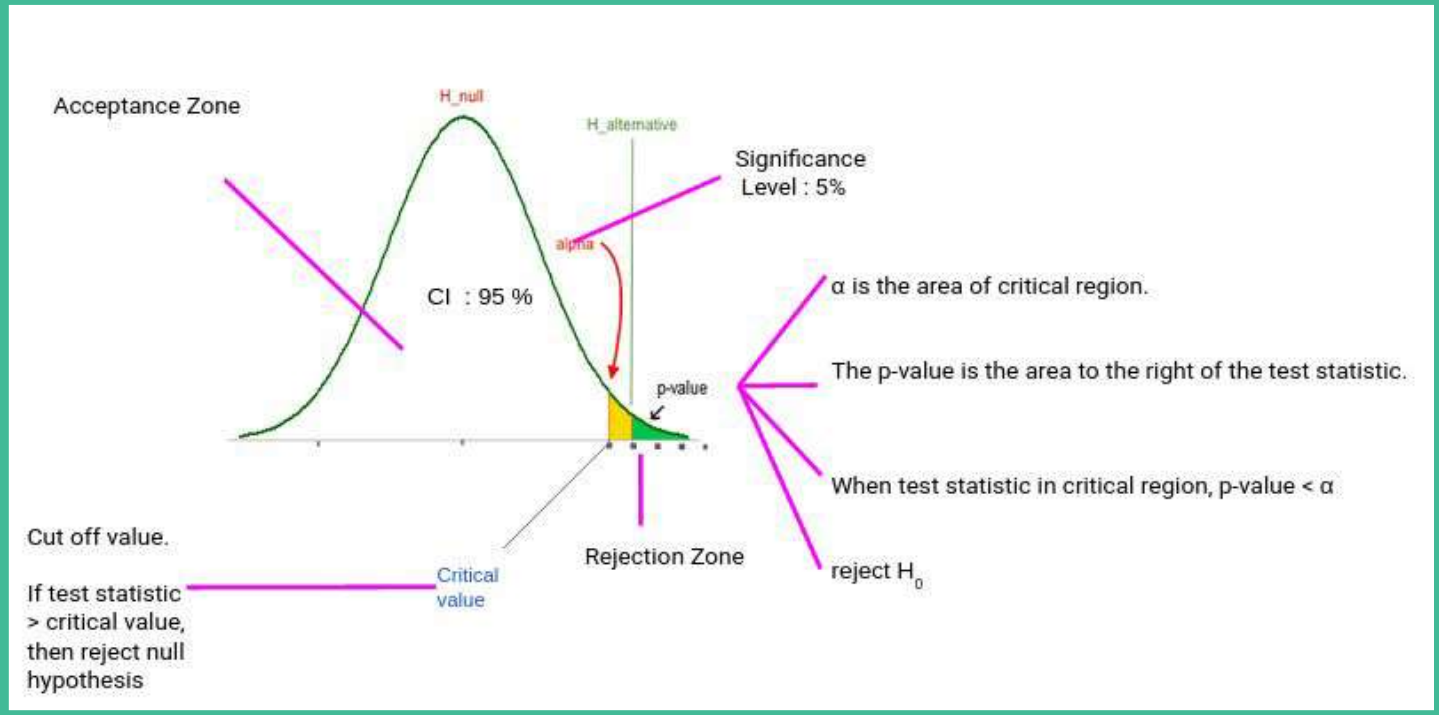
Matriks yg ditentukan untuk validasi

p-value

Nilai dari tes hipotesis

α (Significant Level)

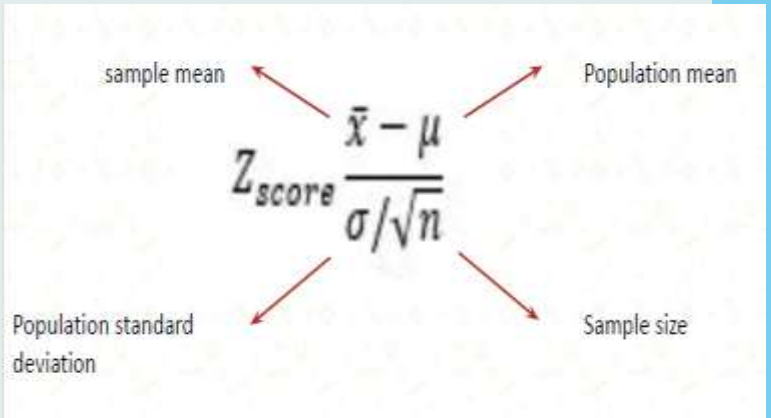
Potongan nilai antara penerimaan dan penolakan hipotesis

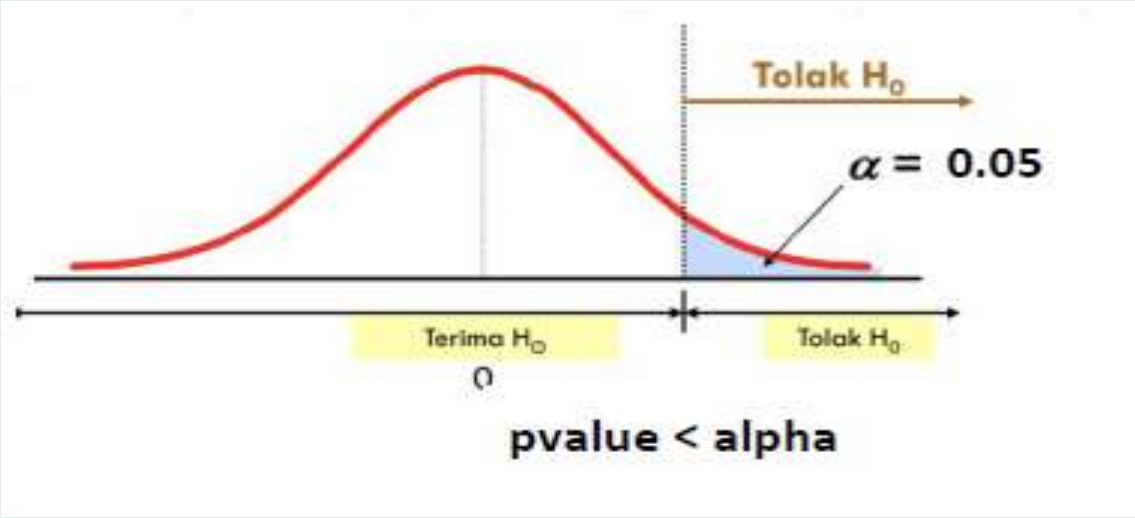


Z-Test

Digunakan saat:

- ✗ Diketahui varians dari populasi.
- ✗ Jika tidak ada varians dari populasi, ukuran sampel harus melebihi 30 data.





Contoh kasus:

Transformer memiliki rata-rata CO_2 sebesar 1186, pada saat bekerja transformer apakah CO_2 yang dihasilkan lebih tinggi dari rata-rata CO_2? Sampel yang diambil adalah 35 data.

Hipotesis:

- Ho: rata-rata > 1186, rata-ratanya lebih besar dari 1186
- H1: rata-rata < 1186, rata-ratanya tidak lebih besar dari 1186

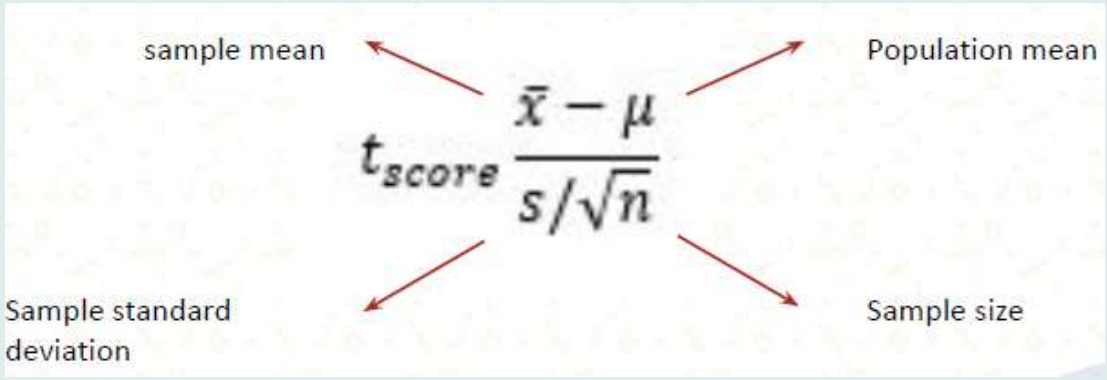
Jawabannya:

Misalkan alpha = 0.05 yang digunakan untuk uji hipotesis ini dan n = 35, maka area-nya sebagai berikut

T-Test

Digunakan saat:

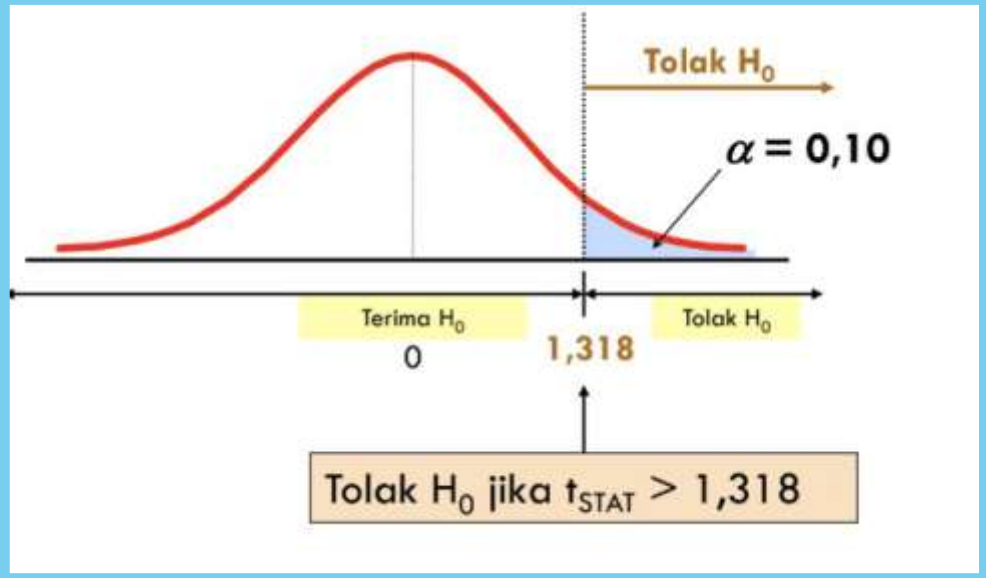
- ✗ Tidak diketahui varians populasi
- ✗ Jumlah sampel data kecil, n ≤ 30



Contoh Kasus:

Seorang manajer penyedia layanan telepon selular berpendapat bahwa telah terjadi peningkatan tagihan telepon pelanggan, sehingga rata-ratanya menjadi lebih dari $52 per bulan. Perusahaan ingin menguji pernyataan ini.

Terdapat 25 sampel. (Diasumsikan populasi berdistribusi normal)



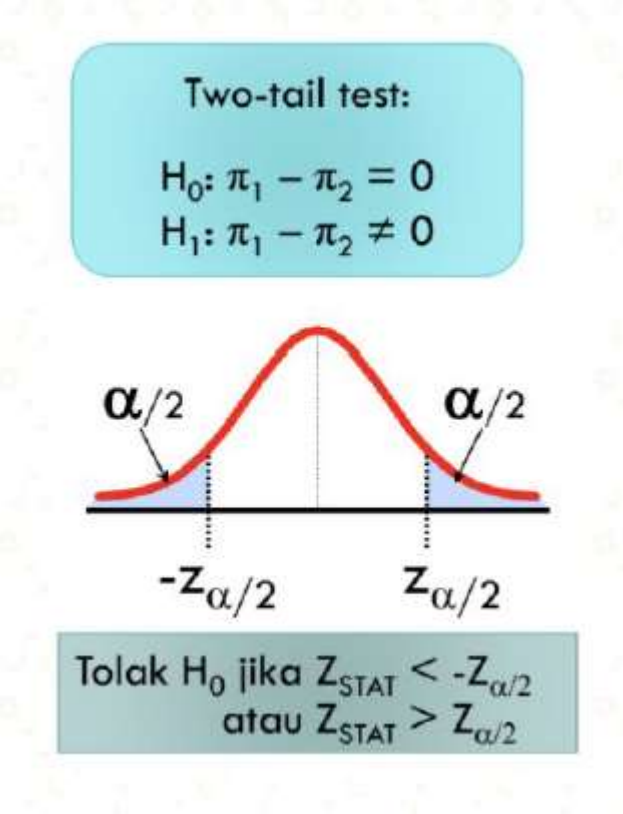
Hipotesis:

- Ho: rata-rata <= 52, rata-ratanya tidak lebih dari $52 per bulan
- H1: rata-rata >52, rata-ratanya lebih dari $52 per bulan

Jawabannya:

Misalkan alpha = 0.1 yang digunakan untuk uji hipotesis ini dan n = 25, maka area-nya sebagai berikut

Two Sample Testing



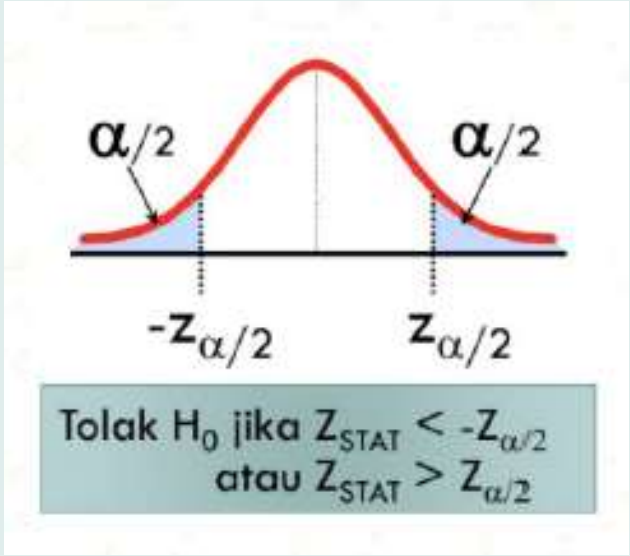
Transformer memiliki karakter metana yang dihasilkan, ketika transformer bekerja dan tidak bekerja apakah rata-rata metana yang dihasilkan sama?

Hipotesis:

- Ho: rata-rata metana saat bekerja = rata-rata metana saat tidak bekerja
- H1: rata-rata metana saat bekerja != rata-rata metana saat tidak bekerja

Jawabannya:

Misalkan alpha = 0.05 yang digunakan untuk uji hipotesis ini dan n = 30, maka area-nya sebagai berikut



AB Testing

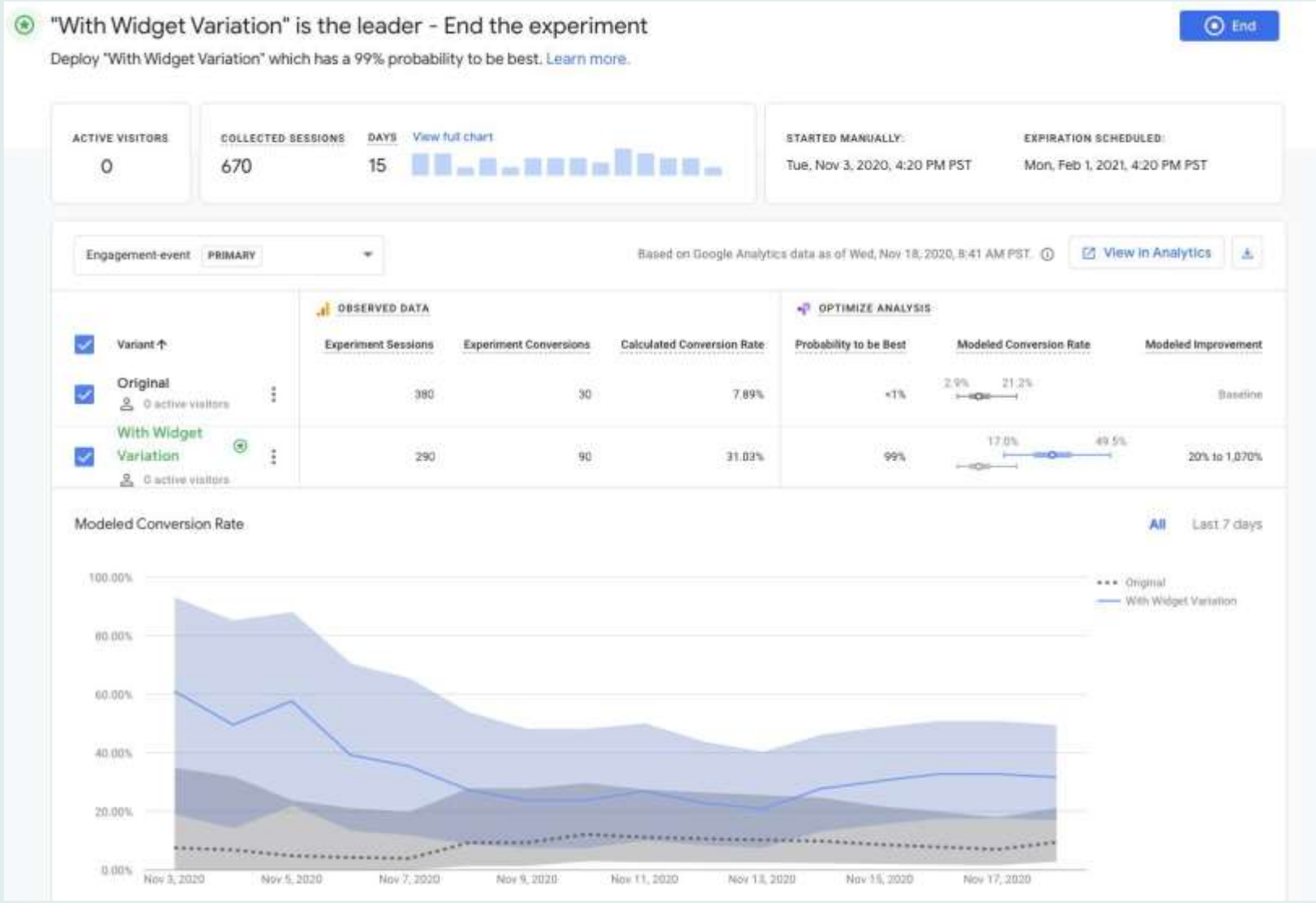
Menguji varian pengamatan anda dan melihat bagaimana kinerjanya terhadap tujuan yang ditentukan

DigitalSkola

Step-by-Step AB Testing in Website

- **Apa tujuan anda?** mendefinisikan tujuan dari eksperimen anda. Contohnya adalah jumlah klik, jumlah submit, user dan pageview
- **Apa hipotesis anda?** dengan menambahkan form submit antara kedua artikel dapat meningkatkan jumlah user dalam subscribe
- **UI/UX team and front end engineer** meminta mereka dalam mendesain fitur A dan B dalam website anda
- **Setting platform AB testing** sebagai seorang data scientist/data analyst perlu paham untuk memasang tracker untuk mengambil data
- **Pada pengaturan platform AB testing** anda dapat melakukan setting 50% - 50% (usually), dan setingan lainnya terkait data apa yang ingin diperoleh

Platform AB Testing



DigitalSkola

THANK YOU!

DigitalSkola