

STATISTIKA

TERAPAN

Disampaikan oleh:

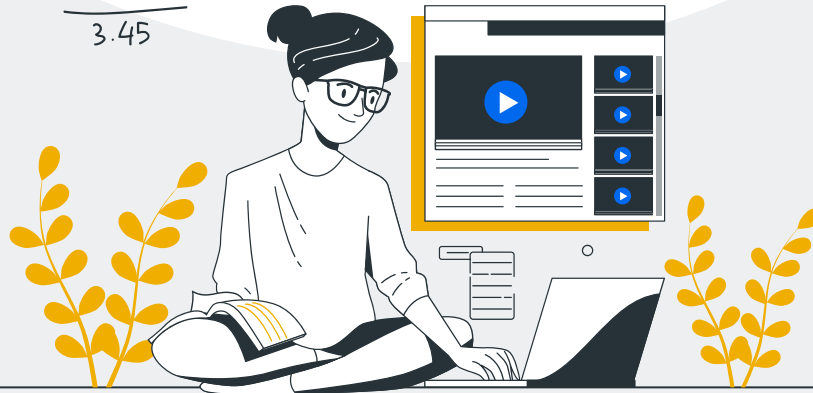
Yohanes Bella Kurniawan

Nuriko Chandra Septiorini

$$\frac{10+17}{3.45}$$

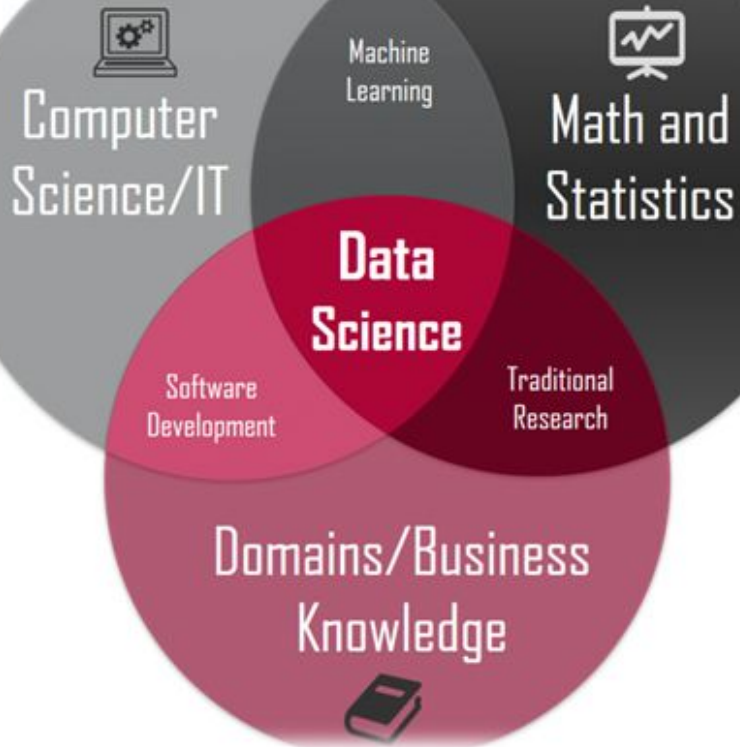
$$C = \frac{B^3 + C^2 + A}{3BA}$$

$$\left(\frac{C-B}{3-D} \right) = \left(\frac{A}{3B} \right) = \frac{3C(2)^4}{X+Y+C}$$



PJJ Data Analytics IV

September 2022



$$\frac{4+6+(2\sqrt{3})}{\sqrt{276}}$$

$$\frac{10+17}{3.45}$$

$$\frac{\sqrt{2.8}}{3+2^+}$$

Outline

01

Intro to Statistics

02

Descriptive Stats

03

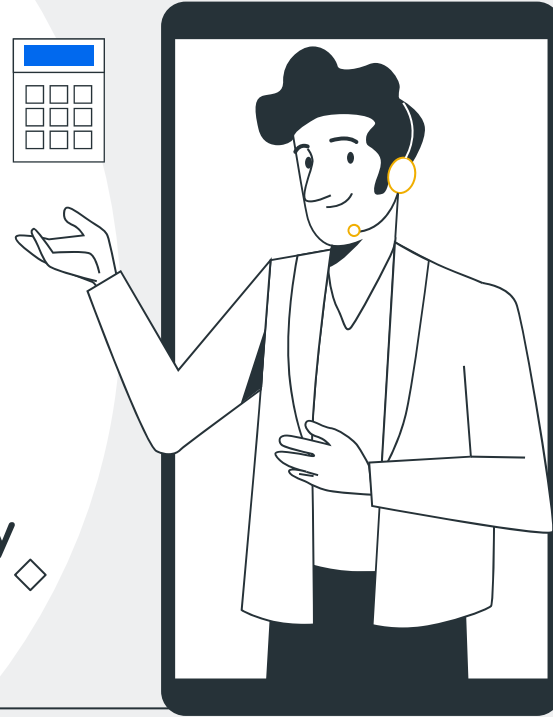
Visualizing Data

04

Sampling and Probability

05

Inference Stats



$$\frac{3C(2)^4}{x+y+c}$$

$$\frac{\sqrt{2.8}}{3+2^+}$$



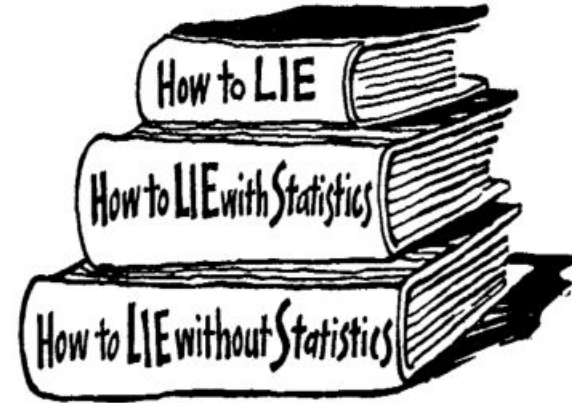
01

Intro to Stats

$$\frac{A}{3B}$$



“There are three kinds of lies:
lies, damned lies, and statistics”
(B. Disraeli / M. Twain)



$$\frac{5 \pm \sqrt{3-4}}{2}$$

Terminologi

◇ Statistika (Statistics)

Teknik mengumpulkan data, menganalisa, menyimpulkan dan menafsirkan data yang berbentuk angka (Hall, 1892)

Populasi

Keseluruhan objek penelitian yang menjadi sumber data

Parameter
Sama dengan statistik namun perbedaannya adalah sumber data berasal dari populasi

Sampel

Bagian dari populasi yang dipilih dengan menggunakan metode tertentu dan diharapkan dapat menggambarkan karakteristik populasi

Statistik

Data hasil pengukuran dalam statistika yang dapat menggambarkan suatu keadaan atau masalah

Tipe Data

Berdasarkan Skala Pengukuran

Kategorik

Nominal

Tidak memiliki urutan
Contoh: Warna

Ordinal

Memiliki urutan
Contoh: Rating Ojek Online

Numerik

Interval

Memiliki urutan
Tidak memiliki nol mutlak
Contoh: Suhu 0 derajat bukan berarti tidak ada suhu sama sekali, melainkan suhu di tempat tersebut sangat dingin

Rasio

Memiliki urutan
Memiliki nol mutlak
Contoh: Tinggi badan peserta PJJ



Karakteristik Tipe Data

Karakteristik	Nominal	Ordinal	Interval	Rasio
Modus	👍	👍	👍	👍
Median		👍	👍	👍
Mean			👍	👍
Penambahan dan Pengurangan			👍	👍
Perkalian dan Pembagian				👍

$$\frac{\sqrt{2}}{(\frac{1}{2})^2}$$

$$\frac{10+17}{3.45}$$

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leon	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elis)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. William H	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D King)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S

$$\frac{10+17}{3.45}$$

Tentukan Tipe Data pada Dataset berikut

Type of Statistic

$$\frac{10+17}{3.45}$$

Statistic

Descriptive Statistic

Summarize data, use the data to provide description of the population, through numerical calculation or graphs and table.



Measure of Location/ central tendencies :

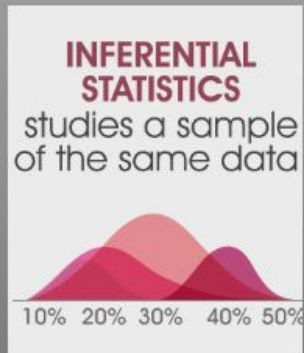
- Median
- Mean
- Modus

Measure of spread (dispersion, skewness):

- Range
- Inter quartile Range
- Variance
- Standard deviation

Inferential Statistic

which test for significant *differences* between groups and/or significant *relationships* among variables within the sample



Hypothesis Testing
t-ratio, chi-square, beta-value

$$\sqrt{276}$$

$$\frac{\sqrt{2.8}}{3+2^+}$$



02

Statistika Deskriptif

Statdes Vocabulary

Central Tendency

01

Mean

02

Median

03

Modus

Dispersion

01

Range

02

Varians

03

Standar
Deviasi

04

Interquartile
Range

Central Tendency

Mean

Average, the sum of the observed values divided by the number of observations.

Population Mean

$$\mu = \frac{\sum_{i=1}^N x}{N}$$

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^n x}{n}$$

Median

Middle value of data when sorted in order of magnitude, **50th percentile**

Sales Sorted Sales

9	6
6	9
12	10
10	12
13	13
15	14
16	14
14	15
14	16
16	16
17	16
16	17
24	17
21	18
22	18
18	19
19	20
18	21
20	22
17	24

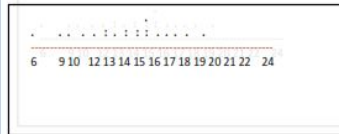
← Median

$$(20+1)50/100=10.5$$

$$16 + (.5)(0) = 16$$

Mode

Most frequently- occurring value



Mode = 16

- Menggambarkan pusat atau nilai tengah dari distribusi
- Mean terpengaruh oleh outlier
- Mode dan Median tidak terpengaruh oleh outlier
- Mean menggambarkan terjadinya redistribusi

Measures of dispersion

<u>Sales</u>	<u>Sorted Sales</u>	<u>Rank</u>	
9	6	1	← Minimum
6	9	2	
12	10	3	
10	12	4	
13	13	5	
15	14	6	← First Quartile
16	14	7	
14	15	8	
14	16	9	
16	16	10	
17	16	11	
16	17	12	
24	17	13	
21	18	14	
22	18	15	← Third Quartile
18	19	16	
19	20	17	
18	21	18	
20	22	19	
17	24	20	← Maximum

Range Maximum - Minimum =
 $24 - 6 =$
 18

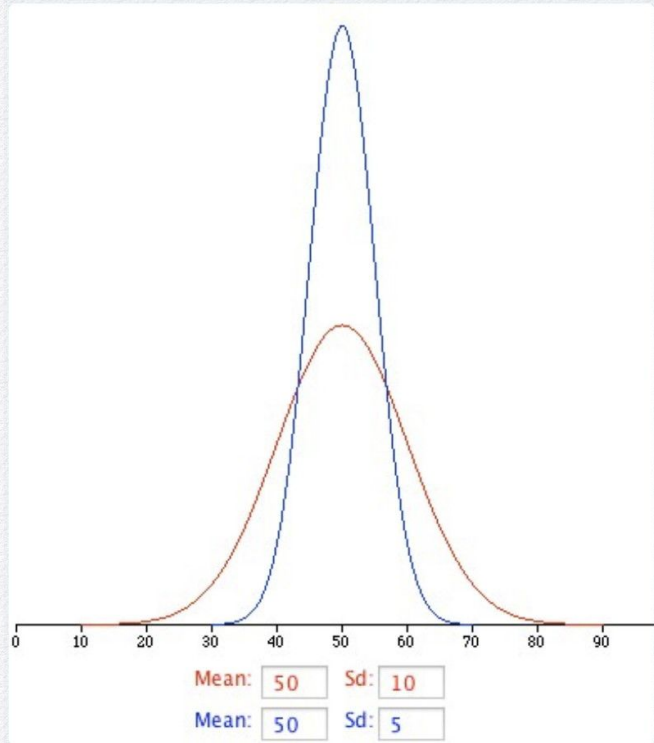
$Q_1 = 13 + (.25)(1) = 13.25$

$Q_3 = 18 + (.75)(1) = 18.75$

Interquartile Range $Q_3 - Q_1 =$
 $18.75 - 13.25 = 5.5$

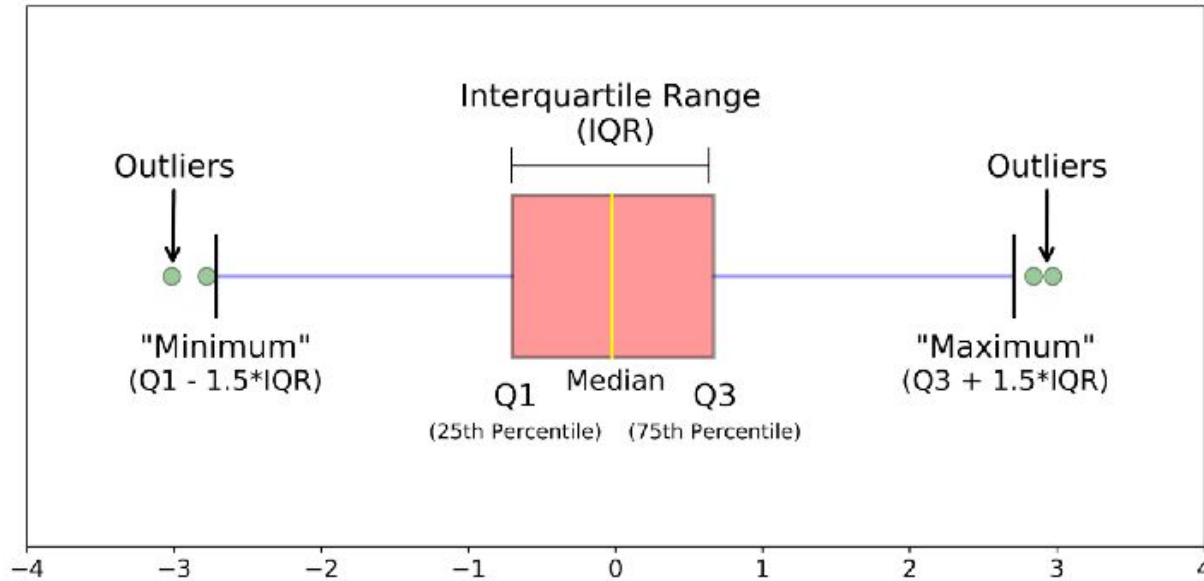
Measures of dispersion

VARIABILITY DEMONSTRATION



- Varians menggambarkan sebaran data
- Semakin besar nilai varians maka data semakin bervariasi
- Standar deviasi mengukur variasi antar data cluster di sekitar rata-rata

Outliers



- Data yang berkarakteristik unik terlihat sangat berbeda jauh dengan data lainnya
- Deteksi bisa menggunakan boxplot atau standardized residual

Diskusi

mean	79.6
median	65
std deviasi	17.71534
max	100
min	60
range	40
q1	63
q2	65
q3	97
count	25

Dari statistic atas nilai ujian matematika berikut

Kira-kira insight apa yang dapat diambil?

Kalau ditambahkan informasi, nilai batas lulus adalah 65, bagaimana?

$$\frac{\sqrt{2.8}}{3+2^+}$$



03

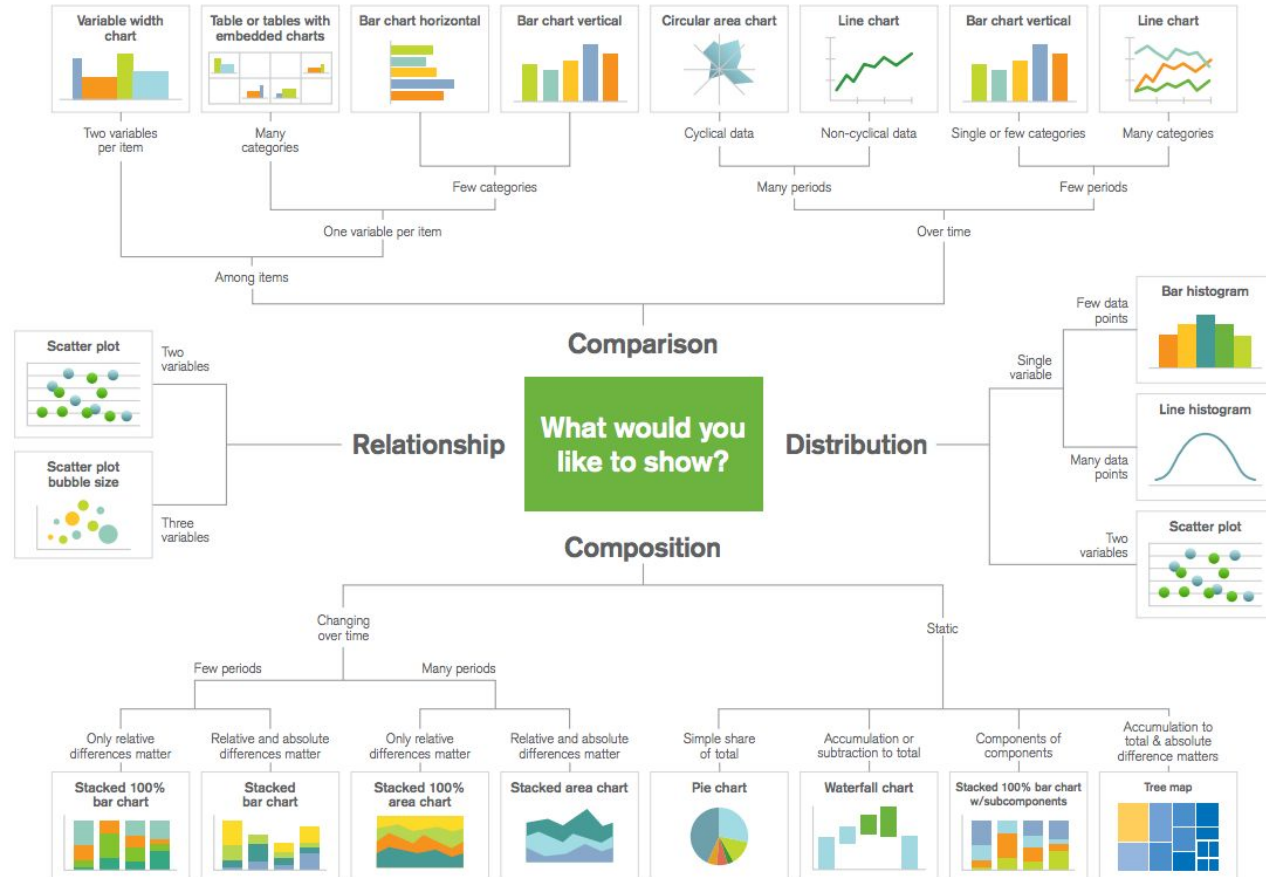
Visualizing Data

Data Visualization

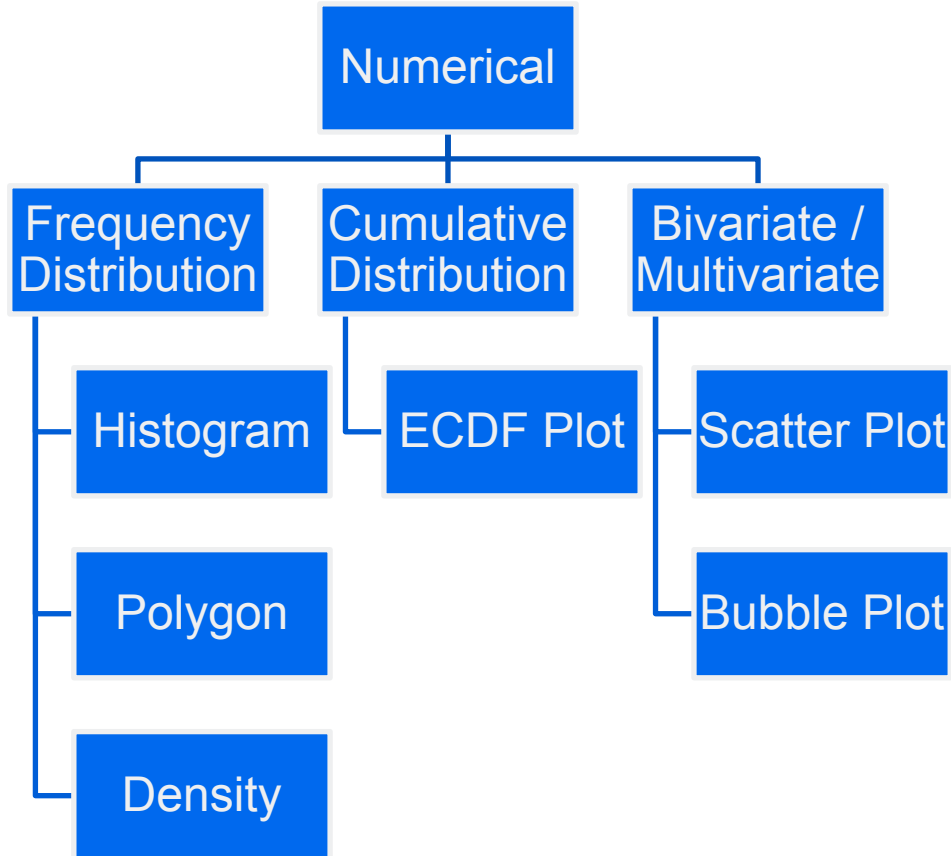
Referensi:

<https://towardsdatascience.com/data-visualization-101-how-to-choose-a-chart-type-9b8830e558d6>

<https://huynp.com/2018/07/19/How-to-choose-data-visualization-techniques.html>



Numerical



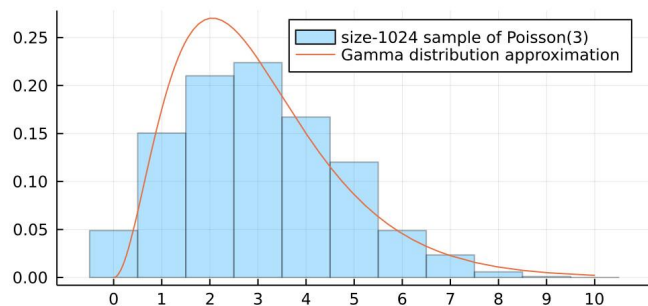
Numerical

Data in ordered array:

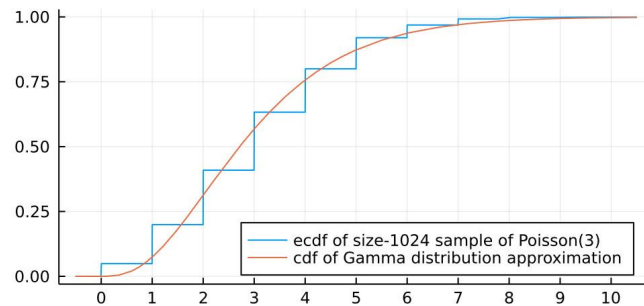
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

Numerical

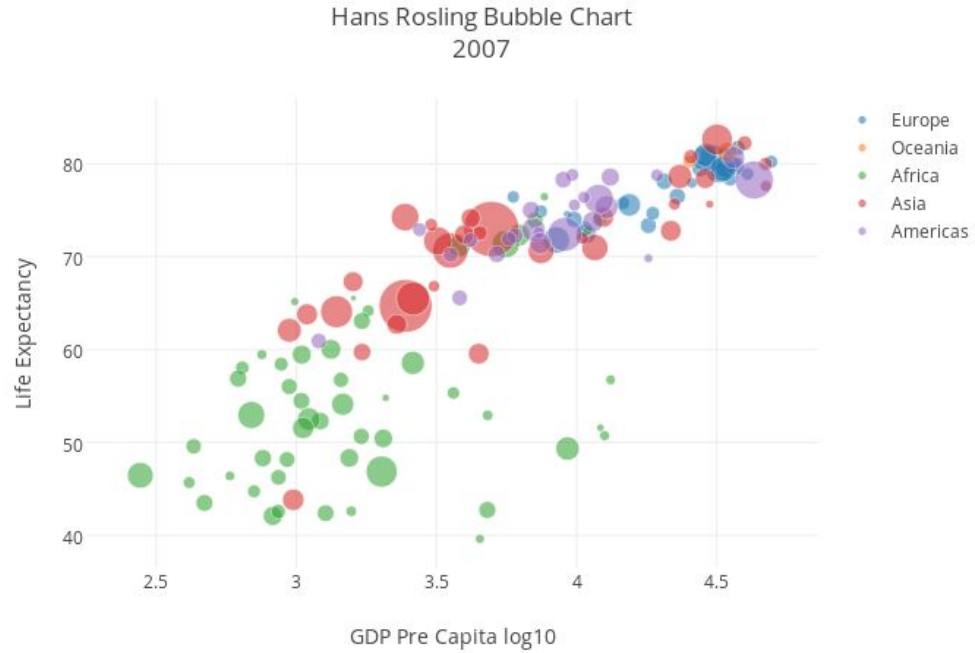


Histogram + KDE

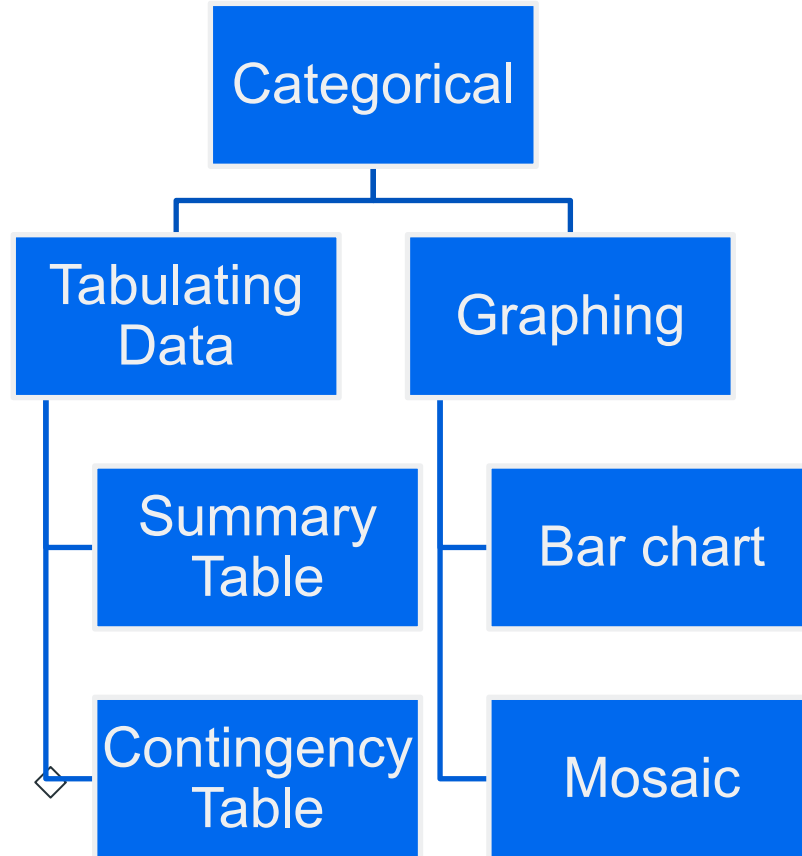


ECDF + CDF

Numerical



Categorical

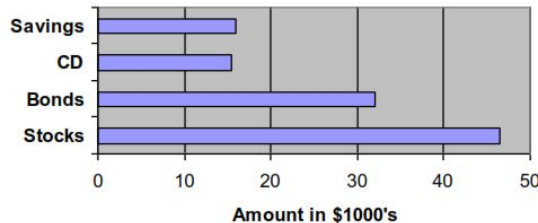


Categorical

Univariate

Investment Type	Amount (in thousands \$)	Percentage (%)
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	110.0	100.0

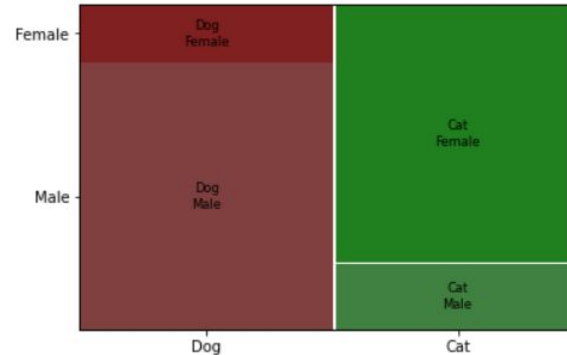
Summary Table



Bivariate

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

Contingency Table / Crosstab



$$\frac{\sqrt{2.8}}{3+2^+}$$

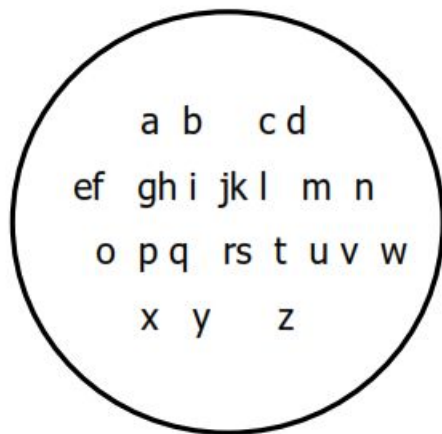


04

Sampling and Probability

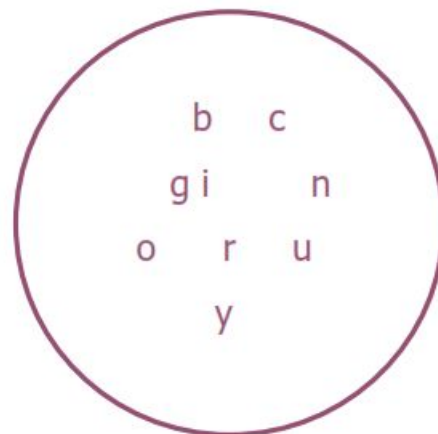
Population & Sample

Population

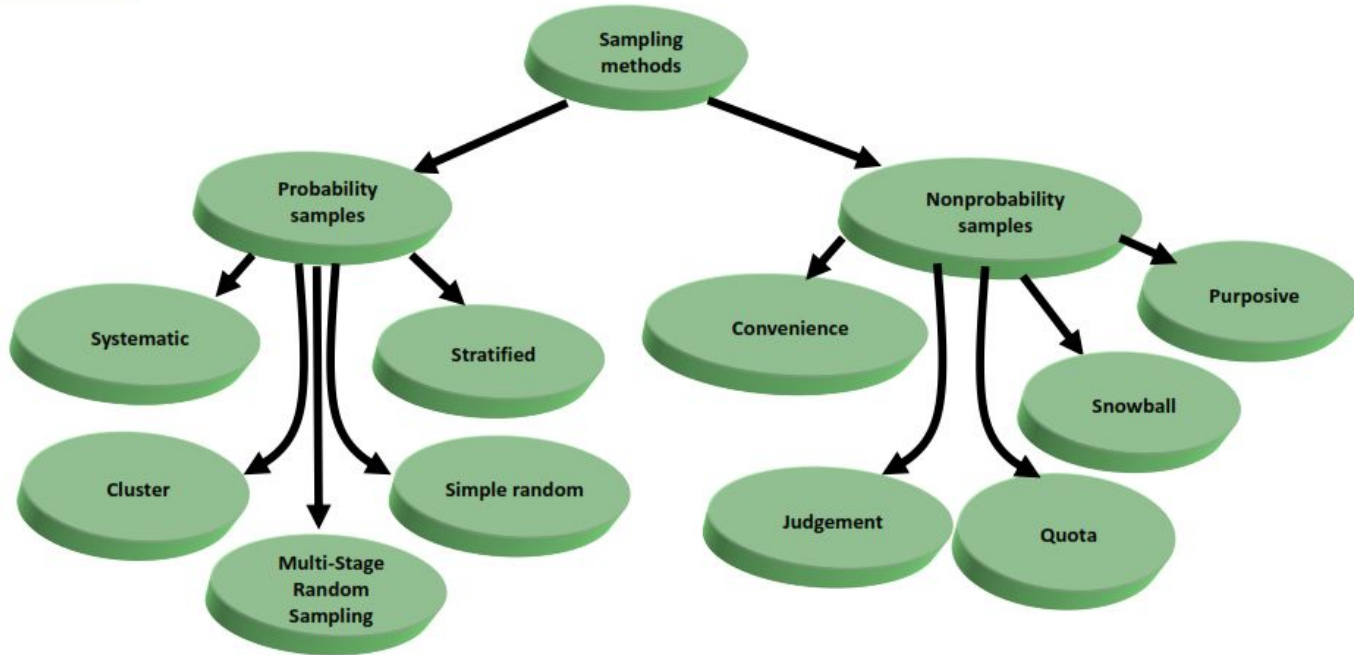


VS

Sample



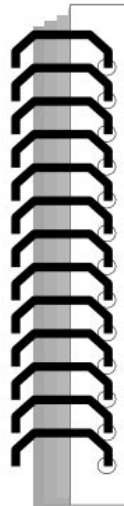
Sampling Method



Probabilistic

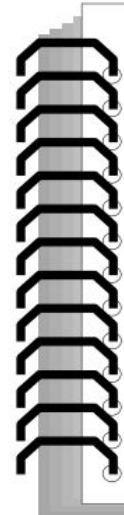
$$\frac{\sqrt{2}}{\left(\frac{1}{2}\right)^2}$$

Simple Random Sampling



1	Albert D.
2	Richard D.
3	Belle H.
4	Raymond L.
5	Stéphane B.
6	Albert T.
7	Jean William V.
8	André D.
9	Jeremy W.
10	Anthony Q.
11	James B.
12	Denis G.
13	Amanda L.
14	Jennifer L.
15	Philippe K.
16	Eve F.
17	Priscilla O.
18	Robert D.
19	Brian F.
20	Hellène H.
21	Isabelle R.
22	Jean T.
23	Samanta D.
24	Berthe L.

Systematic Sampling



1	Albert D.
2	Richard D.
3	Belle H.
4	Raymond L.
5	Stéphane B.
6	Albert T.
7	Jean William V.
8	André D.
9	Jeremy W.
10	Anthony Q.
11	James B.
12	Denis G.
13	Amanda L.
14	Jennifer L.
15	Philippe K.
16	Eve F.
17	Priscilla O.
18	Robert D.
19	Brian F.
20	Hellène H.
21	Isabelle R.
22	Jean T.
23	Samanta D.
24	Berthe L.

$$\frac{4+6+(2\sqrt{3})}{\sqrt{276}}$$



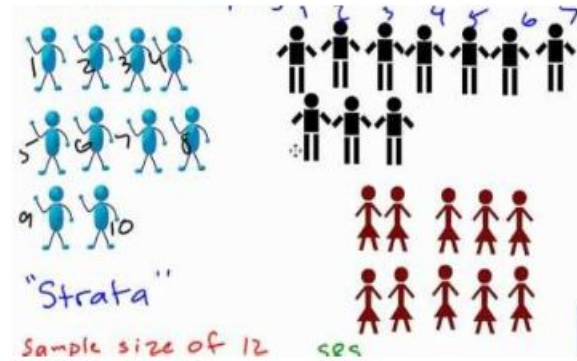
$$\frac{C^3 + 5CA}{2CA}$$

$$\frac{C - B}{3 - D}$$

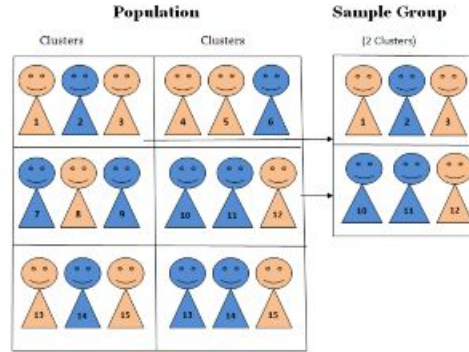


Probabilistic

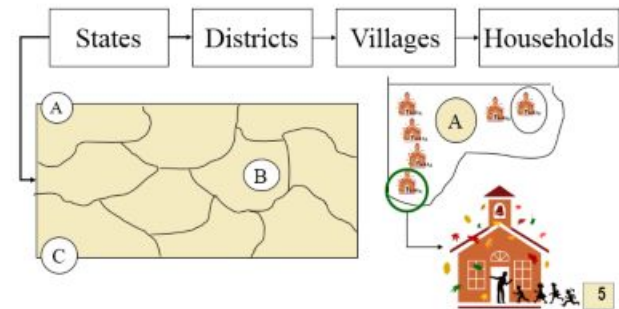
Stratified Sampling



Cluster Sampling



Multistage Sampling





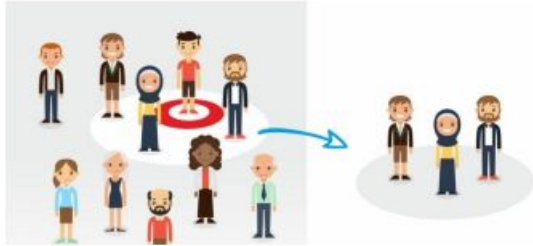
$$\frac{C^3 + 5CA}{2CA}$$

$$\frac{C - B}{3 - D}$$

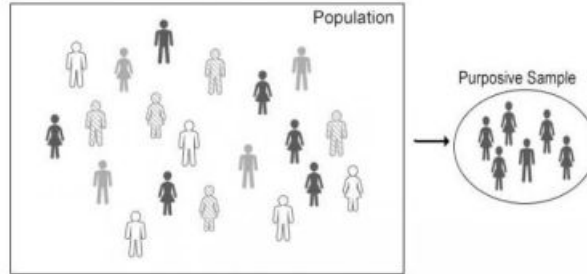


Non - Probabilistic

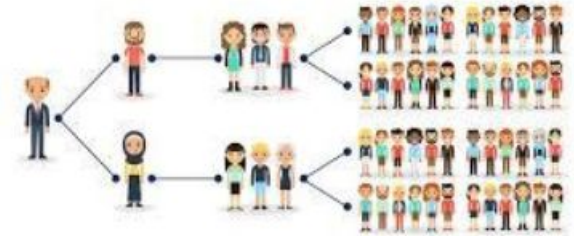
Convenience Sampling



Purposive Sampling



Snowball Sampling





$$\frac{C^3 + 5CA}{2CA}$$

$$\frac{C - B}{3 - D}$$

Sampling Error

Reducing Sampling & Non Sampling Errors

**Sampling
Error**



Cause: Small,
Un-diverse Sample



Solution: Bigger,
More Diverse Sample

**Non-Sampling
Error**



Cause: External
Factors



Solution: Study
Mechanism Design

Peluang

Event

Hasil dari eksperimen

Contoh: Mendapatkan angka 6 dalam melempar satu buah dadu

Sample Space

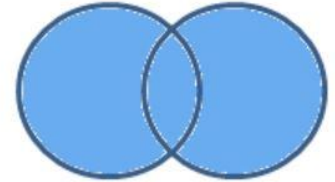
Kumpulan semua kemungkinan hasil eksperimen

Contoh: Kemungkinan angka dalam melempar satu dadu {1,2,3,4,5,6}

Kombinasi antar event

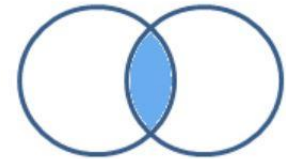
Gabungan

$$P(A \cup B)$$



Irisan

$$P(A \cap B)$$



Disjoint



$$\frac{\sqrt{2.8}}{3+2^+}$$



05

Inference Statistic

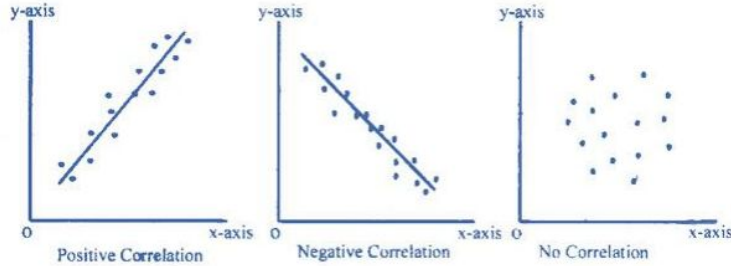
Korelasi

- Digunakan untuk menemukan hubungan antara dua variabel kuantitatif
- **Kausalitas:** variabel X menyebabkan perubahan pada variabel Y
- Memiliki rentang nilai antara -1 hingga 1

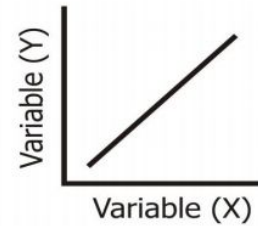
Note:

- Jika X dan Y berkorelasi, bisa jadi X dan Y memiliki hubungan sebab akibat bisa jadi tidak
- Jika X dan Y memiliki hubungan sebab akibat, X dan Y pasti berkorelasi

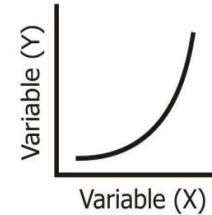
Korelasi



Berdasarkan perubahan proporsi



Linier



Non Linier

Berdasarkan derajat korelasi

- Dua variabel = korelasi bivariat
- > dua variabel = multiple correlation (Koefisien Determinasi/ R-square)
- Korelasi hanya menggambarkan arah dan besaran relatif

Uji Korelasi

Numerical Correlation :

It's a measure of the strength and the direction of a linear relationship between two variables.

Pearson

- Type data interval / ratio
- Outlier sangat mempengaruhi
- Data harus terdistribusi normal

Spearman

- Non parametric test
- Interval, ratio, dan ordinal
- Tidak butuh normal distribusi

Kendall

- Non Parametric test
- Similar dengan spearman
- Statistik atas dependency antar variable

Relationship Test

Categorical Relationship :

Determine if there is an association between two or more categorical variables.

Chi Squared
Test

- Menguji apakah ada hubungan signifikan antar variable

Cramer V

- Menguji kekuatan hubungan antar variable kategorikal

Contingency
Table

- Summary atas hubungan antar variables
- Menampilkan probabilitas antar variable

Uji Hipotesis

Bagian dari Statistika Inferensia yang digunakan untuk mengambil kesimpulan untuk populasi berdasarkan sampel yang representatif

Tujuan : memverifikasi apakah H_0 ditolak atau gagal tolak

- H_n (Null Hypothesis) = tidak ada hal baru yang terjadi pada populasi
- H_1 (Alternative Hypothesis) = negasi dari H_0



- Gagal tolak $H_0 \neq$ Terima H_0
- Jika data yg dikumpulkan tidak mendukung hipotesis alternatif, bukan berarti hipotesis nol benar. Namun belum cukup bukti untuk menolak H_n , maka dari itu istilahnya Gagal menolak bukan menerima

Uji t

- Termasuk uji parametrik (sampel mengikuti distribusi normal)
- Digunakan ketika sampel kecil dan tidak diketahui nilai varians dari populasi
- Data berdistribusi normal

Uji t Satu
sampel

Uji t dua
sampel

Uji t
berpasangan



Uji t Satu sampel

- Membandingkan rata-rata sampel dengan suatu nilai yang spesifik
- Sampel independen
- Berdistribusi normal
- Sampel diambil secara random
- Contoh H_0 :
 - $\mu = 0$, $\mu > xx$, $\mu \leq xx$

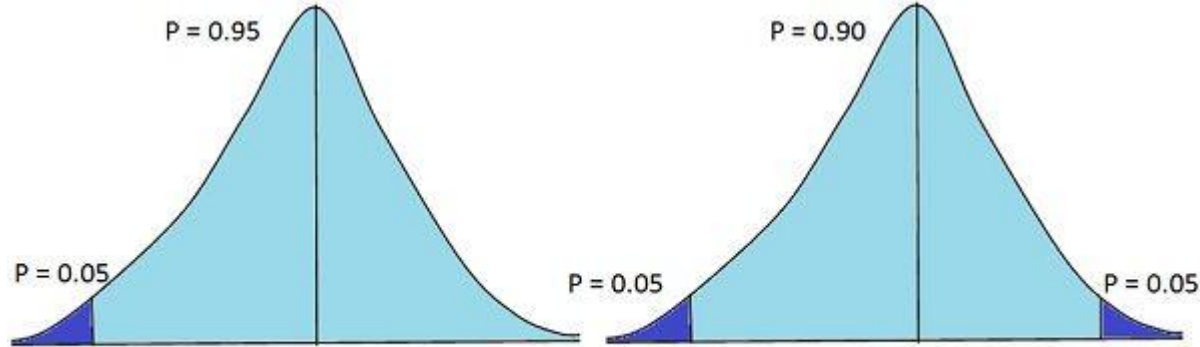
Uji t dua sampel

- Membandingkan rata-rata dua independen sampel
- Sampel independen
- Berdistribusi normal
- Memiliki varians yang sama
- Contoh H_0 :
 - $\mu_1 = \mu_2$, $\mu_1 < \mu_2$

Uji t berpasangan

- Membandingkan dua ukuran entitas yang sama dari waktu ke waktu
- Data berdistribusi normal

Uji t



One-tailed Test Vs Two-tailed Test

Biru tua, signficancy tercapai, H_0 ditolak

Biru muda, signficancy tidak tercapai H_0 gagal ditolak

Type Error

Hypothesis Test	TRUE	FALSE
REJECTED	Type I Error	Correct Decision
NOT REJECTED	Correct Decision	Type II Error

- H_0 benar, ditolak = Error Tipe I (Alpha/False Positif)
- H_0 salah, gagal ditolak = Error Tipe II (Beta/False Negatif)

Besarnya alpha (confident level) mempengaruhi jumlah error
Alpha yang kecil berarti mencari kepercayaan lebih besar untuk mengurangi type I error, namun menambah type II error

◇ $\frac{10+17}{3.45}$

◇ $\frac{\sqrt{2.8}}{3+2^+}$

○

“Statistics is The Grammar of Science”

◇ $\frac{4+6+(2\sqrt{3})}{\sqrt{276}}$

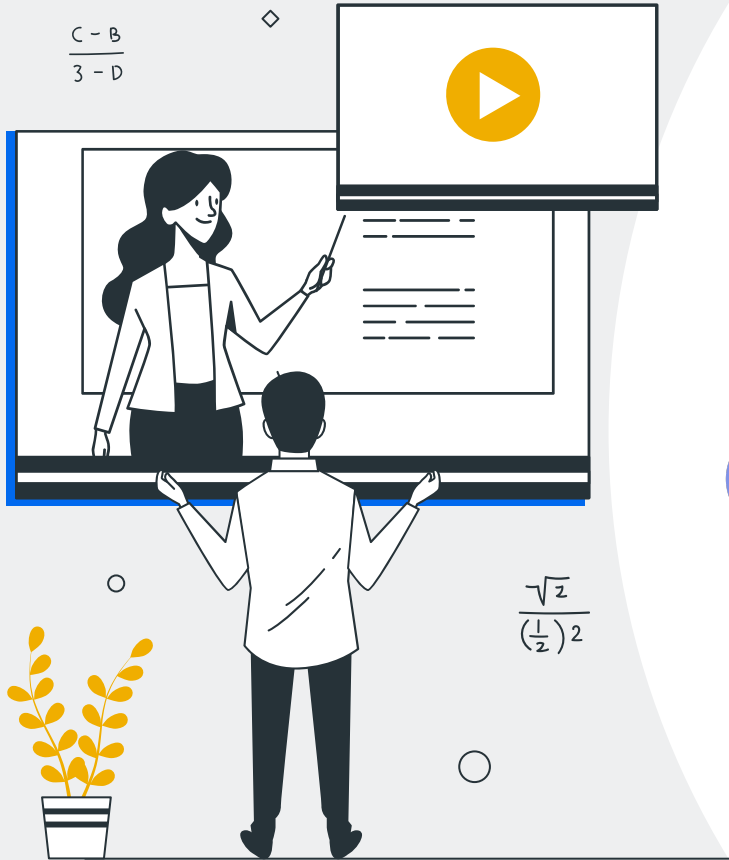
○

○

Karl Pearson

○





Terima Kasih



MoF-DAC



@mof.dac



MoF-DAC | Ministry of
Finance- Data Analytics
Community



mofdac.id

CREDITS: This presentation
template was created by **Slidesgo**,
including icons by **Flaticon**, and
infographics & images by **Freepik**