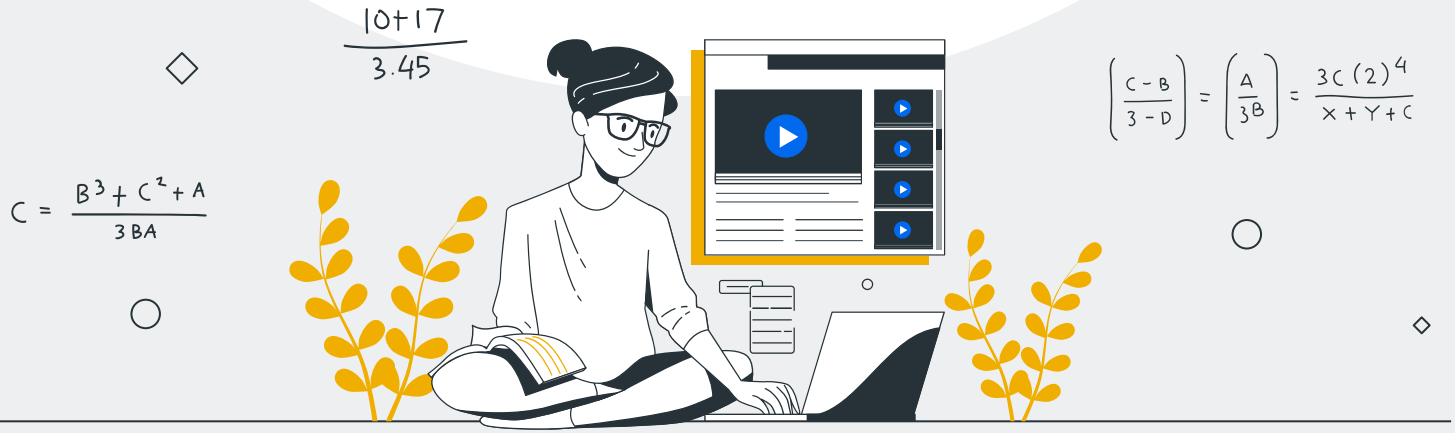
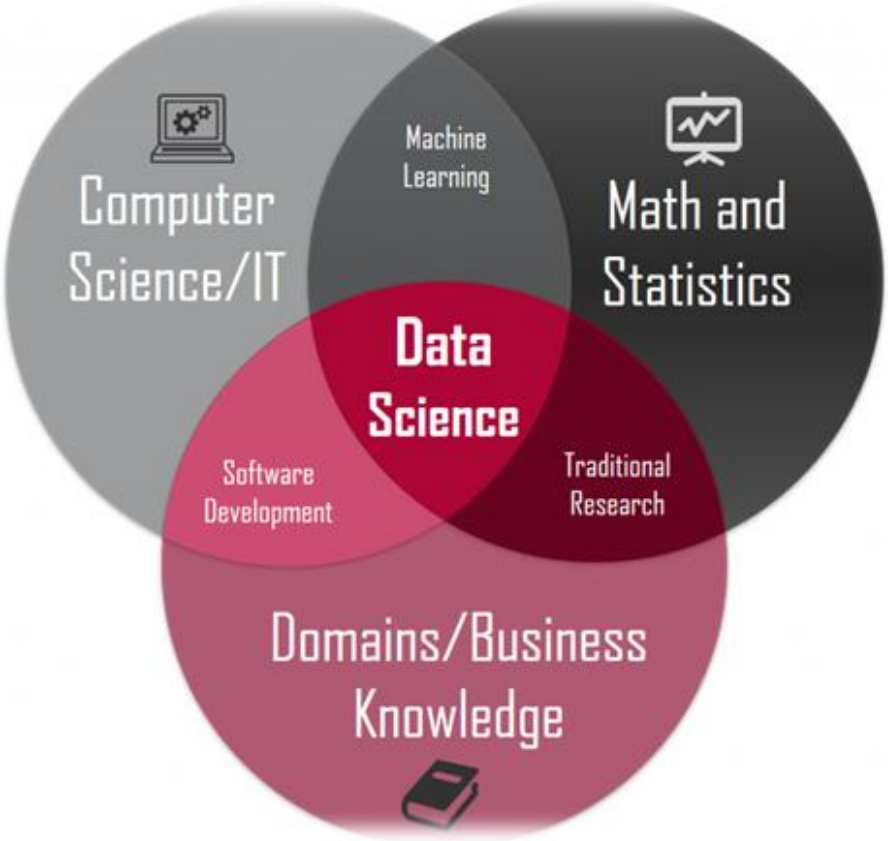
STATISTIKA

TERAPAN

Disampaikan oleh:
Ade Satya Wahana



Pengolahan Data Menggunakan Python
Juli 2023

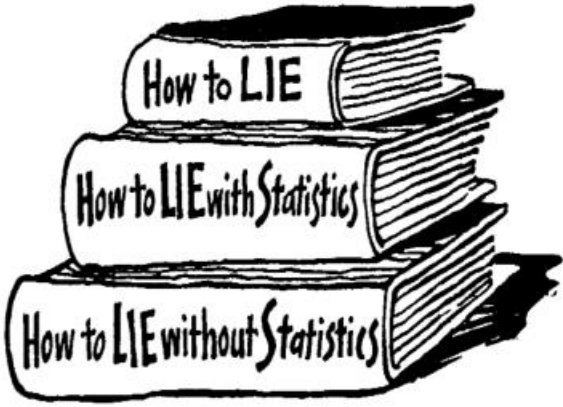

Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

Outline

01

Intro to Stats

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



Terminologi

Statistika (Statistics)

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

Populasi

Keseluruhan objek penelitian yang menjadi sumber data

Sampel

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

Parameter

Sama dengan statistik namun perbedaannya adalah sumber data berasal dari populasi

Statistik

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

Data Type

Text
Boolean
Discrete Number
Datetime

Categorical

Qualitative information classified to their similarities

Nominal

Without order between value

Ordinal

Values can be ordered

Discrete Number
Continuous Numbers

Numerical

values expressing quantitative characteristics

Interval

Lack of absolute true zero

Ratio

Has absolute or true zero

Karakteristik Tipe Data

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

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. Marguerit	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	Saundercock, Mr. William H	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joha	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Aman	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingc	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 Eugen	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

$\sqrt{276}$

02

Statistika Deskriptif

Statdes Vocabulary

Central Tendency

01 Mean

02 Median

03 Modus

Dispersion

01 Range

02 Varians

03 Standar Deviasi

04 Interquartile Range

Asymmetric Distribution

01 Skewness

02 Kurtosis

kdnuggets.com

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**

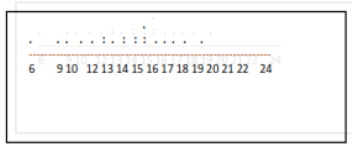
Sales	Sorted Sales
9	6
6	9
12	10
10	12
13	13
15	14
16	14
14	15
14	16 ← Median
16	16
17	16
16	17
24	17
21	18
22	18
18	19
19	20
18	21
20	22
17	24

(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

Sales	Sorted Sales	Rank	
9	6	1	← Minimum
6	9	2	
12	10	3	
10	12	4	
13	13	5	← First Quartile
15	14	6	
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

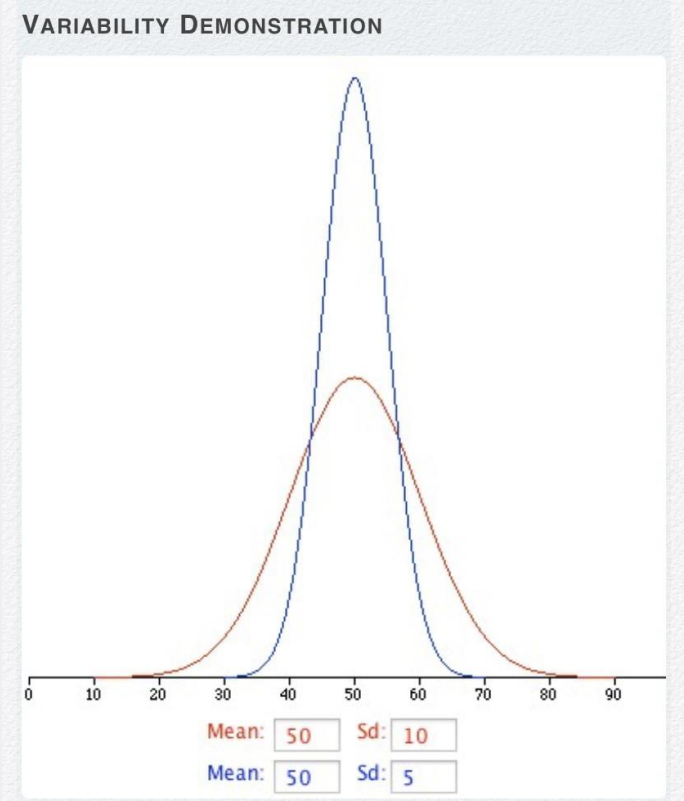
Range Maximum - Minimum =
24 - 6 =
18

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

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

Interquartile Range Q3 - Q1 =
18.75 - 13.25 = 5.5

Measures of dispersion



Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x-\mu)^2}{N} = \frac{\sum_{i=1}^{N} x^2 - \frac{\left(\sum_{i=1}^{N} x\right)^2}{N}}{N}$ $\sigma = \sqrt{\sigma^2}$	$s^2 = \frac{\sum_{i=1}^{n}(x-x)^2}{(n-1)} = \frac{\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}}{(n-1)}$ $s = \sqrt{s^2}$

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

https://onlinestatbook.com/2/summarizing_distributions/spread_sim.html

Diskusi

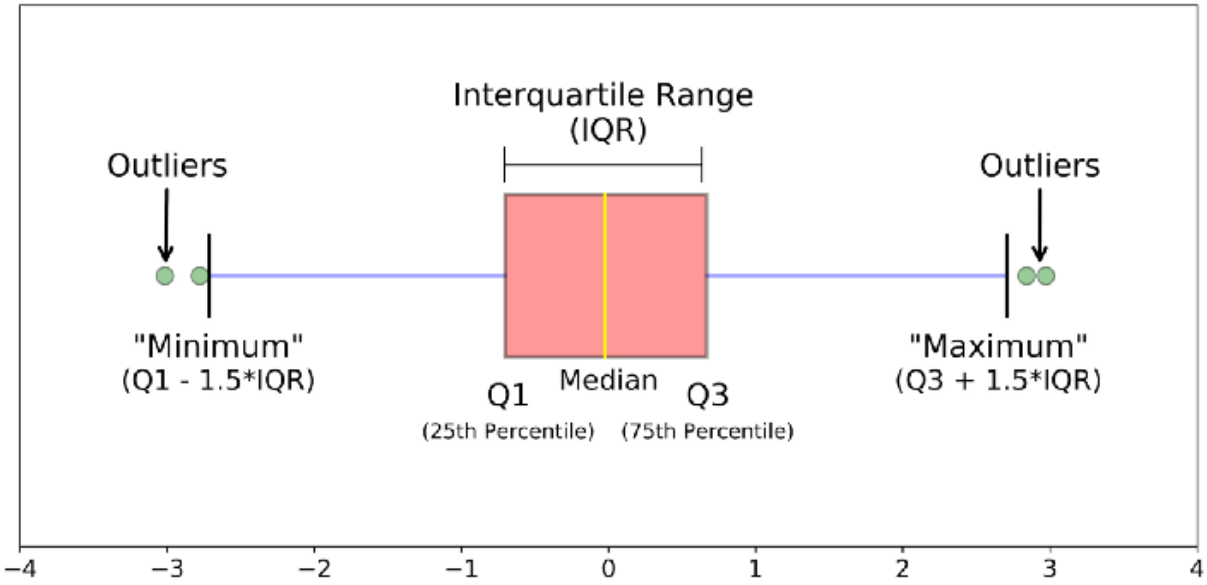
mean	79.6
median	65
std deviasi	17.71534
max	100
min	60
range	40
q1	63
q2	65
q3	97
count	25

Dari statistic atas nilai ujian matematika berikut

Kira-kira insight apa yang dapat diambil?

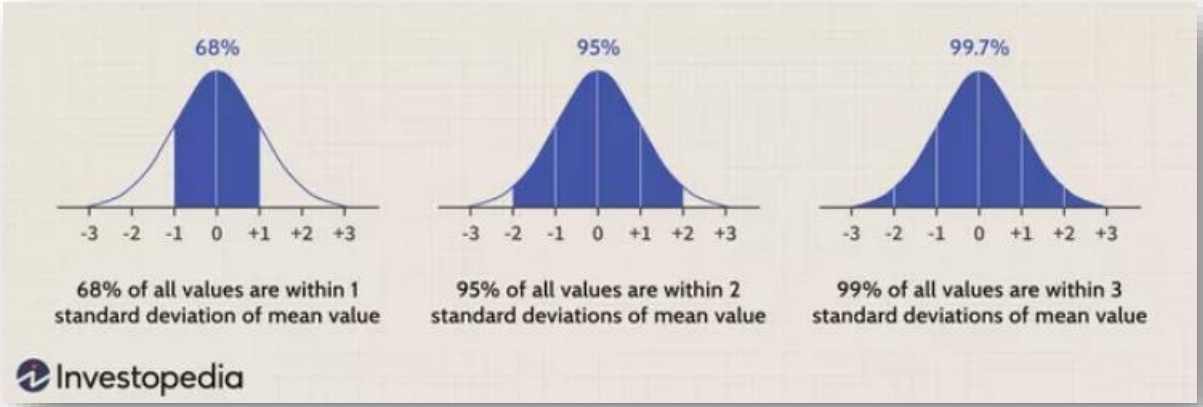
Kalau ditambahkan informasi, nilai batas lulus adalah 65, bagaimana?

Outliers



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

Central Limit Theorem

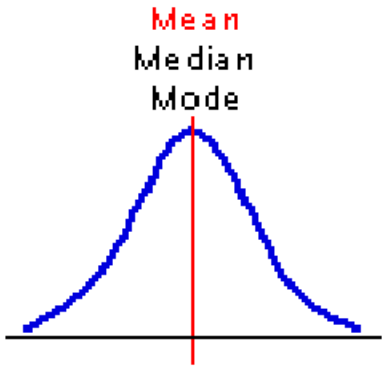


CLT = semakin besar sample, semakin mean dari sample mendekkati mean dari populasi terlepas dari distribusi data yang sebenarnya

data outliers adalah data yang jarak nilainya dengan rata-rata lebih besar dari 3 kali (+-) nilai standar deviasi

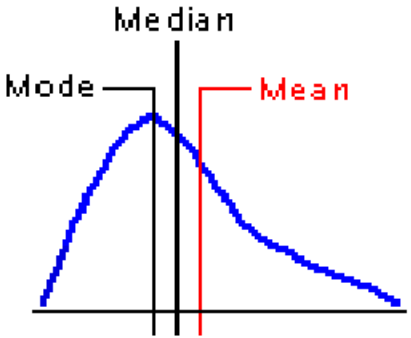
Asymmetric Distribution

Skewness



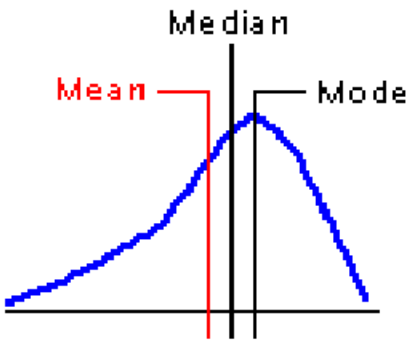
Symmetrical Distribution

Mean = Median = Modus



Positive Skew

Mean < Median < Modus



Negative Skew

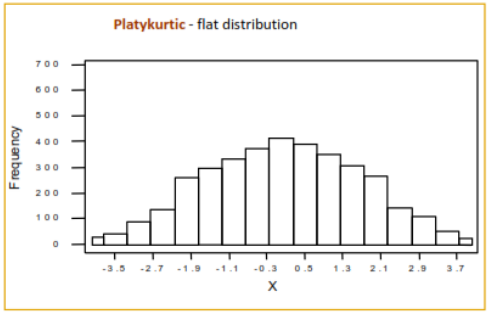
Mean > Median >Modus

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}}$$

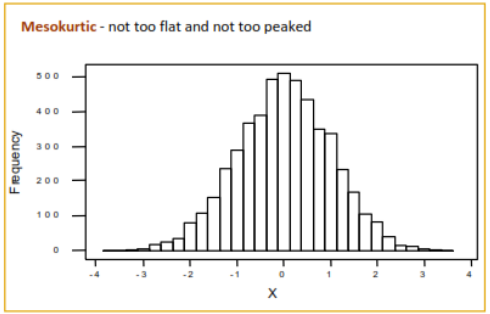
More than 1 highly skewed
0,5 – 1 moderate skewed
0 – 0,5 approcimately symmetric

Asymmetric Distribution

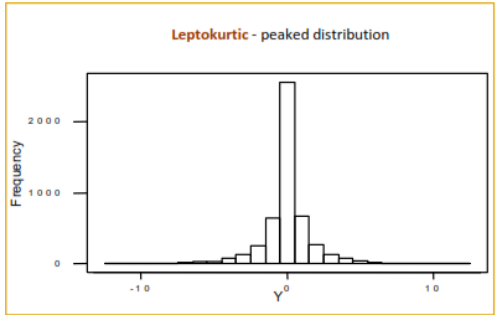
Kurtosis



Negative value



Around Zero



Positive Value

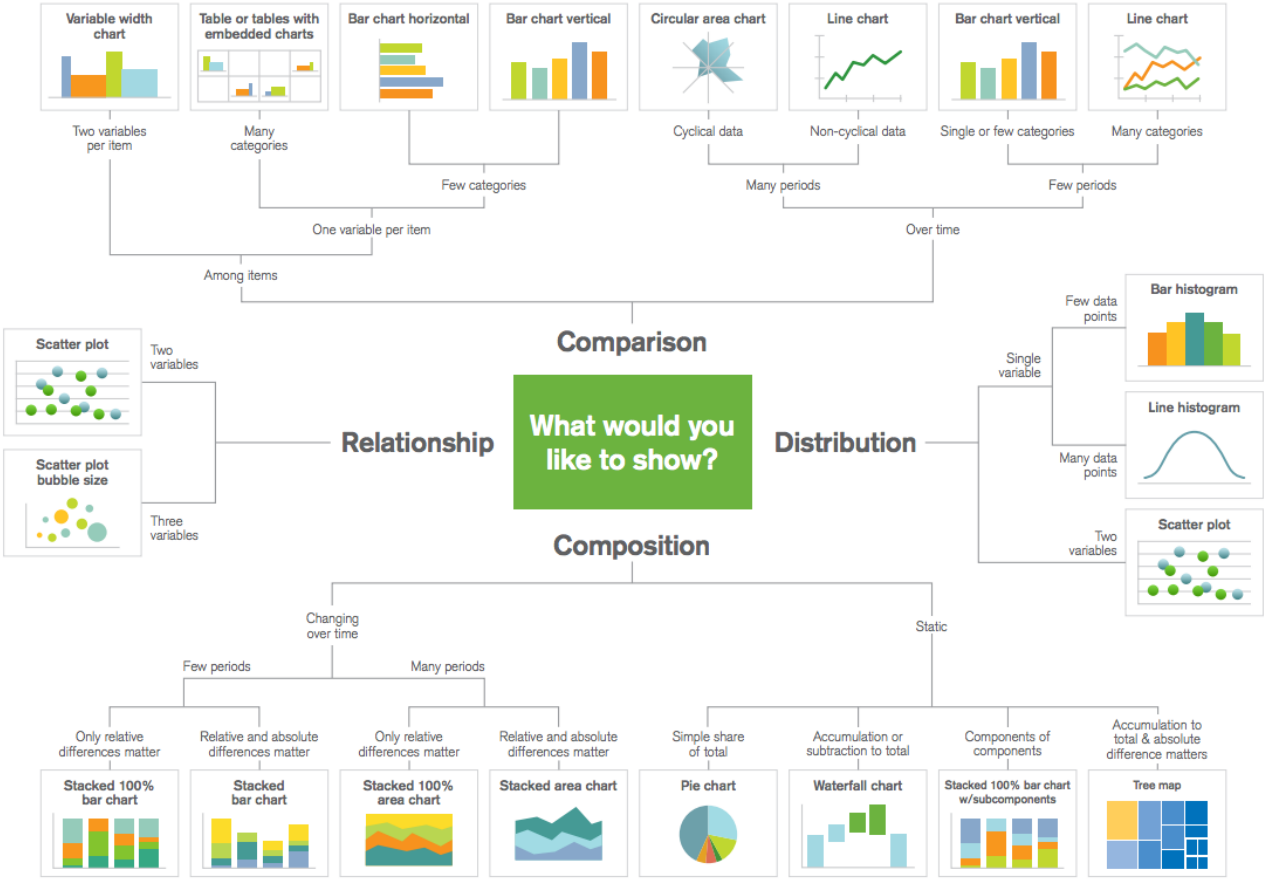
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}-3$$

a measure of the combined weight of the tails relative to the rest of the distribution.

03

Visualizing Data

Data Visualization

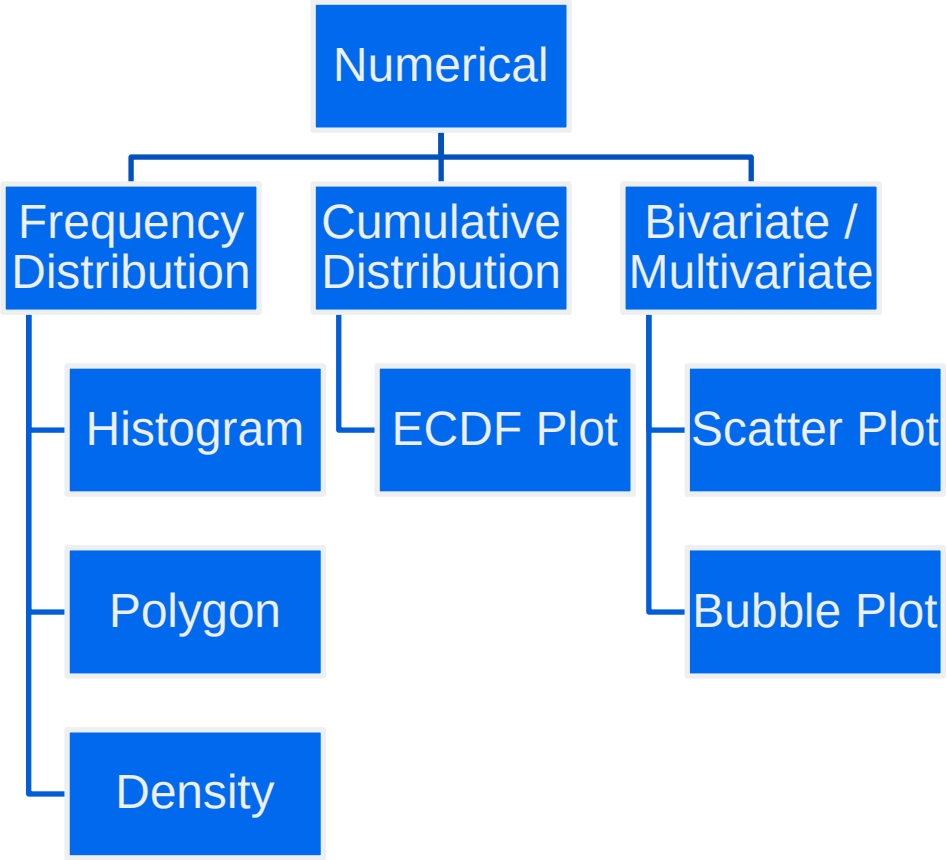


Referensi:

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
Frequency Distribution
Cumulative Distribution
Bivariate / Multivariate
Histogram
Polygon
Density
ECDF Plot
Scatter Plot
Bubble Plot

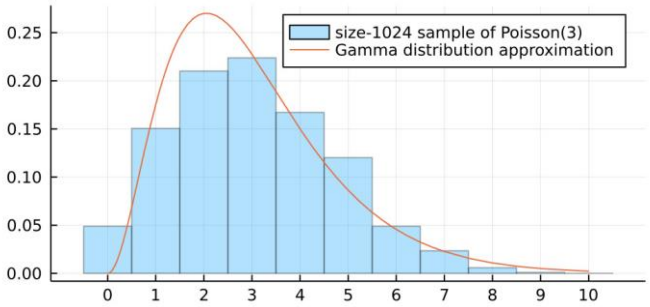
Numerical

Data in ordered array:

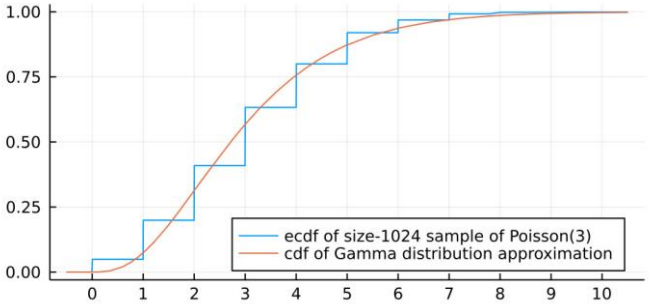
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

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

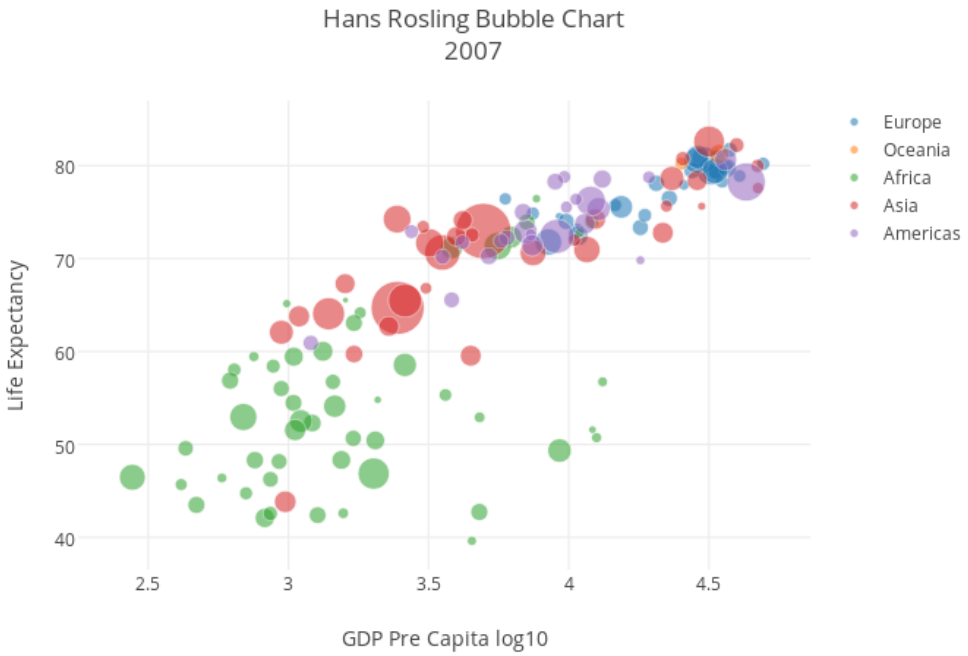


Histogram + KDE

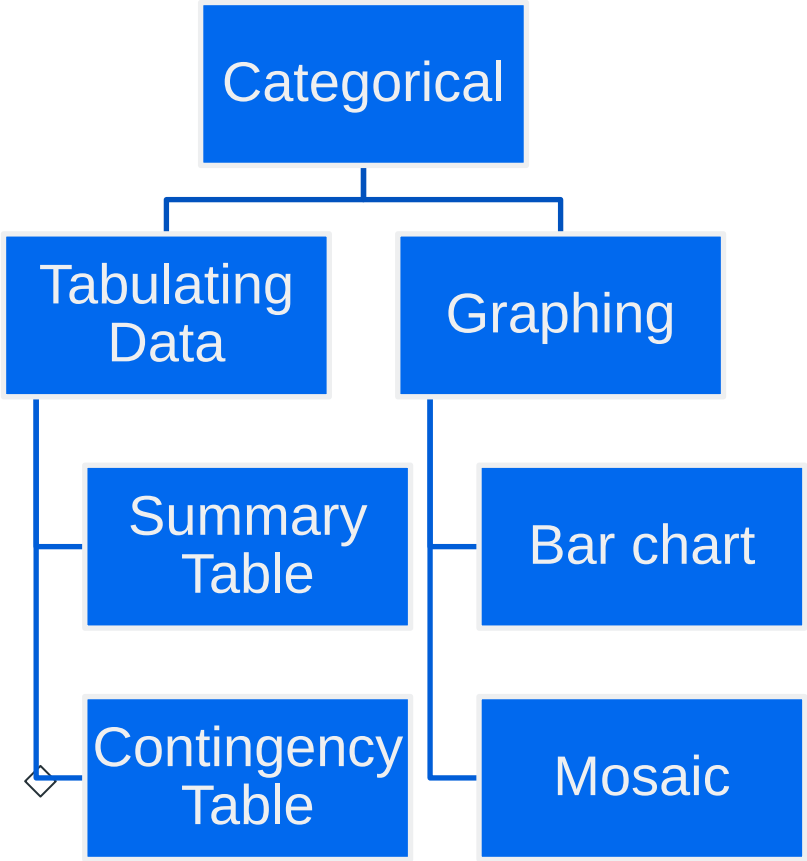


ECDF + CDF

Numerical


Hans Rosling Bubble Chart
2007
Life Expectancy
80
70
60
50
40
2.5
3
3.5
4
4.5
GDP Pre Capita log10
Europe
Oceania
Africa
Asia
Americas

Categorical



Categorical

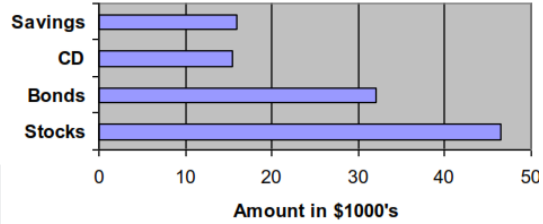
Univariate

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**

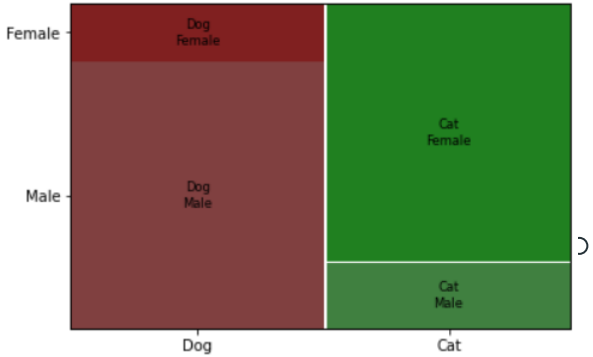
Bivariate

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

Summary Table



Contingency Table / Crosstab

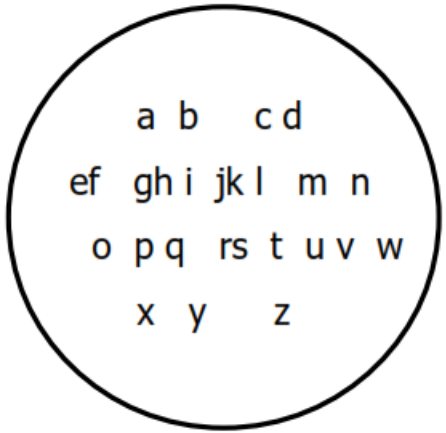


04

Sampling and Probability

Population & Sample

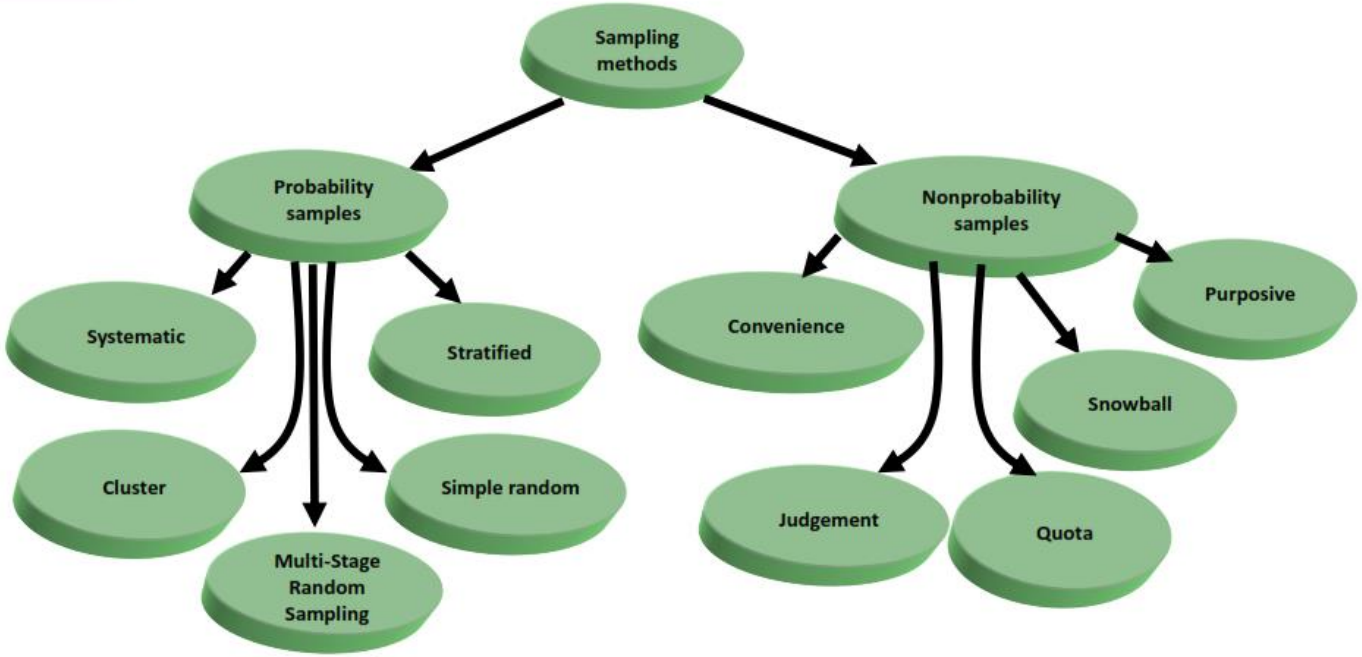
Population



VS

Sample

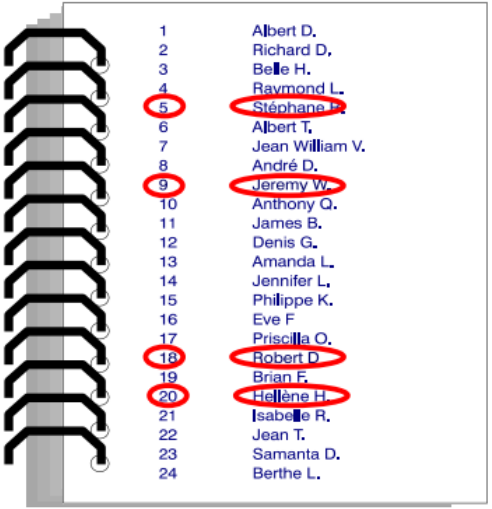
Sampling Method



Probabilistic

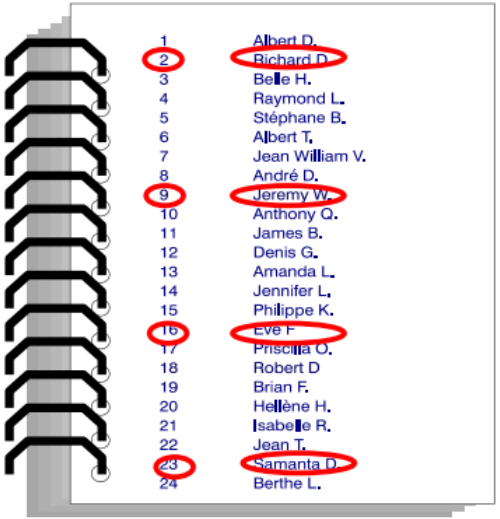
$$\frac{\sqrt{2}}{\left(\frac{1}{2}\right)2}$$

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.

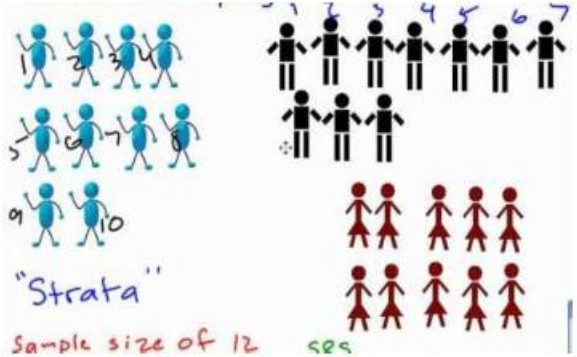
$$\frac{4+6+(2\sqrt{3})}{\sqrt{276}}$$

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

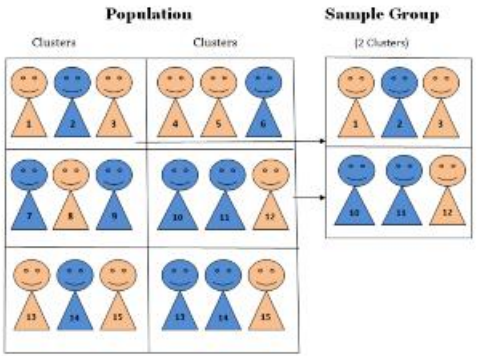
$$\frac{C - B}{3 - D}$$

Probabilistic

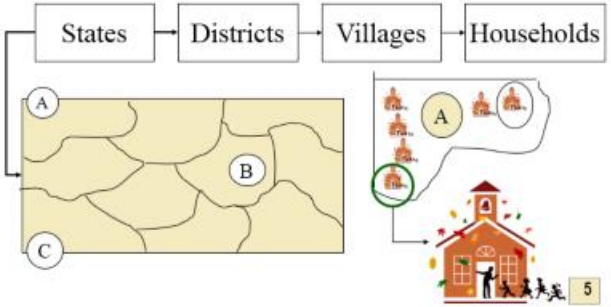
Stratified Sampling



Cluster Sampling



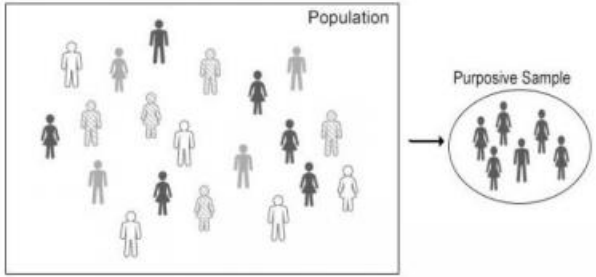
Multistage Sampling



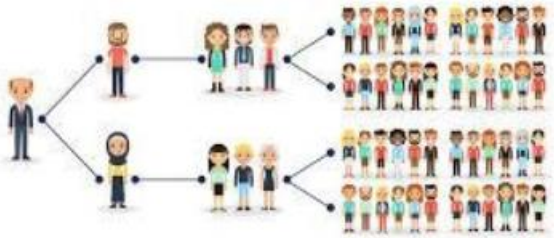
Non - Probabilistic

Convenience Sampling

Purposive Sampling



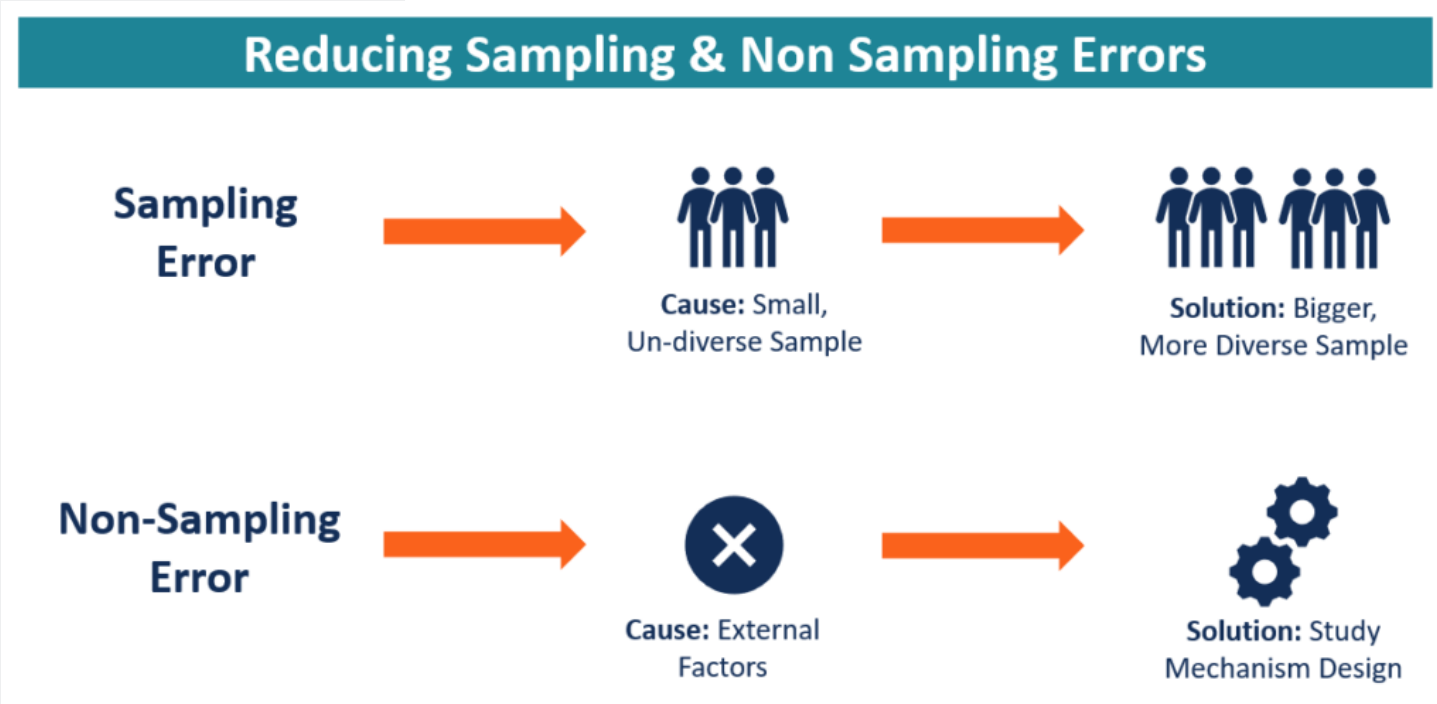
Snowball Sampling



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

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

Sampling Error



Peluang

Event

Hasil dari eksperimen

Contoh: Mendapatkan angka 6 dalam melempar satu buah dadu

Sample Space

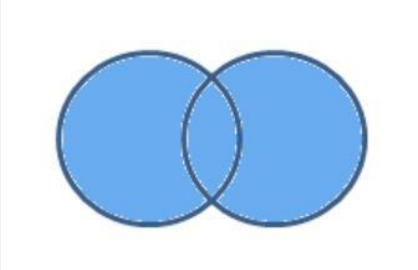
Kumpulan semua kemungkinan hasil eksperimen

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

Kombinasi antar *event*

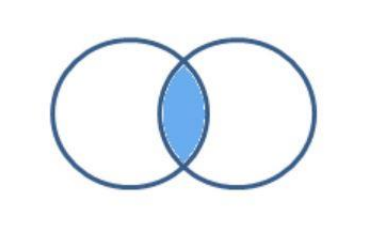
Gabungan

$P(A \cup B)$



Irisan

$P(A \cap B)$



Disjoint

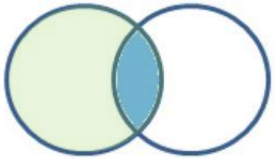
https://www.analyticsvidhya.com/blog/2017/03/conditional-probability-bayes-theorem/

Peluang Bersyarat

kemungkinan hasil yang terjadi, “**bersyarat**”/berdasarkan hasil sebelumnya yang terjadi

Contoh: Peluang alumni perguruan tinggi X Tahun 2021 yang bekerja

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



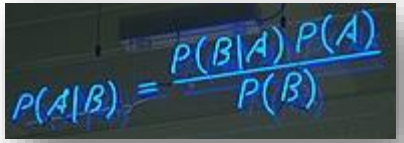
proporsi biru dari seluruh Hijau

- **Peluang Bersyarat merupakan dasar dari Teorema Bayes**

https://medium.com/scivviens/penjelasan-simpel-bayes-theorem-berserta-contoh-soal-69789b7303b4

https://www.rumusstatistik.com/2012/07/rumus-peluang-kejadian-bersyarat.html#:~:text=Peluang%20kejadian%20bersyarat%20adalah%20peluang%20kejadian%20bergantung%20kepada%20peluang%20kejadian%20

Bayesian Theorem

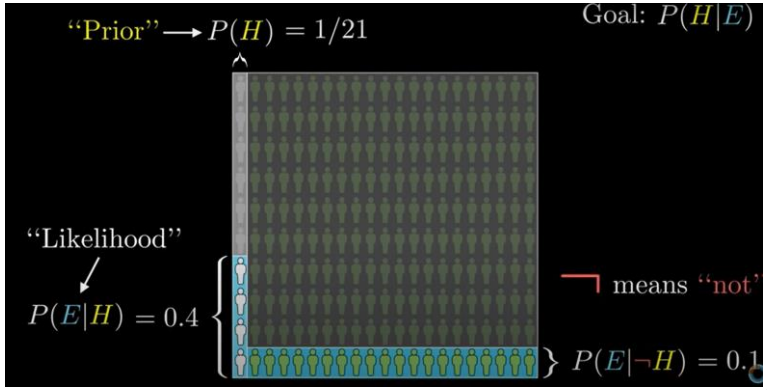


Teorema Bayes merupakan dasar dari Algoritma Naive Bayes pada *Machine Learning*

The geometry of changing beliefs

Secara general menggambarkan bagaimana manusia belajar atau dasar penelitian ilmiah.

Observasi / evidence baru P(B), tidak serta merta menggantikan ilmu / hypothesis yang sudah dipelajari P(A), tetapi mengupdate ilmu /hypothesis tersebut.



https://www.youtube.com/watch?v=HZGCoVF3YvM&t=50s

Pengaplikasian Teori Peluang

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

Gaming Mathematics

- Dice
- Cards

Optimalization

- Machine Learning
- Artificial Intelligent
- Operational Search

Stochastic Process

- Markov Chain
- Renewal Theory (Hypothesis Testing)

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

05

Inference Statistic

Summary of Inference Stats

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

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X2 or binomial	X2	Macnarmar's X2	X2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r

$\sqrt{2.76}$

Korelasi

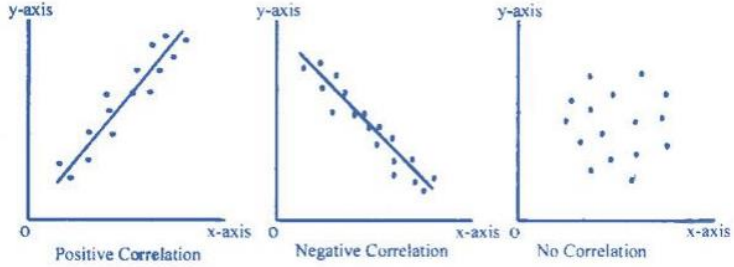
- Digunakan untuk menemukan hubungan antara dua variabel kuantitatif
- **Kausalitas:** variabel X menyebabkan perubahan pada variabel Y
- Memiliki rentang nilai antara -1 hingga 1

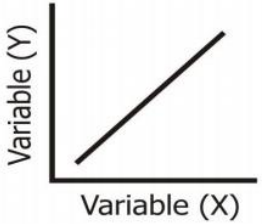
Note:

- **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**

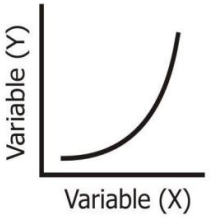
Korelasi

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**

https://www.analyticsvidhya.com/blog/2021/04/intuition-behind-correlation-definition-and-its-types/#:~:text=What%20is%20Correlation%3F,two%20variables%20are%20involved%20here.

Uji Korelasi

Numerical Correlation :
It's a measure of the strength and the direction of a linearrelationship 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 kategorical
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_0 (Null Hypothesis)** = tidak ada hal baru yang terjadi pada populasi
- **H_1 (Alternative Hypothesis)** = negasi dari H_0

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

https://www.greelane.com/id/sains-teknologi-matematika/matematika/fail-to-reject-in-a-hypothesis-test-3126424/#:~:text=Gagal%20menolak%20hipotesis%20nol%20(artinya,hubungan%20konsekuensial%20antara%20kedua%20fenomena)

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:
- $\mu = 0$, $\mu > xx$, $\mu <= xx$

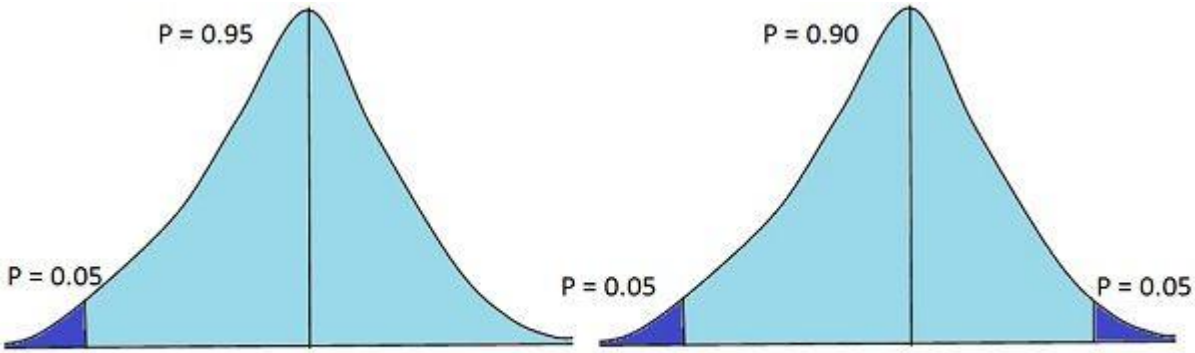
Uji t dua sampel

- 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, significancy tercapai, H0 ditolak
Biru muda, significancy tidak tercapai H0 gagal ditolak

Tipe Error

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

- H0 **benar, ditolak = Error Tipe I (Alpha/False Positif)**
- H0 **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

https://www.kdnuggets.com/2021/09/hypothesis-testing-explained.html

“Statistics is The Grammar of Science”

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**