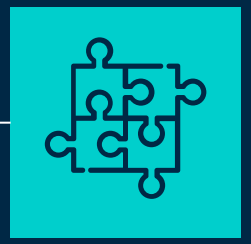
LEARNING PROGRESS REVIEW

Week 8

Diaz Jubairy - Hermulia Hadie
Desi Sulistyowati - Farahul Jannah

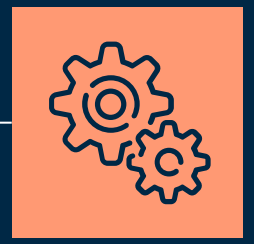
Table of Cotents



01

Basic Statistics

- Pentingnya Statistik
- Korelasi dan sebab akibat
- Populasi, Sample, Parameter
- Tipe Data Statistik
- Frekuensi dan Histogram
- Descriptive Statistics



02

Intermediate Stattistics

- Probabilitas
- Korelasi dan sebab akibat
- Statistical Plot

03

Advanced Statistics

- Teknik Sampling
- Hypothesis Testing
- Uji dalam Hipotesis
- Z-Test
- T-Test
- Chi Square Test

BASIC STATISTICS

01

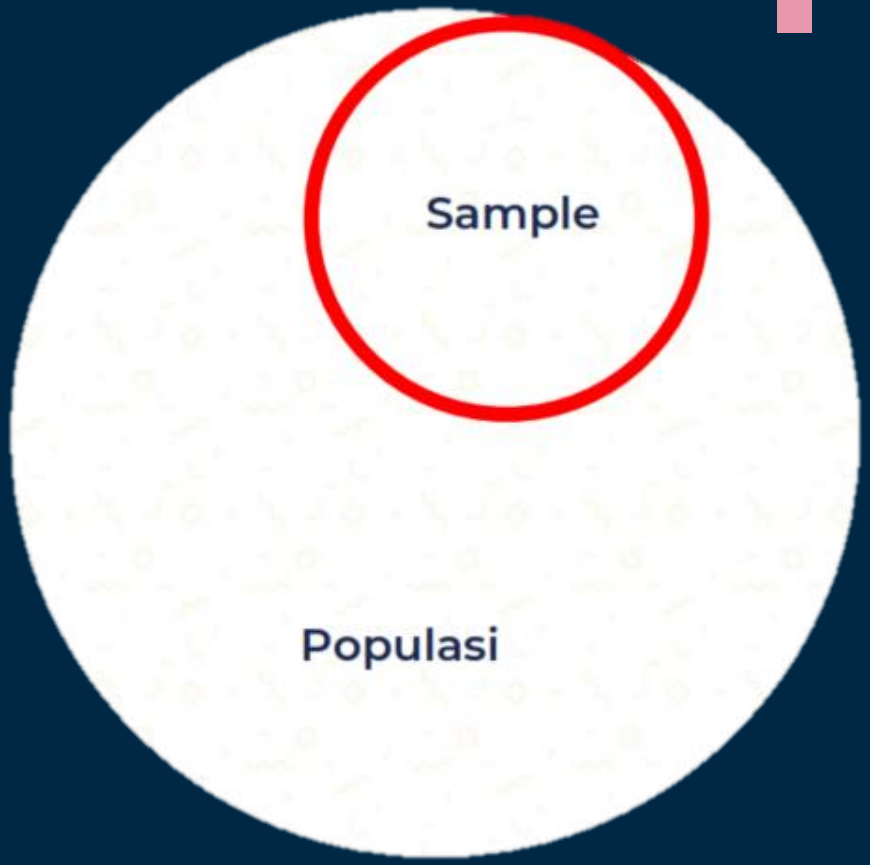
Measure of Central Tendency
Measure of Spread

Pentingnya Statistik

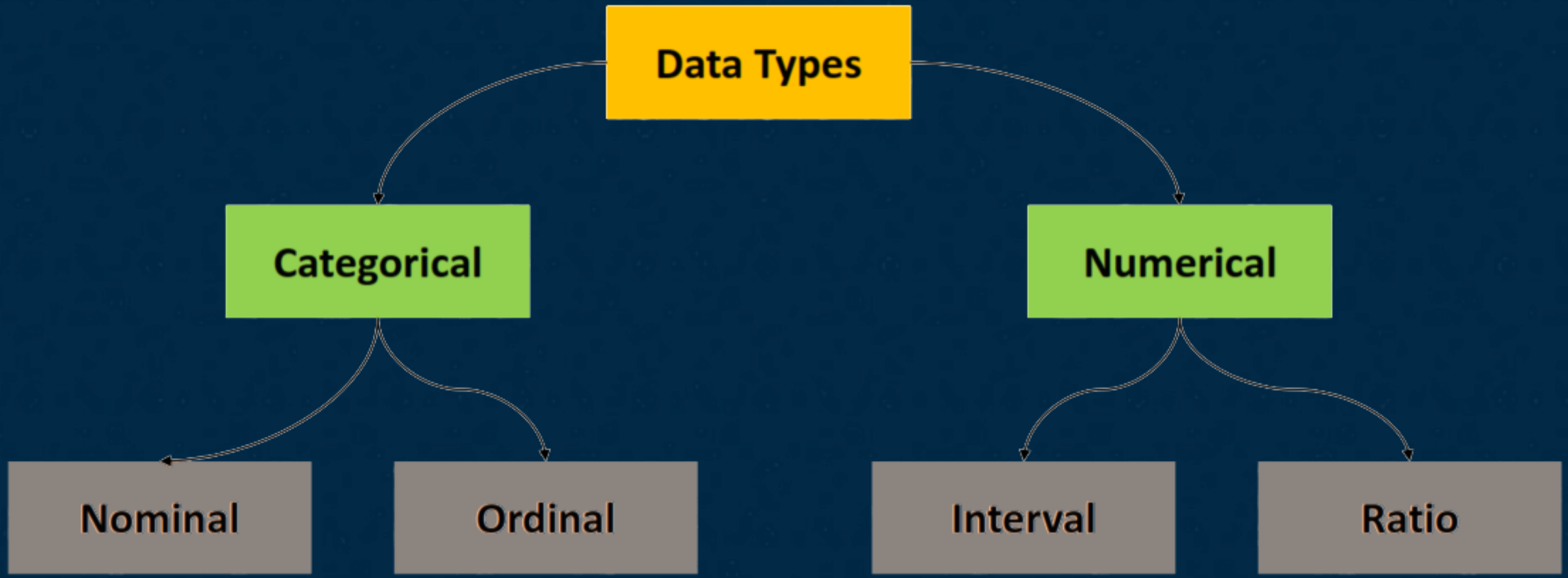
- *Data is everywhere*
- Interpretasi data dalam jumlah besar
- Kebutuhan *Decision Making* yang sangat cepat
- Membantu menilai kebenaran/kewajaran informasi

Populasi, Sample, Parameter, Statistics

- Statistik erat kaitannya dengan sample
- **Sample** adalah bagian populasi.
- **Populasi** menggambarkan keseluruhan anggota elemen.
- Pada sample yang kita observasi, kita dapat mengukur *property dari sample tersebut.* Property ini disebut **Statistics** dari sample.
- Terkadang kita ingin mendapatkan statistics untuk keseluruhan Populasi. Hal ini disebut sebagai **Parameter**.
- Angka statistics digunakan sebagai estimate terhadap parameter, yang terkadang tidak diketahui nilainya.



Tipe Data dalam Statistics



Tipe Data dalam Statistics

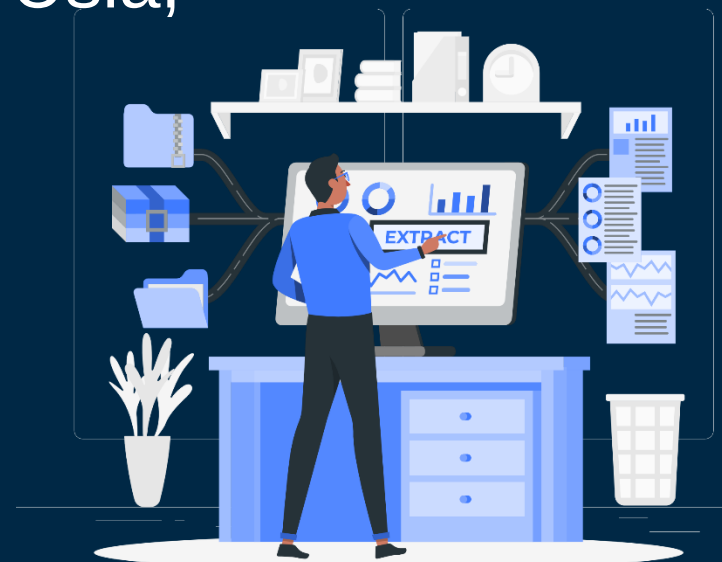
- Tipe data nominal adalah data yang data direpresentasikan sebagai kumpulan event atau obyek dalam kategori yang bersifat diskrit. Contoh:
  - Nama TV show di sebuah channel televise
  - Jenis makanan di sebuah restoran
  - Kumpulan nama bank di Indonesia

- Tipe data ordinal mirip dengan data nominal, namun perbedaannya adalah data tersebut dapat diurutkan. Contoh:
  - Tingkat kepuasan terhadap pelayanan : Bad, Fair, Good, Excellent
  - Tingkat Pendidikan : SD, SMP, SMA, Sarjana

Tipe Data dalam Statistics

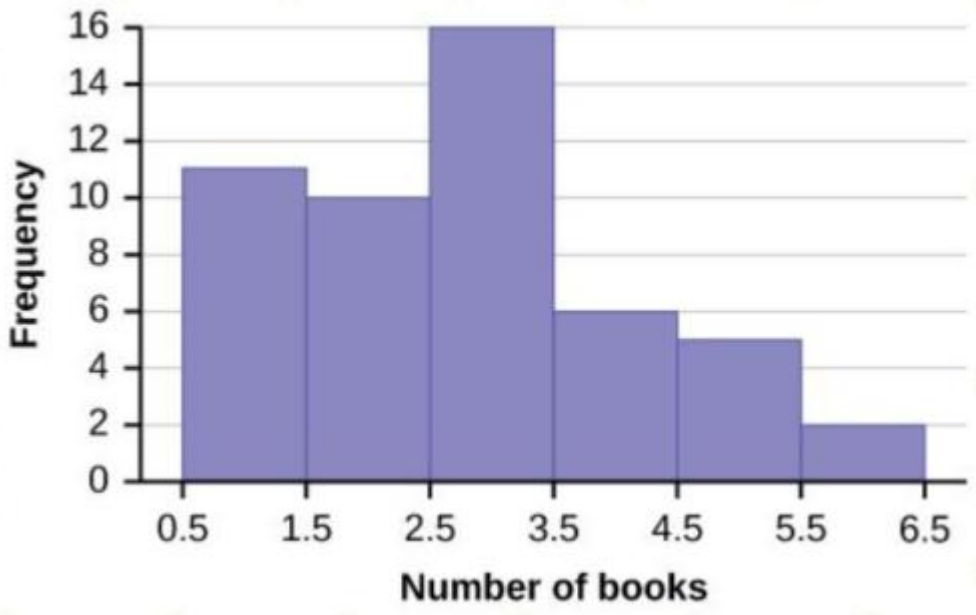
- Tipe data interval adalah data numerik yang tidak memiliki nilai true zero, artinya nilai bisa mencapai di bawah zero (negative). Contoh: Temperature

- Tipe data ratio mirip dengan data interval, namun data ini memiliki angka true zero. Contoh: Usia, Income dalam rupiah, Tinggi badan



Frekuensi dan Histogram

- Frekuensi dan histogram menggambarkan distribusi summary statistic yang melakukan grouping elemen dalam nilai yang sama
- Histogram membantu kita dalam melihat distribusi nilai secara summary dengan lebih mudah dibandingkan melihat dalam bentuk tabel



Nilai ujian	Jumlah siswa
10	5
20	3
30	2
40	1
50	5
60	3
70	2

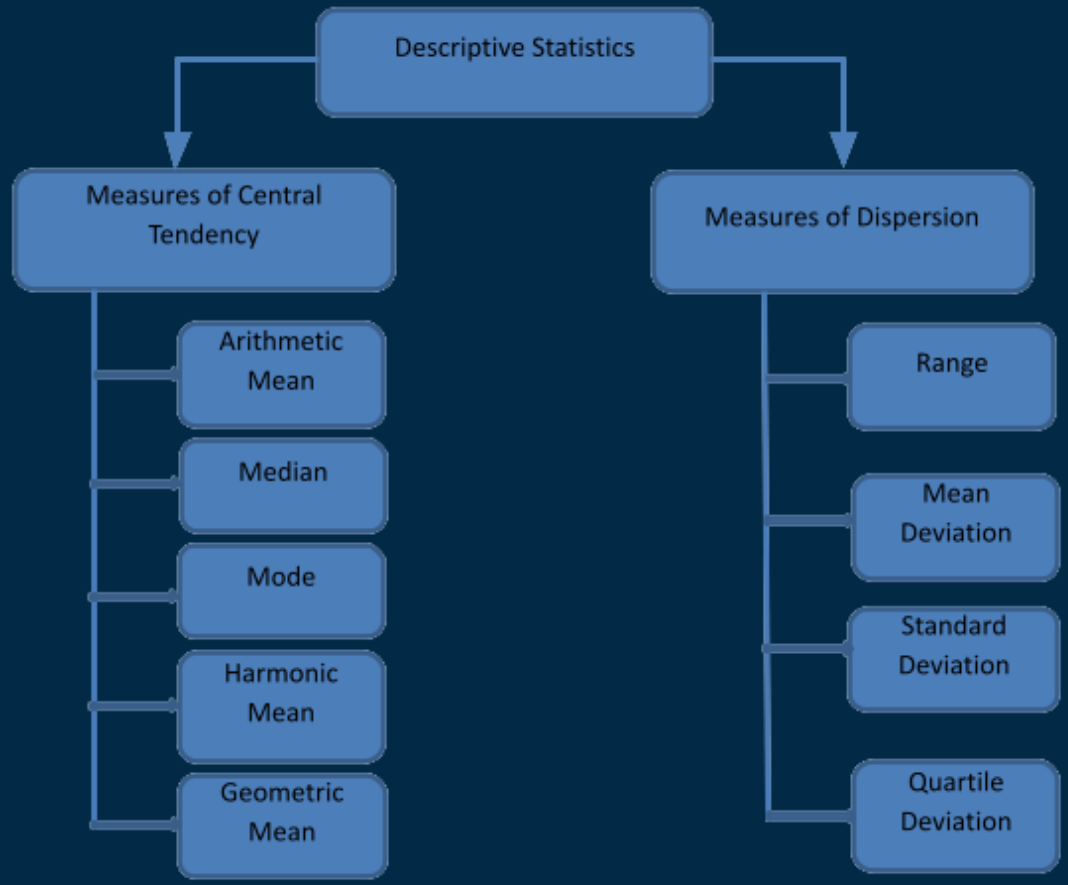
Pembagian Domain Statistics

- Descriptive statistics: Merupakan sebuah metode untuk mendeskripsikan dan menampilkan data secara summary dalam bentuk visual.
- Inferential statistics: Dalam study statistic, seringkali cukup sulit untuk menentukan parameter dari sebuah populasi. Inferential statistics adalah study untuk melakukan generalisasi terhadap suatu sample yang mewakili sebuah populasi

Descriptive statistics

Measures of Spread (Pengukuran variabilitas dalam data):
- Range
- Min-max
- Quartiles/percentiles
- Standard deviation/variance

Measures of Central Tendency (Pengukuran terhadap tendency nilai tengah):
- Mean
- Mode
- Median



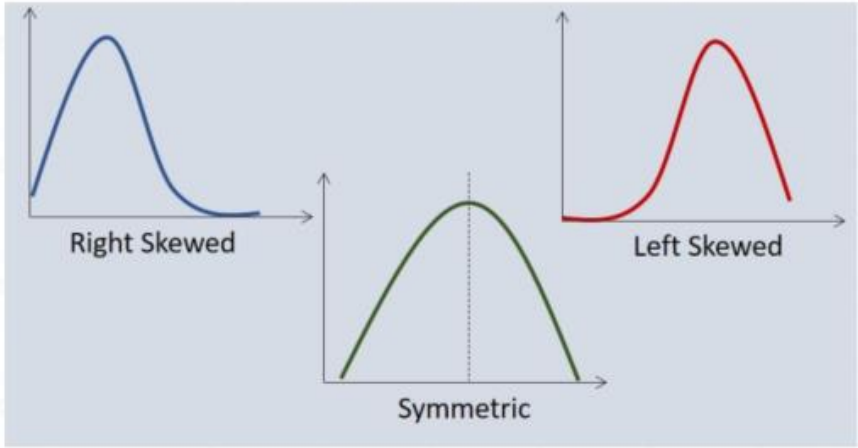
Descriptive Statistics: Mean

- Metric yang paling popular digunakan dalam descriptive statistics
- Mean bisa memberikan informasi yang *misleading*
- Mean sensitive terhadap outlier

Siswa	1	2	3	4	5	6	7	8	9
Usia	15	18	16	14	15	12	17	60	70

Mean = 26.33 (reasonable kah?)

- Mean juga sensitive ketika data yang diberikan skewed, atau cenderung tidak memiliki distribusi normal.
- Semakin skewed, mean bisa jadi kehilangan kemampuan untuk memberikan gambaran nilai tengah dari suatu data



Descriptive Statistics: Median

- Median dapat membantu menyelesaikan isu representasi data nilai tengah dengan Mean ketika terdapat outlier.
- Median diukur dengan cara :
  - Urutkan elemen numeric dari urutan terkecil s/d terbesar
  - Tentukan banyaknya elemen dalam data tersebut, misal ukuran datanya sebesar n
  - Apabila N ganjil, maka median adalah :

$$\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{observation}$$

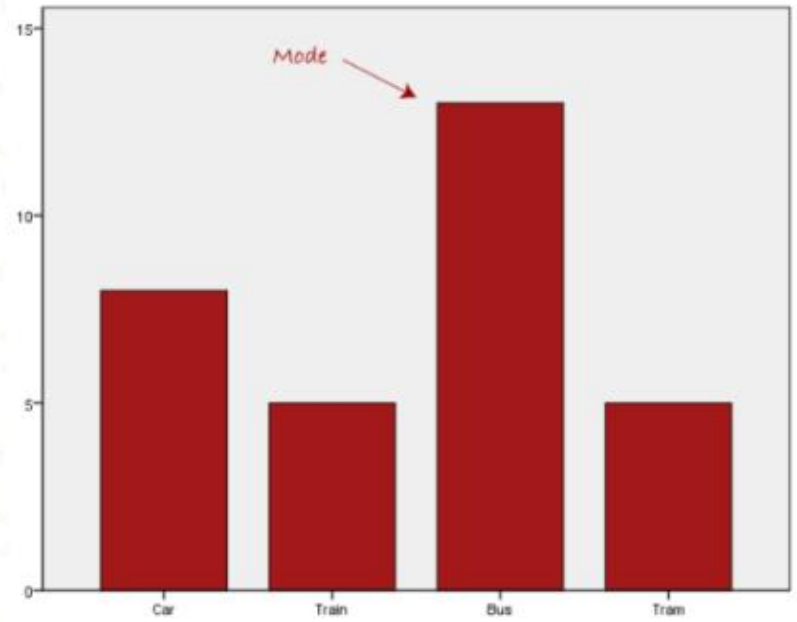
  - Apabila N genap, maka median adalah :

$$\text{Median} = \frac{\frac{n^{\text{th}}}{2}\text{obs.} + \left(\frac{n}{2}+1\right)^{\text{th}}\text{obs.}}{2}$$

Descriptive Statistics: Mode

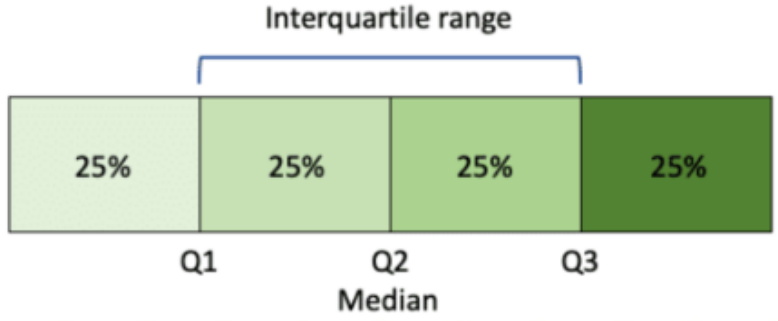
- Mode adalah elemen yang memiliki frequency terbanyak dalam suatu data numeric
- Apabila mean dan median tidak bisa digunakan dalam data yang berbentuk categorical, Mode bisa digunakan untuk data categorical.

Jenis mobil	Jumlah pemilik	
BMW	20	
Mercedes Benz	10	
Honda	**40**	mode
Toyota	5	



Quartiles, Percentiles, Deciles, Interquartile Range

- Mean dan median memberikan nilai tengah, atau nilai paling banyak dalam sebuah data.
- Seringkali, kita membutuhkan informasi yang lebih banyak, seperti, nilai 10% tertinggi dari data, nilai 70% tertinggi dari data, dsb.
- Untuk pertanyaan-pertanyaan tersebut, kita bisa jawab dengan quartiles, percentiles, atau deciles.
- Interquartile range memberikan gambaran mengenai kumpulan nilai tengah (middle fifty) dari sebuah data.
- Penggunaan range Min-Max terkadang bisa terpengaruh oleh outlier
- IQR cenderung lebih tidak sensitive terhadap outlier dibandingkan metric Range lain seperti Min-Max



Descriptive Statistics: Standard Deviation

- Bagaimana apabila kita ingin mengetahui sebaran dari data? Kita bisa menggunakan variance/standard deviation.
- Secara matematis :

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

$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\sum (x_r - \mu)^2}{n}}$$

Contoh ilustrasi

Nilai	(x-mean)	(x-mean)^2
10	-17.8	316.84
15	-12.8	163.84
30	2.2	4.84
43	15.2	231.04
32	4.2	17.64
10	-17.8	316.84
14	-13.8	190.44
24	-3.8	14.44
56	28.2	795.24
44	16.2	262.44
	Jumlah	2313.6
	N	10
	Variance	231.36
	Stdev	15.21052267

Standard Deviation	15.21052267
Mean	27.8

INTERMEDIATE STATISTICS

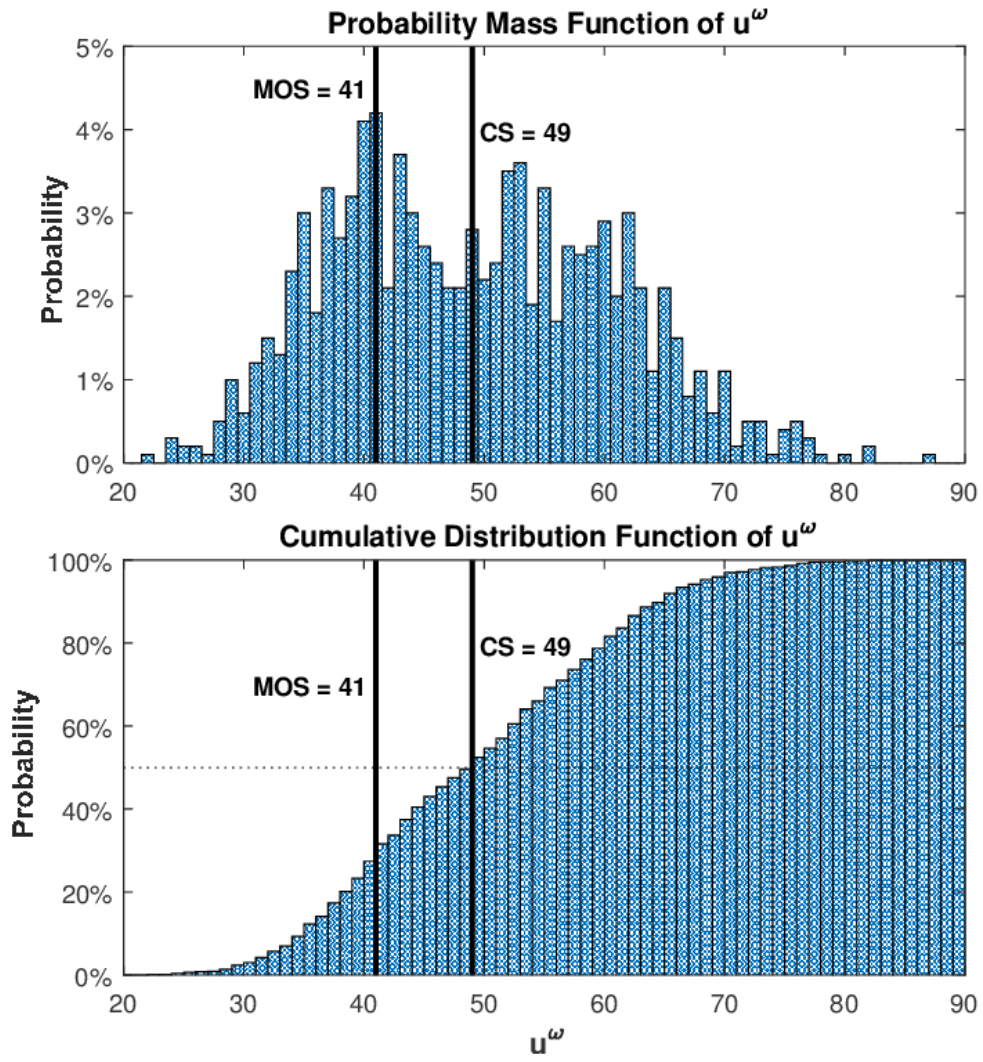
02

Probabilitas
Korelasi dan sebab akibat
Statistical Plot

Probabilitas

mengukur seberapa besar kemungkinan suatu peristiwa terjadi pada skala 0 (tidak pernah terjadi) hingga 1 (selalu terjadi). Probability pada Data Science dibagi menjadi dua :

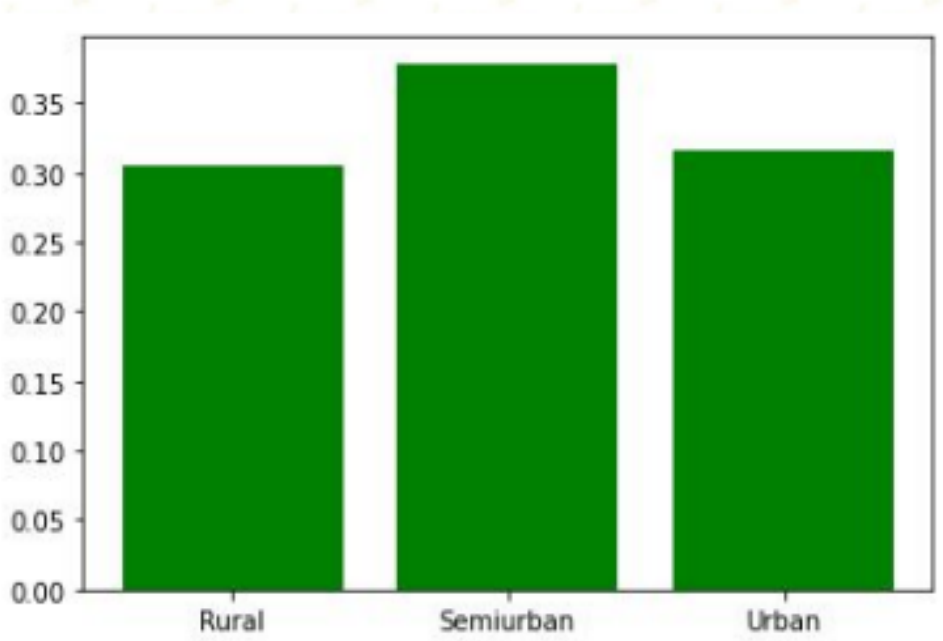
- Probability Mass Function (PMF)
- Cumulative Distribution Function (CDF)



Probability Mass Function

- adalah frekuensi yang dinyatakan sebagai sebuah pecahan dari suatu sampel. Untuk mendapatkan probabilitas, kita membagi datanya sebanyak n, disebut normalisasi.

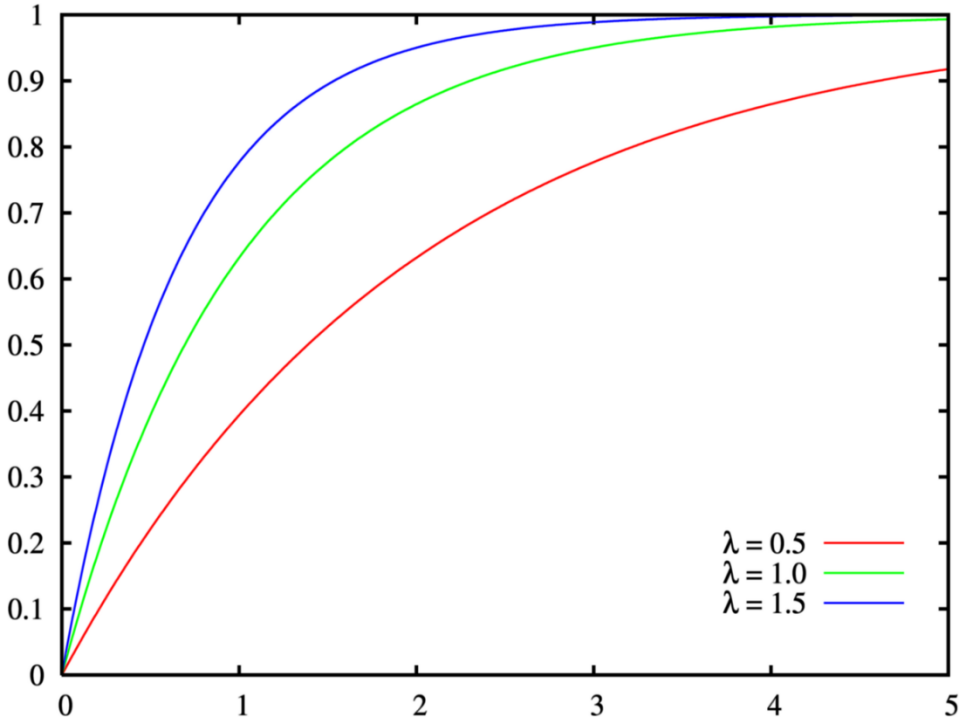
- *Probability Mass Function* hanya digunakan untuk variabel *discrete* atau *categorical*. Sebagai contoh seperti melempar dadu, kasus ini mengikuti *discrete distribution* karena tidak ada nilai yang di tengah-tengah (dadu bermata 6 hanya memiliki kemungkinan angka 1,2,3,4,5,6, tidak mungkin kita mendapatkan angka 1.5).



Cumulative Distribution Function

- PMF bekerja dengan baik jika memiliki variasi data yang sedikit. Jika variasi data meningkat, probabilitas yang terkait dengan setiap nilai yang kecil akan mengalami peningkatan random noise

- Cumulative Distribution Function adalah fungsi yang memberikan penaksiran probabilitas kumulatif dari variabel acak diskrit atau kontinyu hingga nilai tertentu



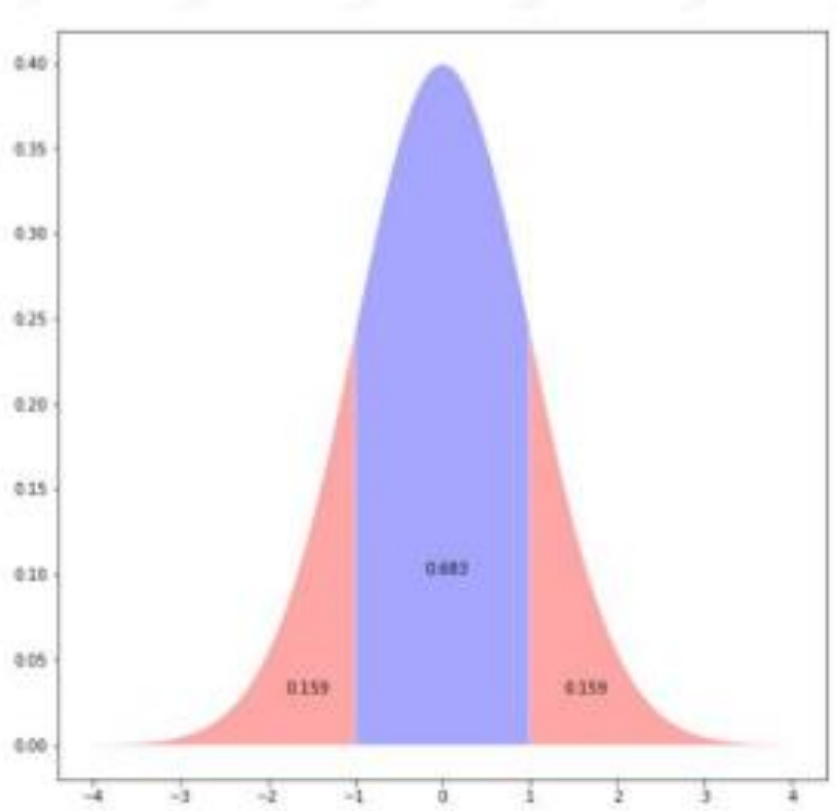
Pembagian Domain Statistics

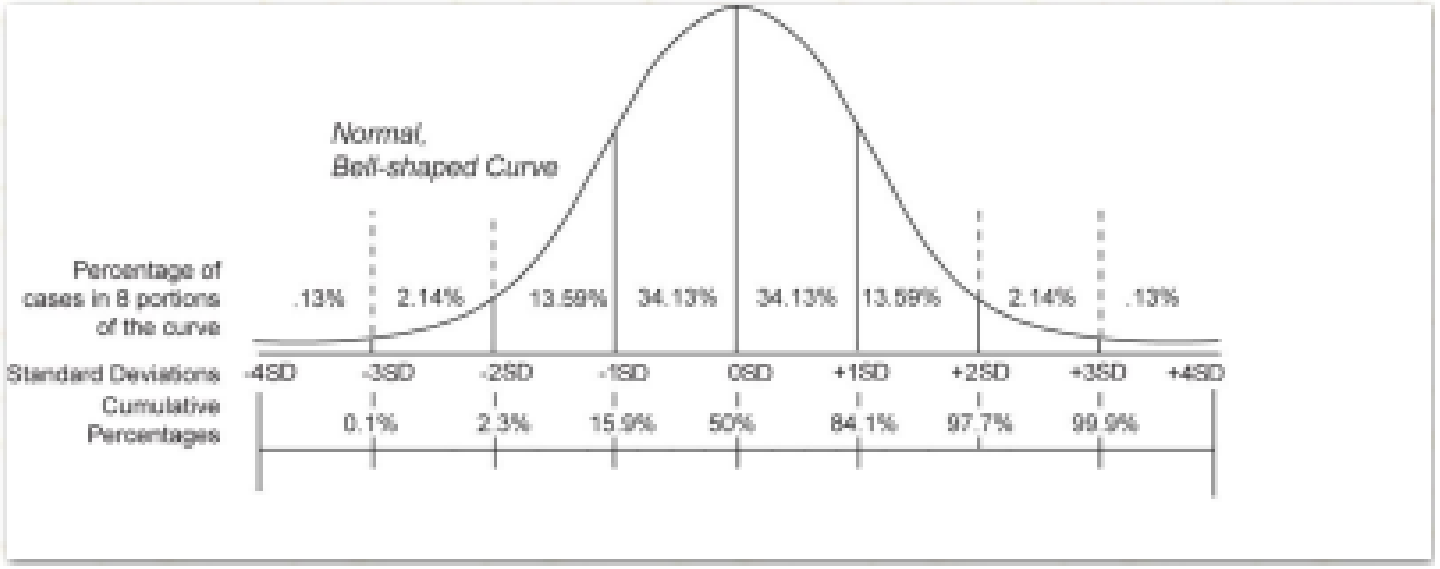
- Distribusi dideskripsikan bagaimana suatu variabel tersebar secara spesifik yang nilai mana yang paling mungkin diambil

- Distribusi normal adalah distribusi probabilitas kontinu yang dicirikan oleh kurva berbentuk lonceng simetris (bell-shaped curve). Terlebih lagi, distribusi normal memiliki karakter dengan di pusatnya sebagai nilai rata-rata dan penyebarannya standar deviasi

- Karakteristik distribusi normal :
  - Simetris jika dibagi dua dari pusatnya
  - Rata-rata dan median hampir sama
  - Setengah dari nilainya sama dengan yang kanan dan kiri

Normal Distribution

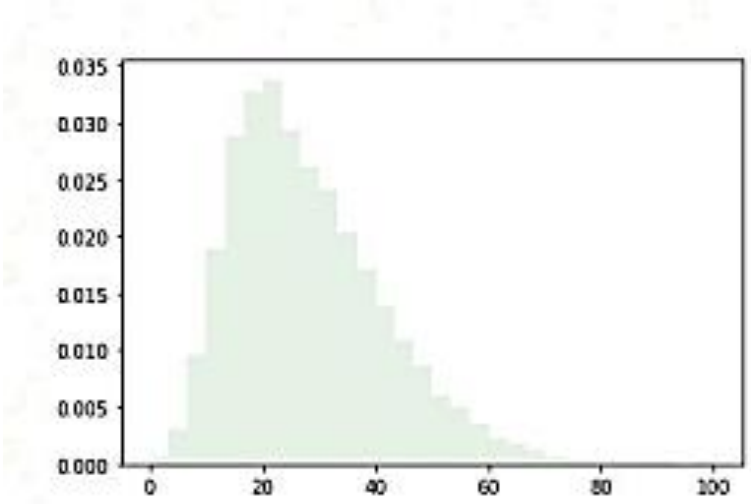


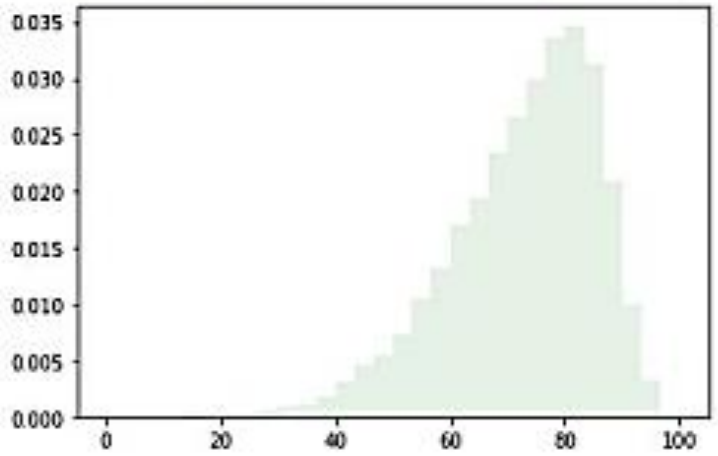


Skewness

Skewness adalah hal yang menggambarkan bentuk dari distribusi. Ada dua hal yang perlu diketahui:

- Right Skewness: ketika nilai memanjang jauh ke kanan
- Left Skewness: ketika nilai memanjang jauh ke kiri





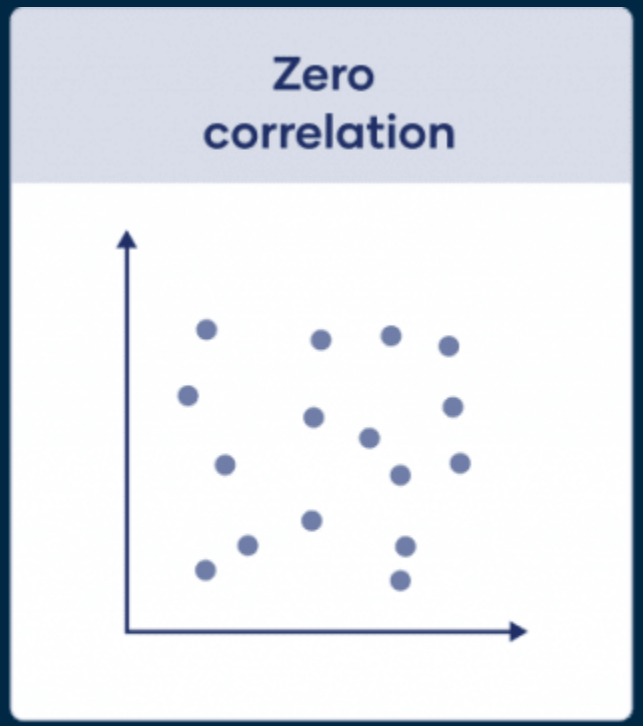
Korelasi dan Sebab Akibat

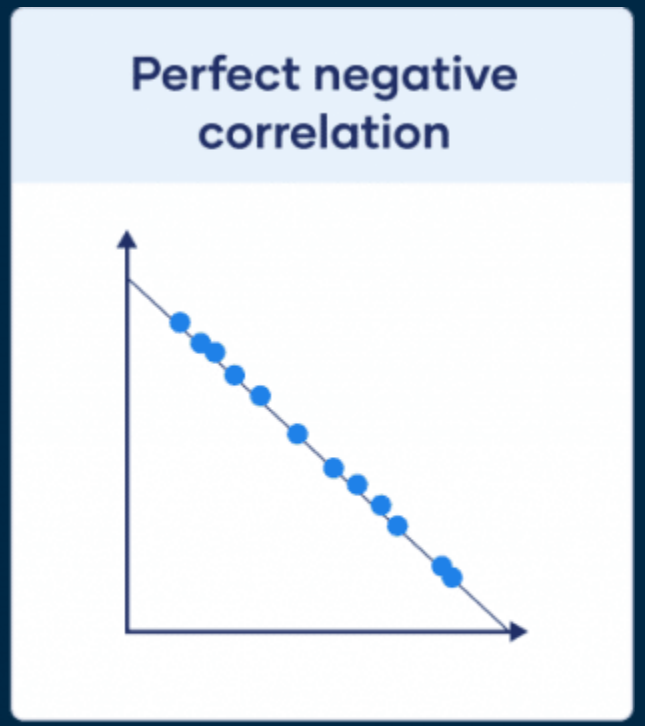
- Korelasi merupakan istilah yang biasa digunakan untuk menggambarkan ada tidaknya hubungan suatu hal dengan hal lain. Analisis korelasi adalah suatu cara atau metode untuk mengetahui ada atau tidaknya hubungan linear antar variabel.

- Apabila terdapat hubungan maka perubahan-perubahan yang terjadi pada salah satu variabel X akan mengakibatkan terjadinya perubahan pada variabel lainnya (Y). Istilah tersebut dikatakan istilah sebab akibat, dan istilah tersebut menjadi ciri khas dari analisis korelasi.

Contoh Kasus Korelasi

- Hubungan antara kenaikan harga BBM (X) dengan harga kebutuhan pokok (Y)
- Hubungan tingkat pendidikan (X) dengan tingkat pendapatan (Y)
- Hubungan umur pernikahan pertama (X) dengan jumlah anak yang dilahirkan (Y)
- Hubungan tingkat pendidikan ibu (X) dengan tingkat kesehatan/tingkat gizi bayi (Y), dsb.





Korelasi Positif

Korelasi positif atinya suatu hubungan antara variabel X dan Y yang ditunjukan dengan hubungan sebab akibat dimana apabila terjadi penambahan nilai pada variabel X maka akan diikuti terjadinya penambahan nilai variabel Y.

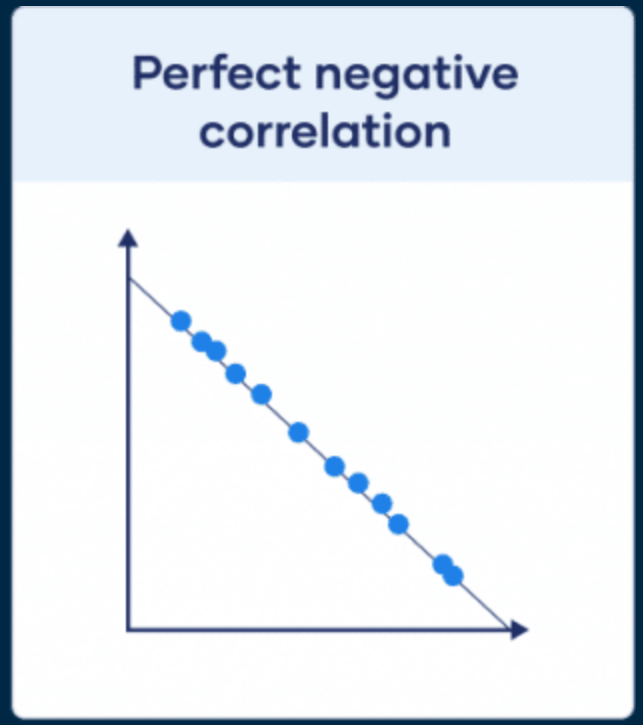
Contoh Korelasi Positif :

- ✓ Dalam pertanian, jika dilakukan penambahan pupuk (X), maka produksi padi akan meningkat (Y)
- ✓ Tentu saja semakain tinggi badan (X) seorang anak maka, berat badannya akab bertambah pula(Y)

Korelasi Negatif

jika pada korelasi positif peningkatan nilai X akan diikuti penambahan nilai Y, korelasi negatif berlaku sebaliknya. Jika nilai variabel X meningkat nilai variabel Y justru mengalami penurunan.

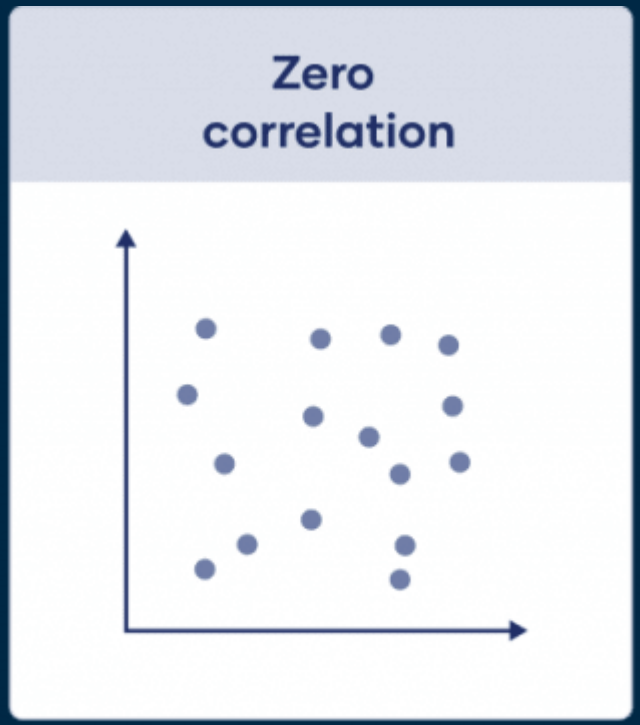
Contoh Korelasi Negatif
Apabila harga barang (X) meningkat maka kemungkinan permintaan terhadap barang tersebut mengalami penurunan.



Tidak ada Korelasi atau Sangat Lemah

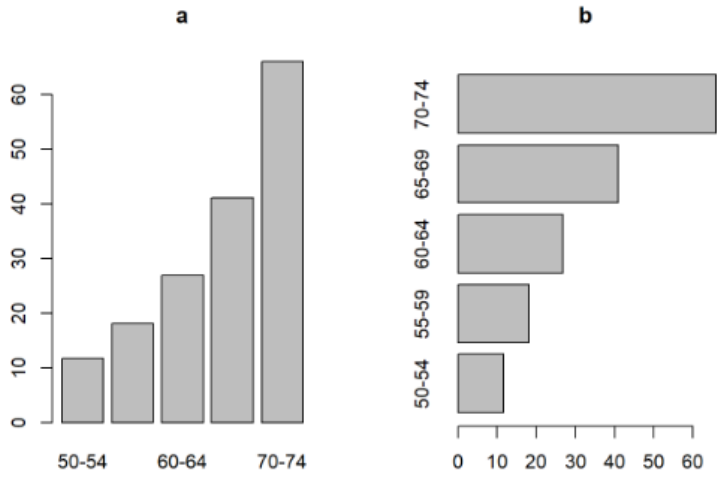
Korelasi ini terjadi apabila kedua variabel (X dan Y) tidak menunjukkan adanya hubungan linear.

Contoh :
Panjang rambut (X) dengan tinggi badan (tidak bisa dihitung hubungannya atau tidak ada hubungannya)

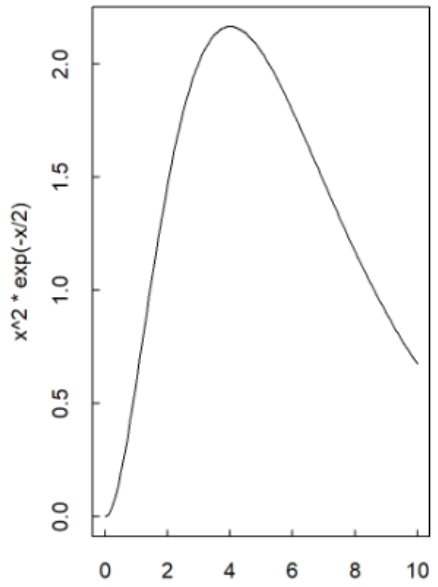


Statistical Plot

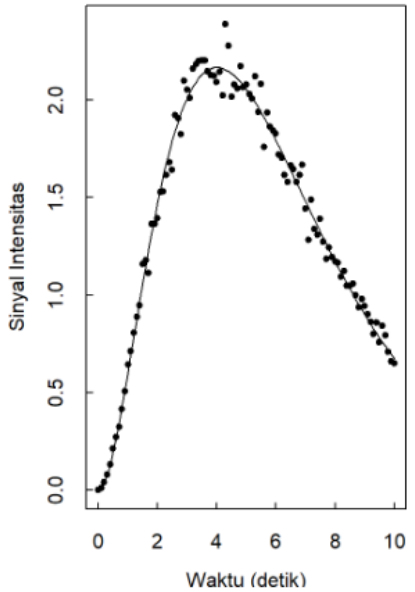
Penyajian suatu data secara statistik dapat dipermudah dengan menampilkan data tersebut secara visualisasi statistik



Bar Plot



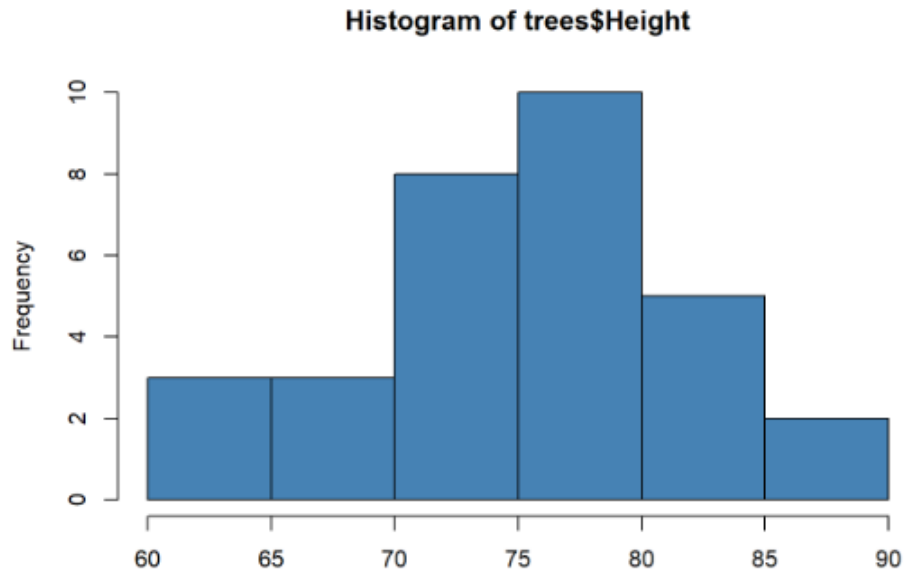
Fungsi Curve



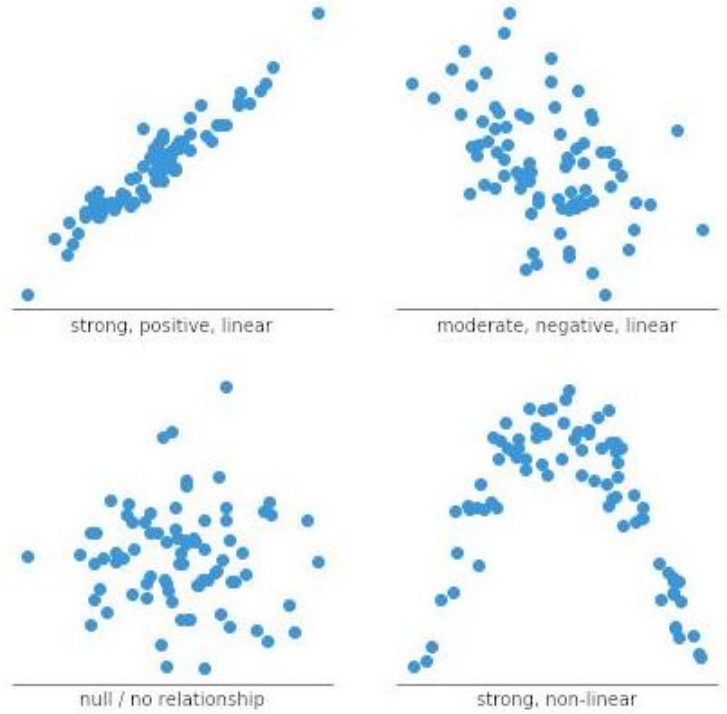
Fungsi Plot & Curve

Statistical Plot

Penyajian suatu data secara statistik dapat dipermudah dengan menampilkan data tersebut secara visualisasi statistik



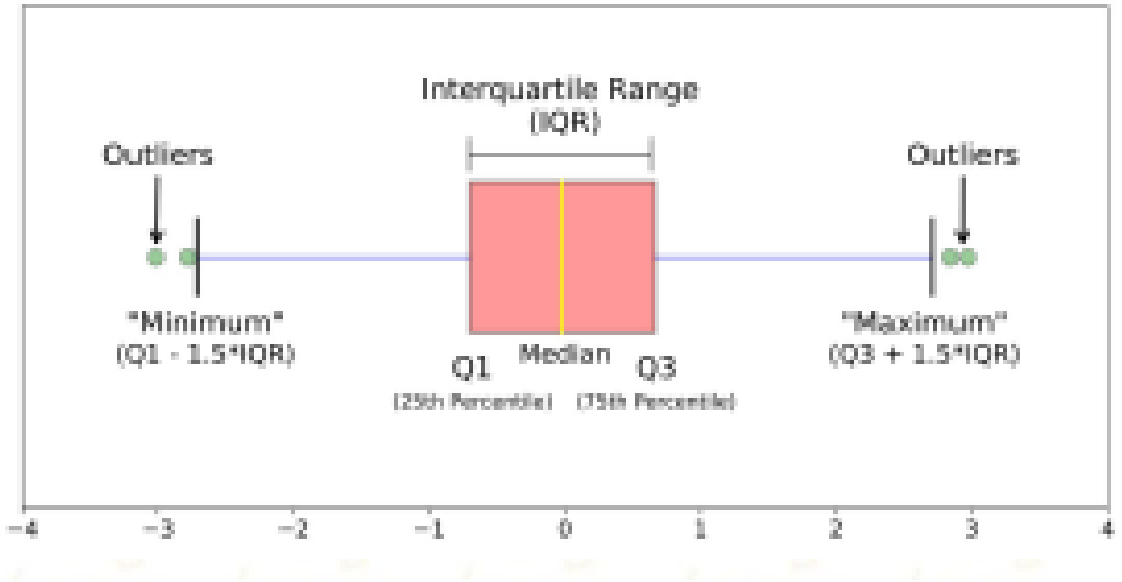
Histogram



Scatter Plot

Box Plot

- Maximum: nilai pengamatan terbesar
- Minimum: nilai pengamatan terkecil
- First Quartile: nilai tengah antara minimum dan median. Penjelasannya 25% data terendah ada di Q1
- Second Quartile (Median): nilai tengah antara titik Q1 dan Q3. Penjelasannya 50% data tersebar antara titik Q1 dan titik Q3
- Third Quartile: nilai tengah antara maximum dan median. Penjelasannya 75% data tertingg ada di Q3
- Whisker: is menunjukkan selisih antara Q1 ke minimum atau Q3 ke maximum



Interquartile Range
(IQR)

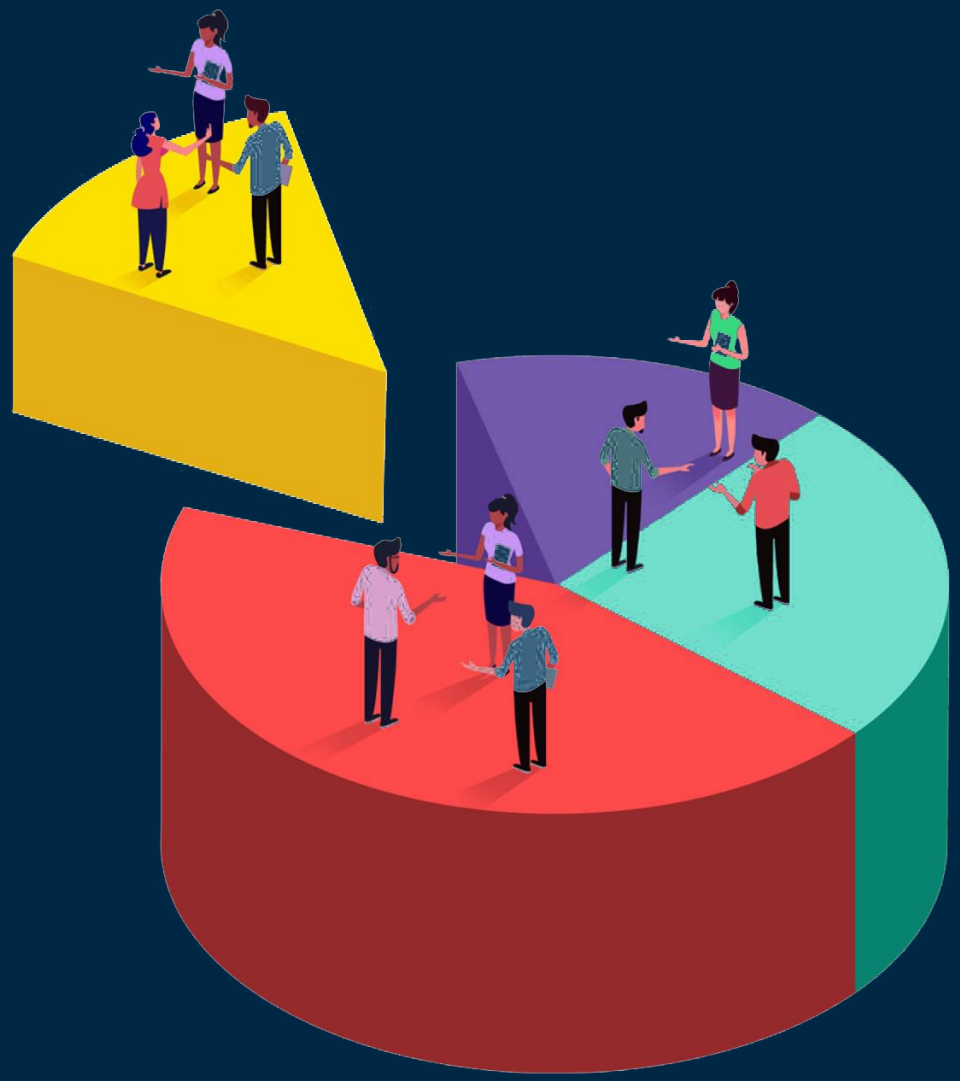
ADVANCED STATISTICS

03

Sampling
Hypothesis Testing

Sampling

Sampling adalah metode yang dilakukan untuk memilih pengamatan dari suatu populasi sehingga sample yang diambil dapat mewakili keseluruhan populasi.



Teknik Sampling

- Simple Random Sampling

  Teknik sampling yang dilakukan secara acak tanpa memperhatikan strata yang ada dalam populasi itu
  Syntax:

```
simple_s = data.sample(n=20)
```

  n adalah jumlah sampel yang ingin diambil

Teknik Sampling

- Systematic Sampling

  Teknik sampling yang dilakukan berdasarkan interval tertentu
  Syntax:

```
ind = np.arange(0, len(data), 4)
system_s = data.iloc[ind]
```

  Langkah pertama, tentukan dulu interval yang diinginkan.
  Kemudian, terapkan interval tersebut ke populasi data yang ada.

Teknik Sampling

- Stratified Sampling

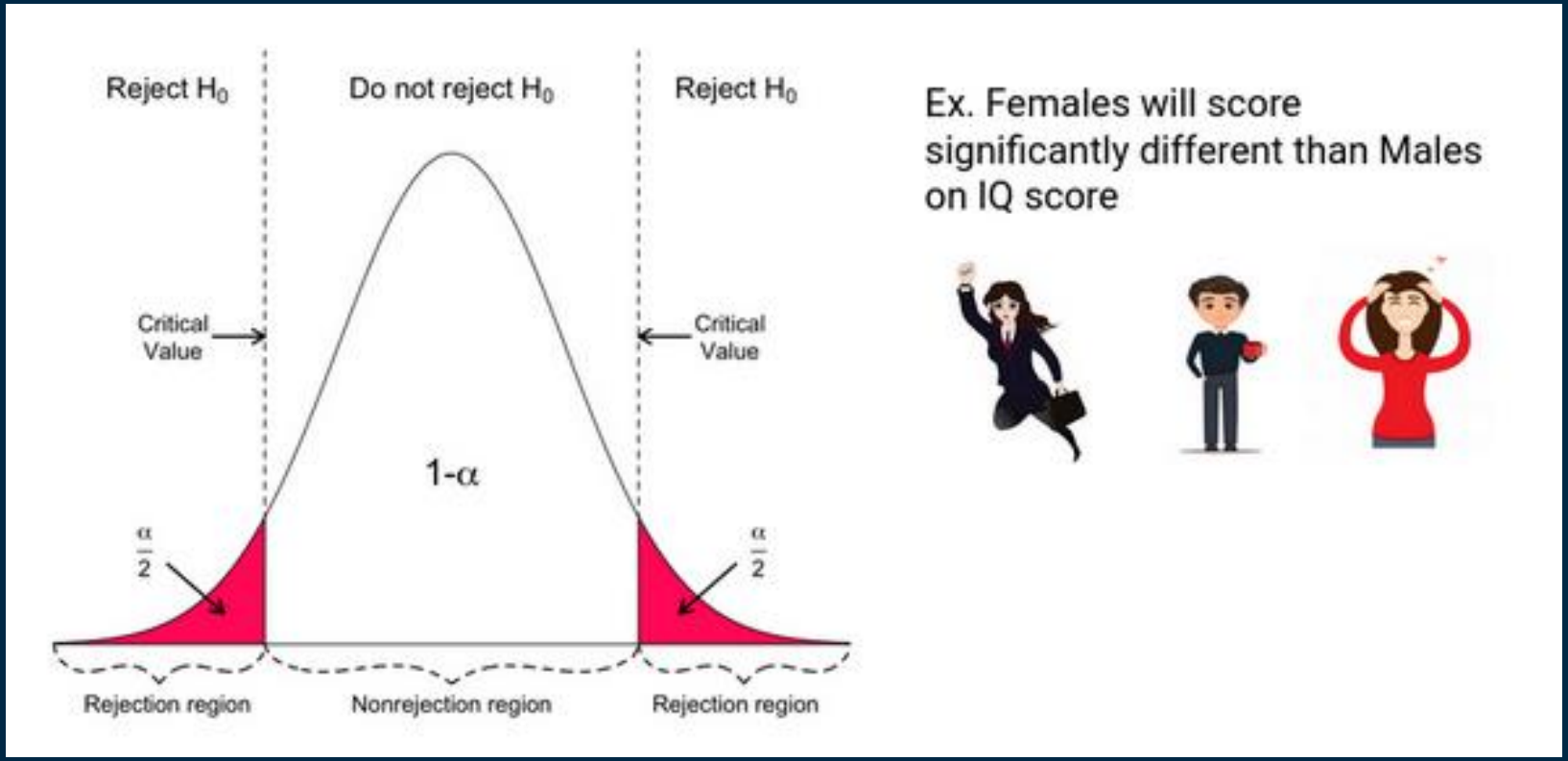
  Teknik sampling yang dilakukanberdasarkan kategori tertentu
  Syntax:

```
data[data['Health index'] <= 80].head()
```

  "<=80" adalah kategori yang diinginkan dari kolom "Health index"
  Sehingga sample-nya adalah Health index yang nilainya kurang atau sama dengan 80

Uji Hipotesis

Uji hipotesis digunakan untuk menguji kebenaran suatu pernyataan (hipotesis) secara statistik dan menarik kesimpulan apakah menerima atau menolak pernyataan (hipotesis) tersebut



Istilah Dalam Uji Hipotesis

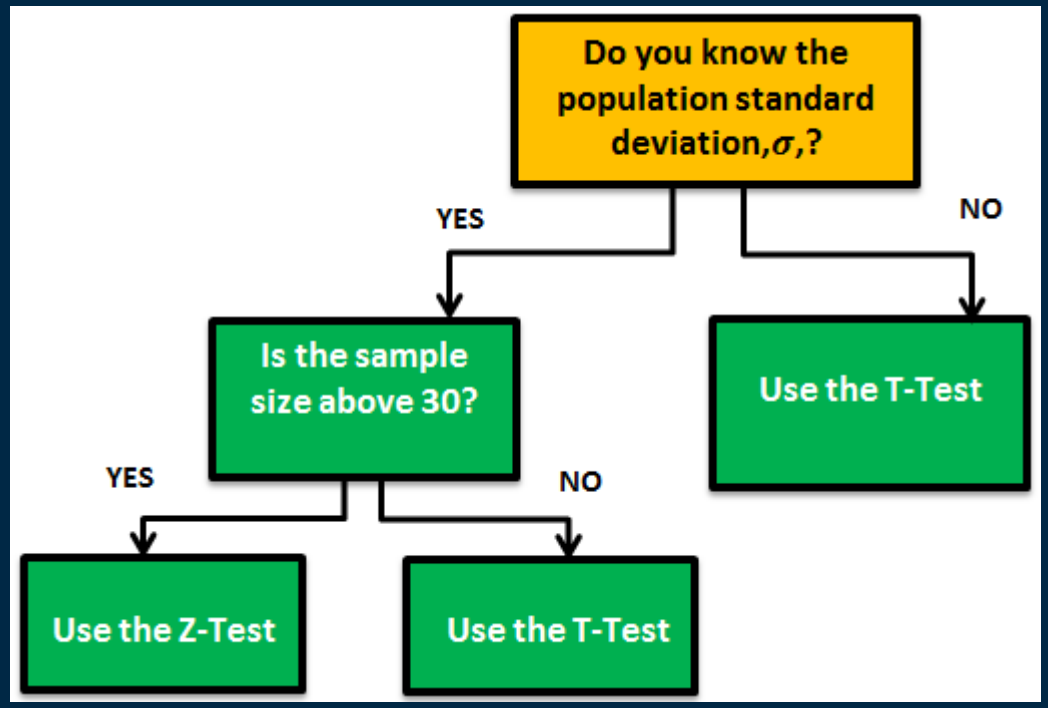
- H0 (hipotesis null) : hipotesis yang menerangkan tidak adanya hubungan antara variabel independen (X) dan variabel dependen (Y).
- Ha (hipotesis alternatif) : hipotesis yang menerangkan adanya hubungan antara variabel independen (X) dan variabel dependen (Y).
- Confidence level : nilai yang menggambarkan tingkat kepercayaan terhadap hasil kalkulasi.
- Alpha (signifikan level) : potongan nilai antara zona menerima atau menolak hipotesis null.
- p-value : nilai dari tes hipotesis yang digunakan untuk menerima atau menolak hipotesis null.
    - Jika p-value < alpha maka H0 ditolak.
    - Jika p-value > alpha maka H0 diterima

Z-Test

Z-test adalah uji statistik yang digunakan untuk mengetahui apakah suatu populasi memiliki rata-rata yang (a) sama dengan; (b) lebih kecil; atau (c) lebih besar dari suatu nilai rata-rata tertentu sesuai dengan hipotesis yang telah ditetapkan.

Syarat Z-test adalah:

- Diketahui varians populasi.
- Data berdistribusi normal.
- Jumlah sampel harus > 30



Z-Test

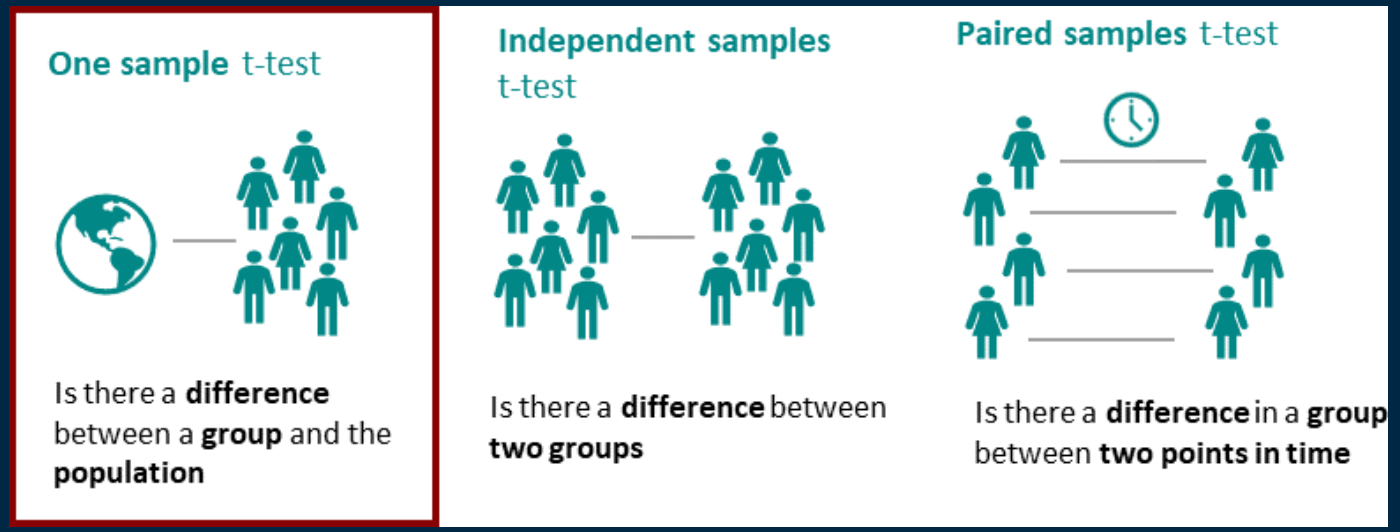
```
# Hitung rata-rata nilai CO2-nya
mean_CO2 = data['CO2'].mean()
print(mean_CO2)

# Ambil data yang transfomernya dalam keadaan "Working"
# Ini termasuk stratified random sampling
working = data[data['event'] == "Working"]

# Ambil sample sebanyak 35 dari data transformer keadaan "Working" menggunakan simple random sampling
sample_CO = working["CO2"].sample(n=35)

# Z-Test
ztest, pvalue = ztest(x1=sample_CO, x2=None, value=mean_CO2, alternative='larger')
print ('ztest: {} and pvalue: {}'.format(ztest, pvalue))

#Pengambilan kesimpulan
CI = 0.95
alpha = 1-0.95

if pvalue < alpha:
  print ("H0 ditolak")
  print ("pvalue {} < alpha {}".format(round(pvalue,2), round(alpha,2)))
else:
  print ("H0 diterima")
  print("pvalue {} > alpha {}".format(round(pvalue,2), round(alpha,2)))
```

T-Test

T -test merupakan alat analisis statistik yang digunakan untuk membandingkan rata-rata (mean) dari sekelompok data atau dua kelompok data.

Syarat T-test adalah:

- Tidak diketahui varians populasi.
- Data berdistribusi normal.
- Jumlah sampel kurang atau sama dengan 30



T-Test

```
# Hipotesis : Apakah O2 lebih kecil dari rata-rata keseluruhan O2 pd saat transformer bekerja?

# Mean Oksigen
mean_O2 = data['Oxigen'].mean()
print(mean_O2)

# Ambil data yang transfomernya dalam keadaan "Working"
# Ini termasuk stratified random sampling
working = data[data['event'] == "Working"]

# Ambil sample sebanyak 25 dari data transformer keadaan "Working" menggunakan simple random sampling
sample_O = working["Oxigen"].sample(n=25)

# T-Test
ttest, pvalue = stats.ttest_1samp(sample_O, mean_O2, alternative='less')
print ('ttest: {} and pvalue: {}'.format(ttest, pvalue))

#Pengambilan Kesimpulan
CI = 0.95
alpha = 1-0.95

if pvalue < alpha:
  print ("H0 ditolak")
  print ("pvalue {} < alpha {}".format(round(pvalue,2), round(alpha,2)))
else:
  print ("H0 diterima")
  print("pvalue {} > alpha {}".format(round(pvalue,2), round(alpha,2)))
```

Chi-square

Chi-square bertujuan untuk melakukan prediksi secara statistik dengan melihat ada atau tidaknya hubungan antar variabel.

```
# set kolom 'event' sebagai index
chi = data[['event', 'CO', 'CO2', "Hydrogen", "Methane"]]
chi = chi.set_index('event')

# Chi-Square
stat, pvalue, dof, expected = chi2_contingency(chi)
print ('Chi square: {} and pvalue: {}'.format(stat, pvalue))

# Pengambilan Keputusan
CI = 0.95
alpha = 1-0.95

if pvalue < alpha:
  print ("H0 ditolak")
  print ("pvalue {} < alpha {}".format(round(pvalue,2), round(alpha,2)))
else:
  print ("H0 diterima")
  print("pvalue {} > alpha {}".format(round(pvalue,2), round(alpha,2)))
```