DigitalSkola

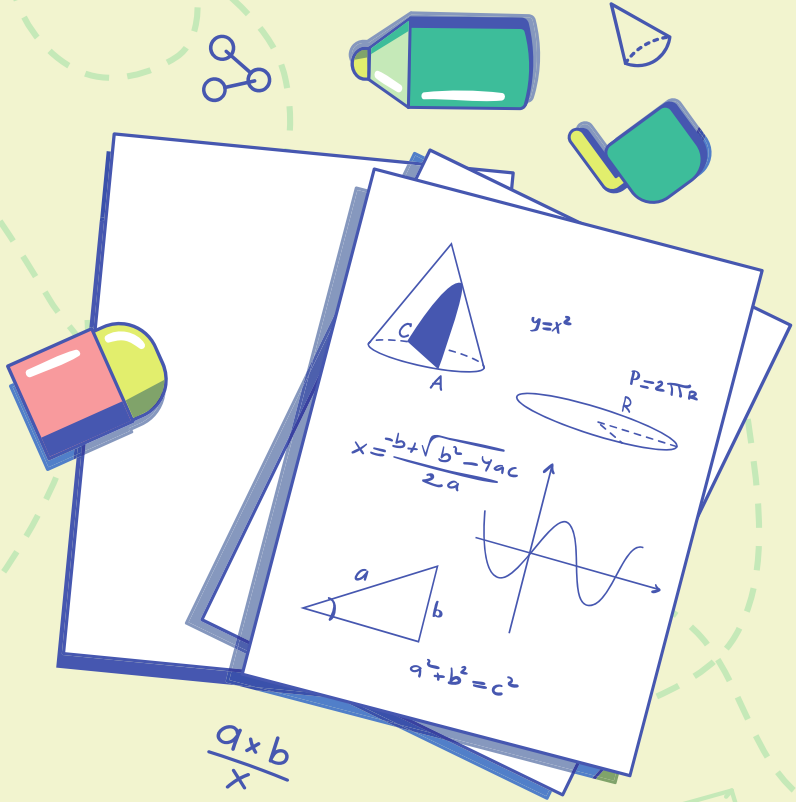
Learning Progres Review Week #8

By Optimistic team

DigitalSkola

Session 22

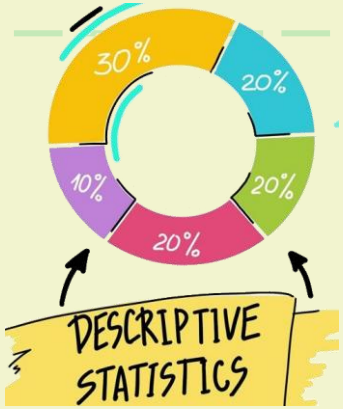
BASIC STATISTICS



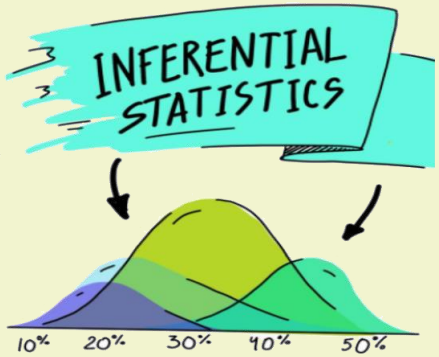
Statistika

Statistika adalah ilmu tentang pengumpulan, analisis, dan interpretasi data dalam rangka pengambilan keputusan.

Ilmu Statistika

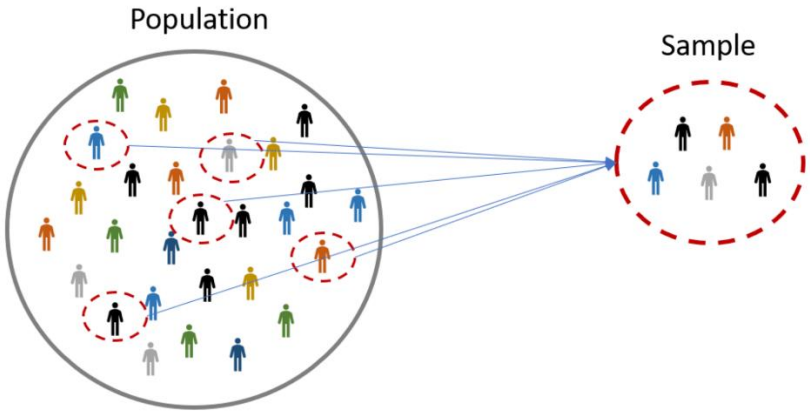


Statistika Deduktif / Deskriptif
Merupakan statistika sederhana. Hanya menyajikan dan mengolah data. Tanpa membuat kesimpulan tentang karakteristik populasi.



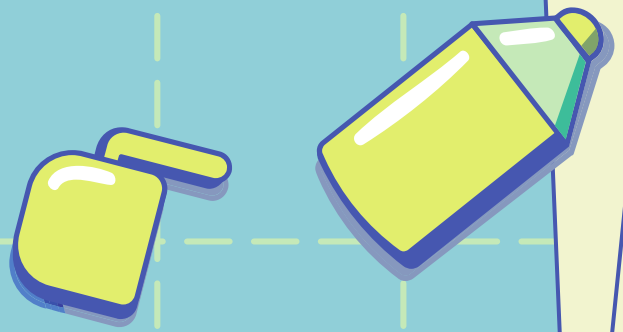
Statistika Induktif / Inferensial
Merupakan statistika lanjutan. Berkaitan dengan penarikan kesimpulan tentang karakteristik populasi

- Populasi adalah seluruh objek yang diamati dari seluruh kriteria tertentu.
- Sampel adalah bagian dari anggota populasi, dianggap menjadi gambaran bagi populasi asalnya.



Populasi dan Sampel

Jenis Data



1. **Data Kualitatif**

   Data berupa kata-kata yang berhubungan dengan karakteristik dalam bentuk sifat (bukan angka). Data Categorical seperti, Data Nominal, dan Data Ordinal

2. **Data Kuantitatif**

   Data berupa angka atau bilangan. Data Numerical seperti, Data Diskrit, dan Data Kontinu.

Ukuran Pemusatan Data

MEAN

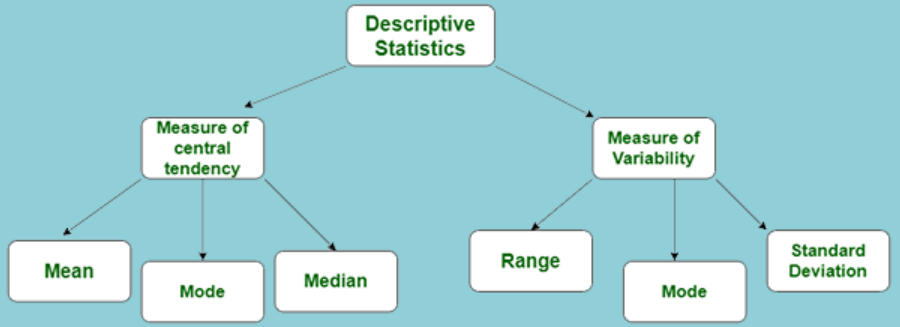
Mean atau istilah lainnya nilai rata-rata adalah jumlah keseluruhan data dibagi banyaknya data.

MODUS

Nilai yang paling banyak muncul.

MEDIAN

Median atau nilai tengah adalah pemusatan data yang membagi suatu data menjadi setengah (50%) data terkecil dan terbesarnya.



Penyajian Data

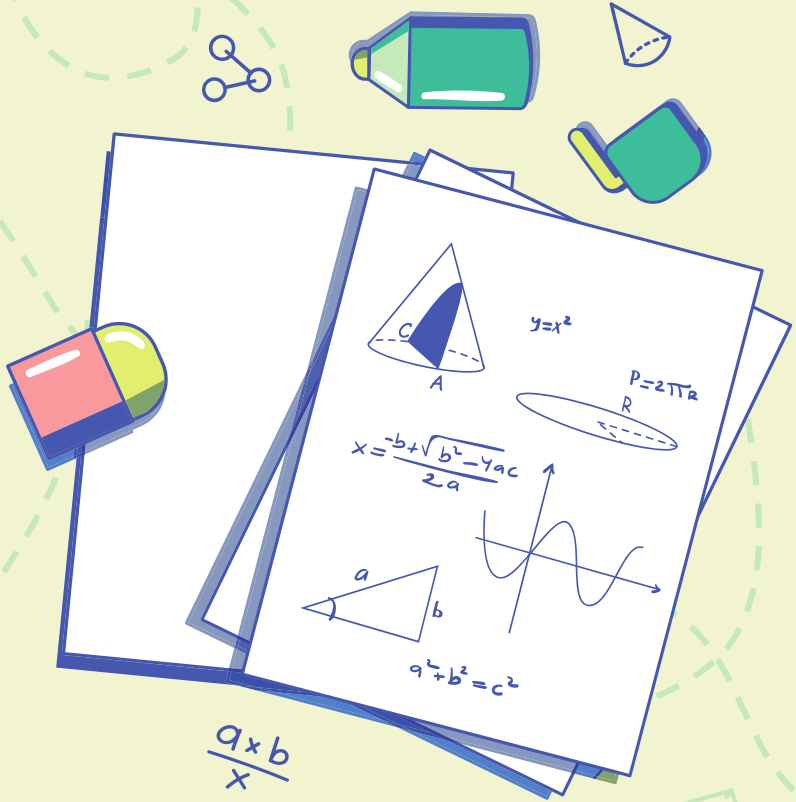
Macam-macam Diagram

- Diagram Batang
- Diagram Garis
- Diagram Lingkaran dan Pastel
- Diagram Lambang / Piktogram
- Diagram Pencar / Titik
- Diagram Peta / Kartogram

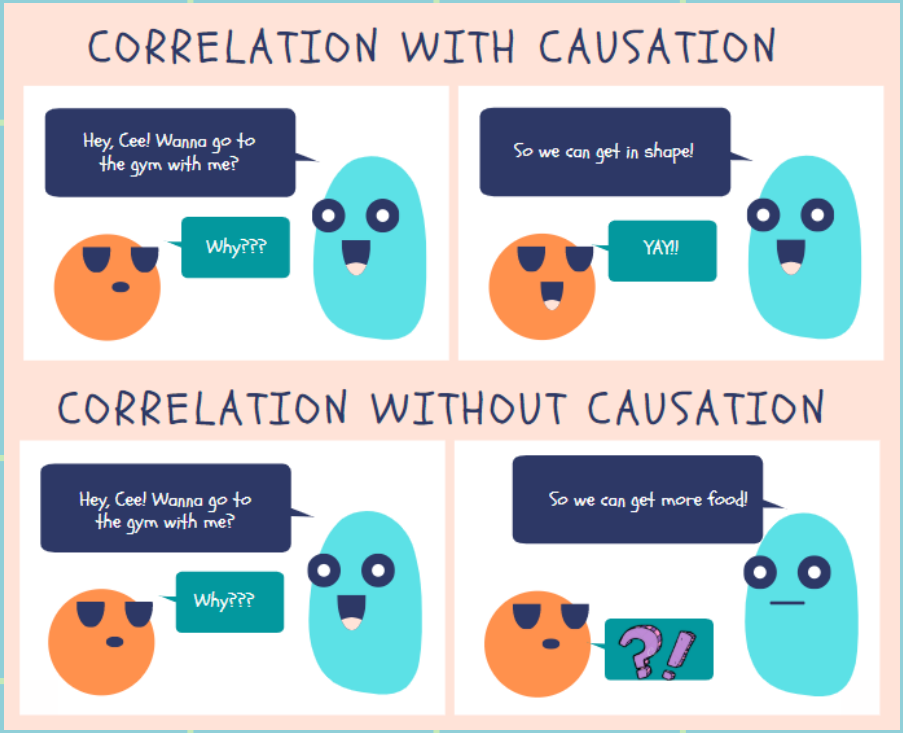
DigitalSkola

Session 23

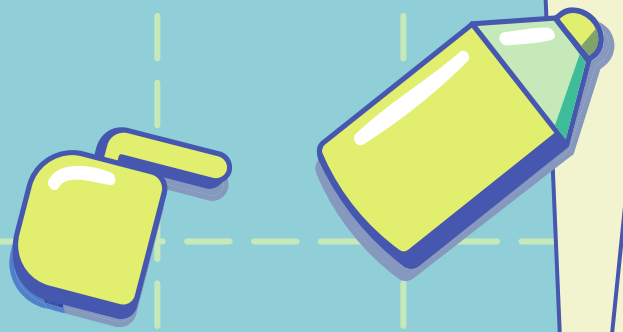
Intermediate Statistics



Correlation and Causality



Correlation

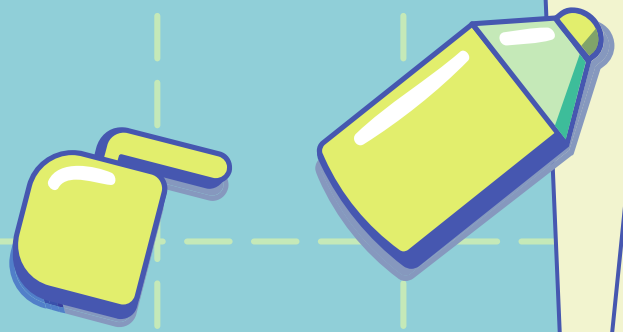


Analisis korelasi adalah suatu cara atau metode untuk mengetahui ada atau tidaknya hubungan linear antar variabel.

Contoh Kasus yang Memiliki Korelasi.

- Hubungan antara kenaikan harga BBM (X) dengan harga kebutuhan pokok (Y)
- Hubungan tingkat pendidikan (X) dengan tingkat pendapatan (Y)

Causality



kausalitas adalah perihal sebab akibat, hubungan **kausalitas** dalam teks eksplanasi adalah hubungan yang dibentuk atas suatu kejadian (sebab) dan dampak (akibat) dari kejadian tersebut.

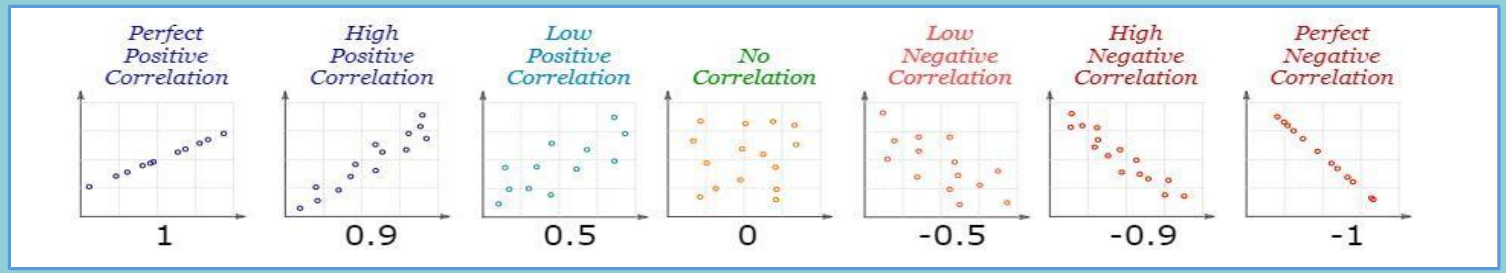
Contohnya dalam teks eksplanasi mengenai pemanasan global.

Menghitung Correlation

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Korelasi dapat memiliki nilai:

- 1 adalah korelasi positif sempurna
- 0 tidak ada korelasi (tidak ada hubungan)
- -1 adalah korelasi negatif sempurna



Probability and Distribution

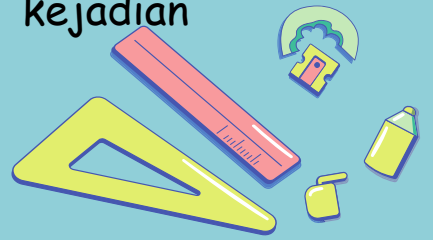
Probablity adalah sebagai suatu cara untuk menyatakan kepercayaan atau pengetahuan terhadap seberapa besar peluang terjadinya suatu kejadian yang akan atau yang telah terjadi.

- **Probability mass function (pmf)**

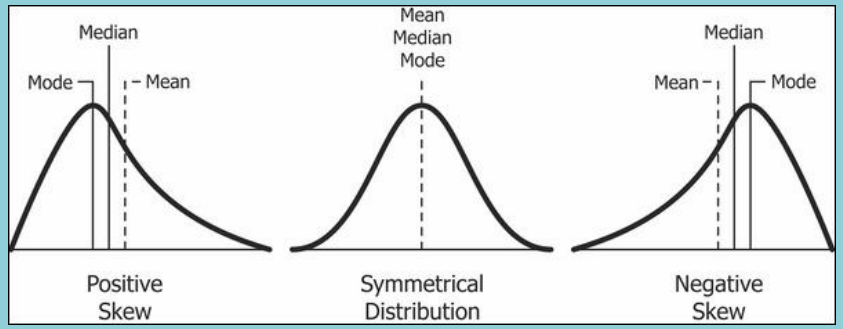
Fungsi matematika yang memberikan kemungkinan dari variabel acak untuk memiliki suatu nilai. Digunakan untuk variabel discrete atau categorical.

- **Cummulative density function (cdf)**

Fungsi yang menjumlahkan nilai kemungkinan sampai suatu kejadian tertentu.

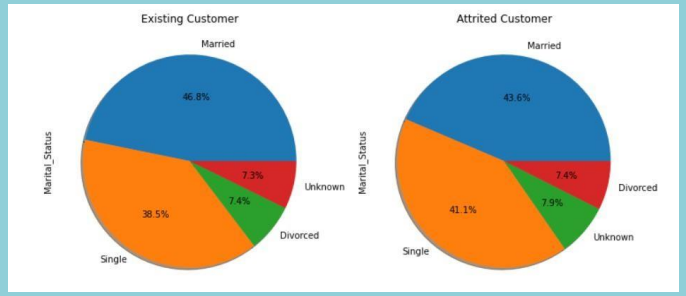


Skewnes

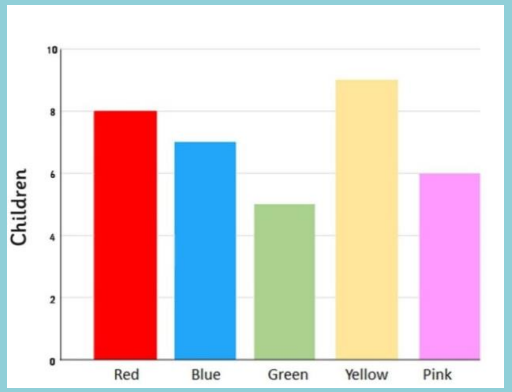


Skewness adalah derajat ketidaksimetrisan suatu distribusi. Jika kurva frekuensi suatu distribusi memiliki ekor yang lebih memanjang ke kanan (dilihat dari meannya) maka dikatakan miring ke kanan (positif) dan jika sebaliknya maka miring ke kiri (negatif). Secara perhitungan, skewness adalah momen ketiga terhadap mean.

Statistical Plot



Pie Chart

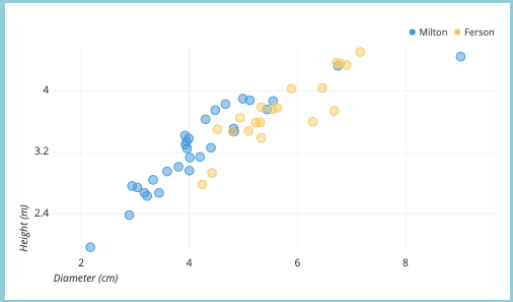


Bar plot

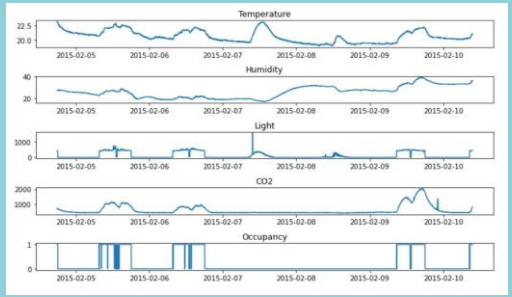
Histogram

Statistical Plot

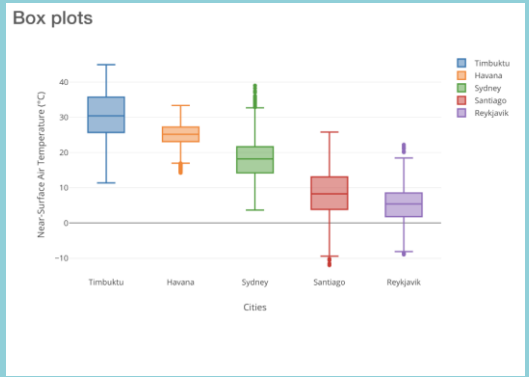
Scatter Plot



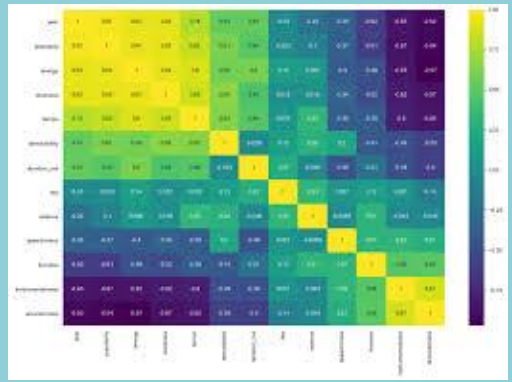
Line plot



Boxplot



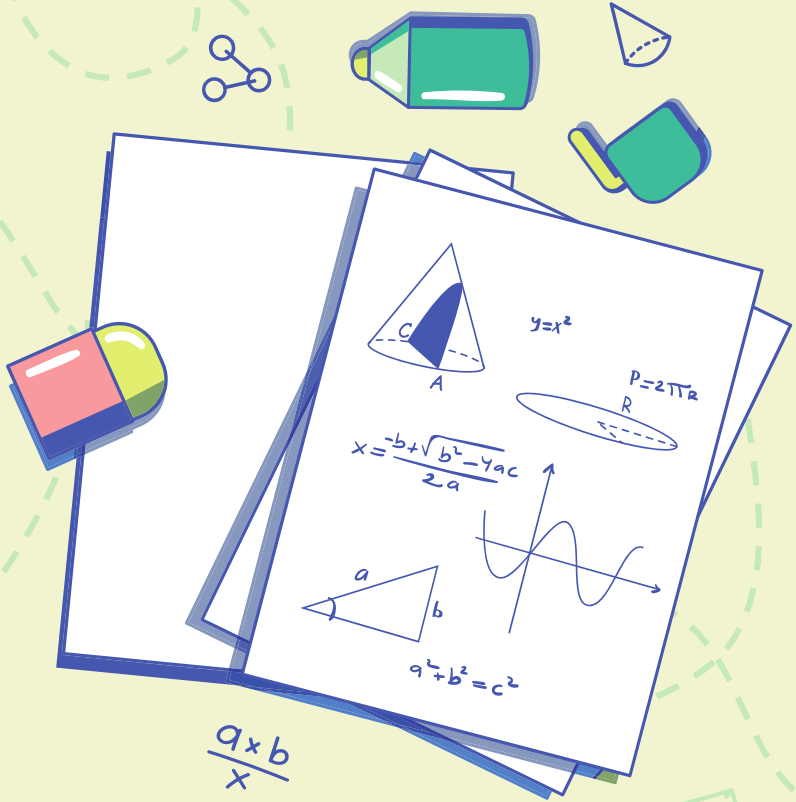
Heatmap



DigitalSkola

Session 24

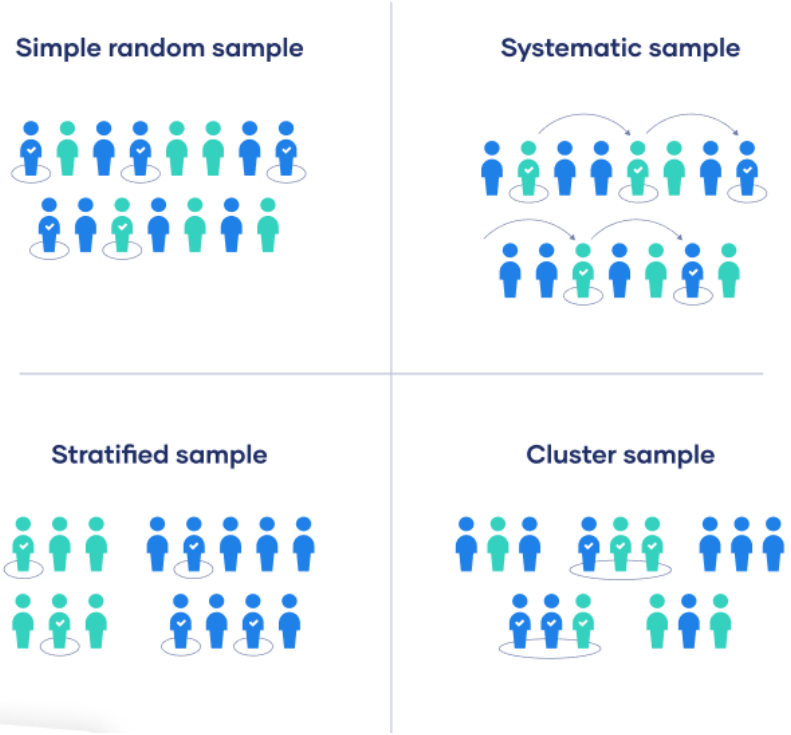
Advance Statistics



Sampling

Sampling, adalah proses mengambil beberapa sampel yang mengacu pada metode statistik, dari populasi agar dapat memperkirakan karakteristik populasi. Sampel harus representative agar dapat menggambarkan populasi secara tepat

1. Simple Random sampling
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling



Teknik Sampling

1. Simple Random Sampling

Pada sistem random sampling setiap objek memiliki peluang yang sama untuk dijadikan sampel. Teknik ini baik digunakan pada populasi homogen atau yang memiliki karakteristik mirip/hampir sama.

Sample 400 observation

```
df.sample(n=400)
```

	user_id	timestamp	group	landing_page	converted
84561	889913	2017-01-10 03:29:44.877020	control	old_page	1
158240	795331	2017-01-05 02:28:31.324120	control	old_page	0
16396	931481	2017-01-14 10:27:44.036521	treatment	new_page	0
214735	678545	2017-01-10 20:09:08.870690	control	old_page	1
270680	938785	2017-01-05 16:55:07.131904	treatment	new_page	1
...	...	...	...	...	...
189394	634341	2017-01-20 12:30:19.787521	control	old_page	0
49744	873063	2017-01-07 08:50:00.929656	treatment	new_page	1
158707	783264	2017-01-06 06:11:55.341772	treatment	new_page	0
110721	870153	2017-01-11 14:05:31.435297	treatment	new_page	0
218199	703692	2017-01-24 05:53:55.056226	control	old_page	0

400 rows × 5 columns

Sample 30% from dataset

```
df.sample(frac=0.3)
```

	user_id	timestamp	group	landing_page	converted
186583	893743	2017-01-11 04:39:37.172078	control	old_page	0
256031	853511	2017-01-03 05:00:45.694946	control	old_page	0
283006	814654	2017-01-05 21:04:18.990880	treatment	new_page	0
254110	751639	2017-01-08 14:59:31.537678	control	old_page	0
181889	665122	2017-01-06 17:44:30.179297	control	old_page	0
...	...	...	...	...	...
52803	670220	2017-01-06 11:44:33.323235	control	old_page	0
26239	888236	2017-01-23 05:55:08.816955	treatment	new_page	0
47497	845510	2017-01-03 16:58:13.046340	control	old_page	0
77750	653130	2017-01-22 18:56:55.587907	treatment	new_page	0
119070	785120	2017-01-14 05:27:37.836025	control	old_page	0

88343 rows × 5 columns

2. Stratified Sampling

Prosesnya dengan cara membagi objek dulu ke dalam berbagai strata, kemudian setelah itu dilakukan sampling. Hal ini dilakukan agar sampel tidak kehilangan karakteristik dari strata tertentu. Teknik ini baik digunakan untuk data heterogen dan sebaran strata yang tidak seimbang.

Sampling for Each Group

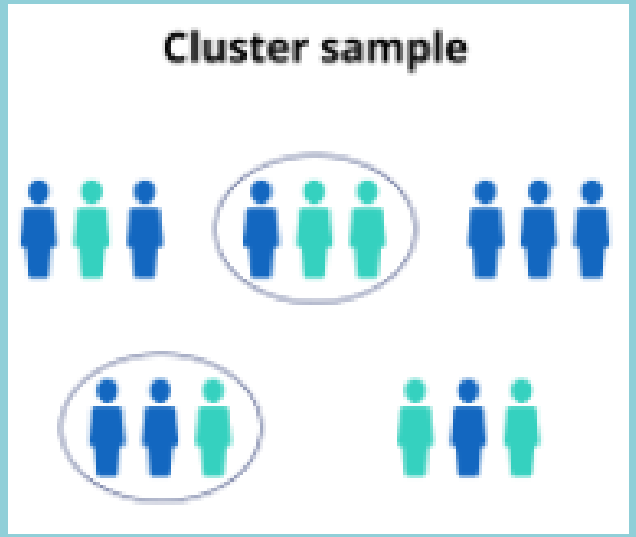
```
df.groupby(['group'],as_index=True).apply(lambda x: x.sample(n=200,random_state=123))
```

group		user_id	timestamp	group	landing_page	converted
control	95574	704344	2017-01-08 06:33:15.620318	control	old_page	0
	282637	903218	2017-01-07 16:40:31.904242	control	old_page	0
	201262	724634	2017-01-05 18:38:31.257679	control	old_page	0
	93315	750623	2017-01-21 19:20:32.814948	control	old_page	0
	16163	651056	2017-01-04 03:17:39.846424	control	old_page	0
...	...	...	...	...	...	...
treatment	16034	665227	2017-01-18 06:10:37.832101	treatment	new_page	1
	241972	818984	2017-01-23 01:45:24.506789	treatment	new_page	0
	135298	843757	2017-01-04 03:10:19.433517	treatment	new_page	0
	200501	659763	2017-01-24 13:21:56.026713	treatment	new_page	0
	158648	788418	2017-01-14 05:09:32.246838	treatment	new_page	0

400 rows × 5 columns

3. Cluster Sample

Sampling dilakukan terhadap cluster tertentu. Teknik ini baik digunakan pada populasi yang homogen secara antar-cluster, tetapi heterogen secara intra-cluster. Dapat mengurangi biaya secara signifikan



4. Systematic Sample

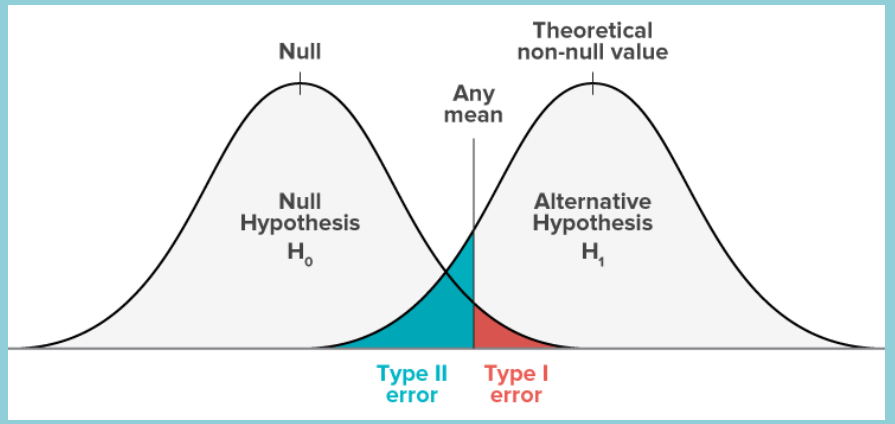
Sampling yang dilakukan secara sistematis, misal menggunakan nomor urut tertentu perlu diantisipasi, karena teknik ini dapat menyebabkan bias jika sampel yang digunakan telah diatur sedemikian rupa.

Hypothesis Testing

Hypothesis testing atau uji hipotesis adalah bentuk statistical inference yang menggunakan data dari sampel untuk menarik kesimpulan atas parameter populasi. Hypothesis Testing terbagi menjadi 2 yaitu, Null Hypothesis, dan Alternative Hypothesis

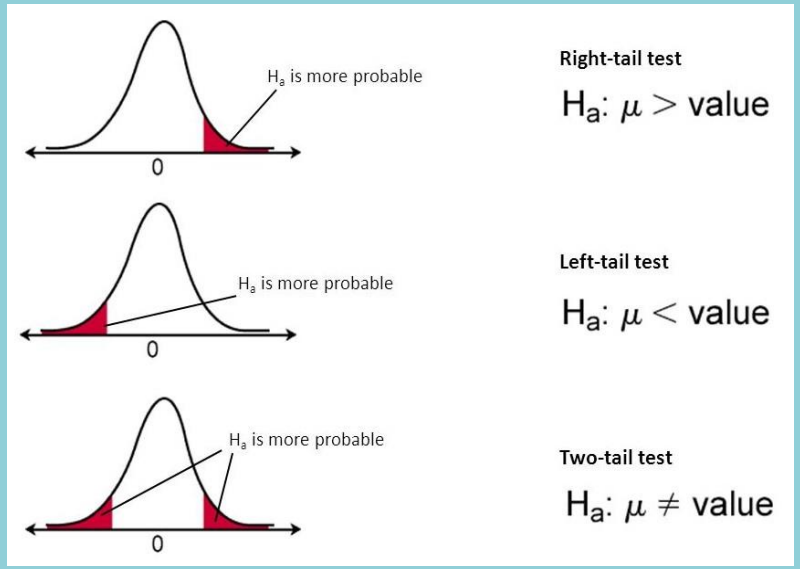
Poin utama dalam pengujian hipotesis:

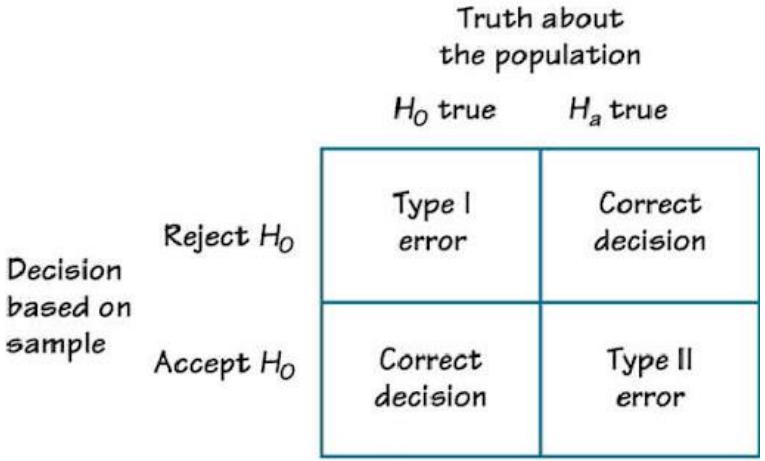
- Menemukan null hypothesis
- Menyatakan null hypothesis
- Memilih jenis uji yang perlu dilakukan
- Hasil menolak null hypothesis atau tidak



Jenis Hipotesis

- Simple Hypothesis, Hipotesis yang disertakan dengan nilai persis.
- Composite Hypothesis, Hipotesis yang disertakan dengan selang nilai.
- One-tailed Hypothesis, Hipotesis alternative memiliki satu arah perbandingan nilai (hanya lebih besar atau lebih kecil).
- Two-tailed Hypothesis, Hipotesis alternative memiliki dua arah perbandingan nilai.





Type I and II Errors Testing

Metrik untuk mengukur error untuk ditolak H0 (Error Type I) disebut P-Value (Probability Value). P-Value adalah probabilitas/kesempatan yang mewakili untuk menerima H1 atau menolak H0.

Bagaimana untuk memutuskan,

Ada dua keputusan:

- Jika P-Value < a maka H1 dapat diterima
- Jika P-Value > a maka H0 dapat diterima

Biasanya nilai a adalah 1%, 5%, 10%

Itu tergantung pada keputusan bisnis atau seberapa yakin analisisnya.

Statistical Methods

T-test

T-test adalah metode statistik yang menggunakan rata-rata sampel dan distribusi untuk membandingkan antara 2 populasi.

Tipe T-test:

- Tes untuk 1 Populasi -> H0: avg = 10 vs H1: avg != 10
- Tes untuk 2 Populasi Independen-> H0: avg1 = avg2 vs H1: avg1 != avg2

Asumsi

- Sample adalah distribusi normal
- Atau sejumlah besar sampel (teorema limit pusat)

Statistical Methods

Chi-square

Uji Chi-square adalah metode yang digunakan untuk menguji apakah ada hubungan antara dua variabel kategori. Ini juga digunakan untuk menyelidiki apakah distribusi variabel kategori berbeda satu sama lain..

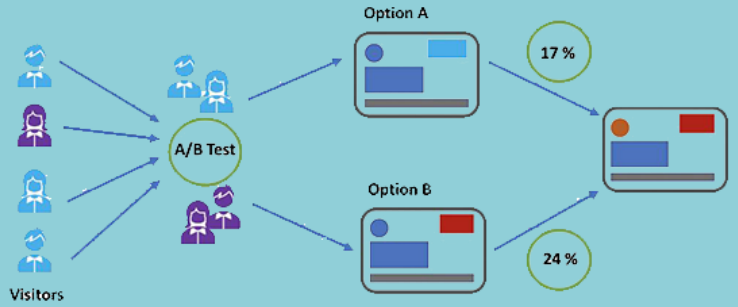
Hipotesis:

H0: X dan Y adalah independen

H1: X dan Y adalah independen

A/B Testing

A/B testing adalah eksperimen terkontrol terhadap dua variabel atau lebih yang dilakukan secara bersamaan untuk melihat variabel mana yang memberikan performa terbaik. Performa di sini diukur dengan conversion rate, variabel mana yang menghasilkan conversion rate lebih tinggi.



Proses A/B Testing

- **Membuat hipotesis**

H0 : tidak ada perbedaan conversion rate antara variabel A dan variabel B

H1 : grup pada variabel B memiliki conversion rate lebih tinggi dari grup pada variabel A

- **Membuat grup kontrol dan grup tes**

Kontrol : grup yang menerima variabel A (tidak ada perubahan)

Tes : grup yang menerima variabel B (dengan perubahan) Umumnya pengambilan sampel menggunakan random sampling

- **Melakukan A/B Testing** dan mengumpulkan data Setelah data didapatkan, untuk membuktikan statistical significance dapat digunakan T-test atau Z-test. Uji hipotesis ini biasa digunakan untuk membandingkan rata-rata perbedaan pada dua grup.

DigitalSkola

Thank you!

Our Team

1. **Aldiva Wibowo**
2. **Asprizal Rizky**
3. **Gilang Rahmat R**
4. **Lutfia Humairosi**
5. **Millenia Winadya P**