

MySkill

#RintisKarirImpian

Intro to Statistics

Achmad Rozie

Data Analyst at Flip



Nanti kita akan exercise
bareng-bareng, jadi tolong buka
Google Sheet ya!



Sub Topik



- Populasi dan sampel
- Pengantar Statistika dan Analisis Data
- Uji Hipotesis dan Regresi



Populasi dan Sampel



STATISTIK

*Nilai-nilai ukuran data (atau fakta) yang **mudah untuk dimengerti**.*

Misal: Statistik kasus positif harian Covid-19 Periode Januari-Juni 2020.

STATISTIKA

*Ilmu yang berkaitan dengan **cara pengumpulan, pengolahan, analisis, dan penarikan kesimpulan** atas data.*

STATISTIKA DESKRIPTIF

.Describe and summarize of data

STATISTIKA INFERENSI

Use a sample of data to make *inferences* (dugaan) about a larger population

CONTOH

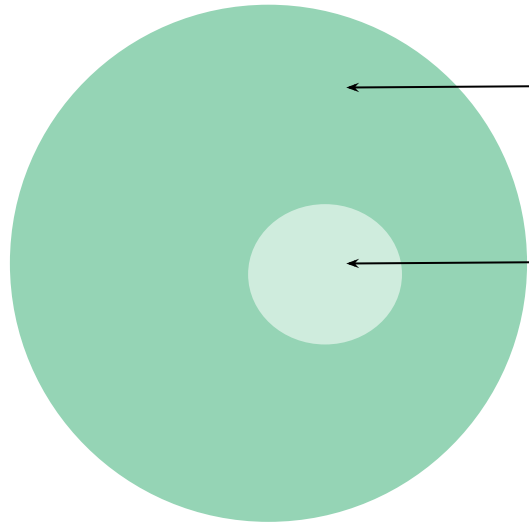
Data banyaknya pelanggan yang beli gorengan di warung A di kompleks X di 15 hari terakhir pada bulan September 2020:

26 37 76 49 95 69 83 87 39 95 59 83 83 87 46

Pertanyaan:

1. Berapa rata-rata banyak pelanggan yang datang di 10 hari terakhir tersebut? - ***Statistika Deskriptif***
2. Apakah rata-rata pelanggan yang datang ke warung A adalah representasi rata-rata pelanggan warung A di kompleks X secara keseluruhan? - ***Statistika Inferensi***

Populasi, Sampel, Sampel Acak, Data



POPULASI



keseluruhan dari objek penelitian yang menjadi pusat perhatian dan menjadi sumber data penelitian.



SAMPEL



bagian dari populasi yang dipilih dengan menggunakan aturan tertentu, yang digunakan untuk mengumpulkan informasi/data yang menggambarkan sifat atau ciri yang dimiliki populasi.



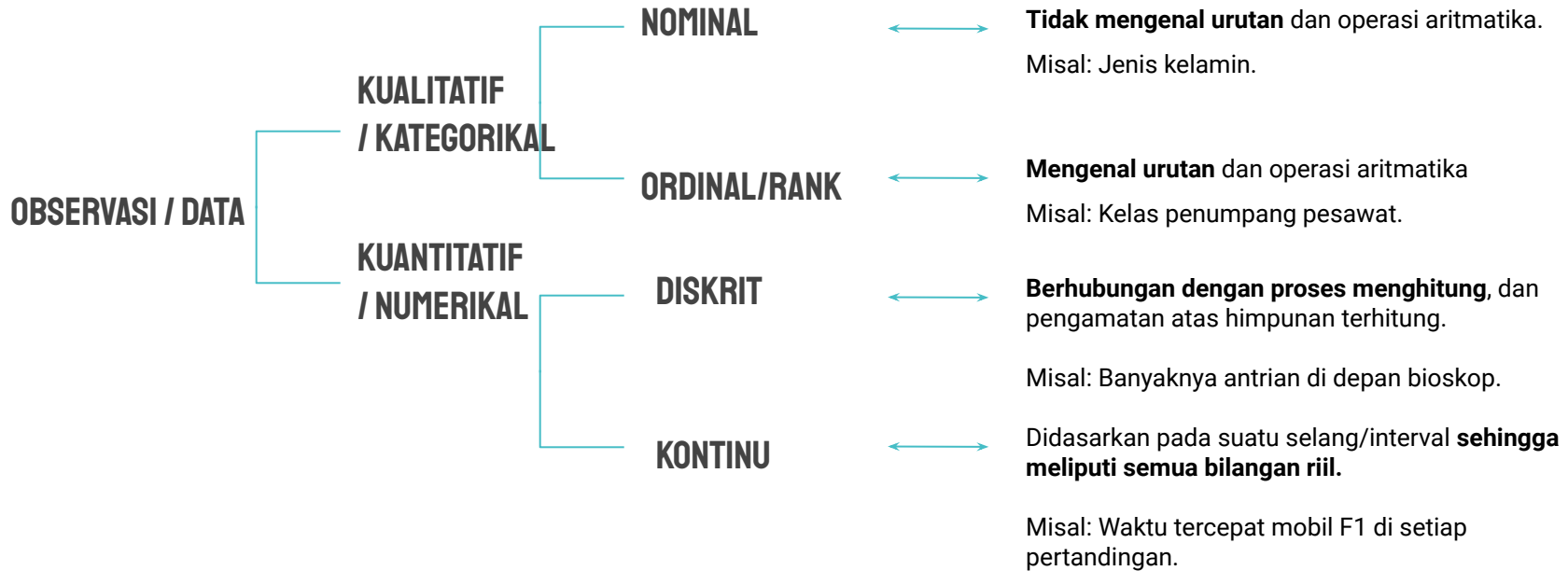
DATA



Hasil pengukuran atau pengamatan



Type of Data



Type of Data (Con't)

Numeric (Quantitative)

- **Continuous (Measured)**
 - Airplane speed
 - Time spent waiting in line
- **Discrete (Counted)**
 - Number of pets
 - Number of packages shipped

Categorical (Qualitative)

- **Nominal (Unordered)**
 - Married/unmarried
 - Country of residence
- **Ordinal (Ordered)**
 - ☐ Strongly disagree
 - ☐ Somewhat disagree
 - ☐ Neither agree nor disagree
 - ☒ Somewhat agree
 - ☐ Strongly agree

Type of Data (Con't)



Nominal (Unordered)

- Married/unmarried (1 / 0)
- Country of residence (1 , 2 , ...)

Ordinal (Ordered)

- Strongly disagree (1)
- Somewhat disagree (2)
- Neither agree nor disagree (3)
- Somewhat agree (4)
- Strongly agree (5)



SAMPEL ACAK

EKSPERIMEN ACAK

- Dapat diulangi baik oleh si pengamat sendiri maupun orang lain
- Proporsi keberhasilan dapat diketahui dari hasil sebelumnya
- Bisa diukur (diamati)
- Hasilnya tidak bisa ditebak karena adanya galat/error

RUANG SAMPEL (S)

Himpunan dari semua kemungkinan hasil dari suatu eksperimen acak.

DISKRIT

Banyaknya elemen dapat dihitung (countable).

KONTINU

Elemen-elemen dari ruang sampel adalah bagian dari suatu interval

KEJADIAN (EVENT)

Himpunan bagian (subset) dari suatu ruang sampel (S).

Random Sampling

RANDOM SAMPLING

Seluruh proses pengambilan **sampel dilakukan secara independen** dari anggota populasi lainnya.

Contoh: Mengambil angka pengeluaran rumah tangga di Jawa Tengah dengan mengambil sampel dari beberapa kabupaten dan kota.

SYSTEMATIC SAMPLING

Setelah kita memutuskan **ukuran sampel seperti apa**, atur elemen populasi dalam beberapa urutan dan **pilih secara berkala dari daftar**.

Contoh: Mengambil setiap pelanggan kelipatan ke-10 dalam supermarket untuk mencatat karakteristik berbelanja.

STRATIFIED SAMPLING

Populasi **dibagi menjadi beberapa karakteristik**, lalu populasi diambil sampel secara acak dalam setiap kategori.

Contoh: dari data didapatkan 40% perempuan dan 60% laki-laki, lalu sampel dipilih secara acak dengan proporsi yang sama.

Random Sampling (Con't)

RANDOM SAMPLING

Seluruh proses pengambilan **sampel dilakukan secara independen** dari anggota populasi lainnya.

SYSTEMATIC SAMPLING

Setelah kita memutuskan **ukuran sampel seperti apa**, atur elemen populasi dalam beberapa urutan dan **pilih secara berkala dari daftar**.

STRATIFIED SAMPLING

Populasi dibagi menjadi beberapa karakteristik, lalu populasi diambil sampel secara acak dalam setiap kategori.

Give 1 example of each random sampling above!

Random Sampling (Con't)

RANDOM SAMPLING

SYSTEMATIC SAMPLING

STRATIFIED SAMPLING

Random Sampling: Mengambil 1 buah dalam keranjang yang terdiri dari banyak buah.

Systematic Sampling: Seorang peneliti ingin mengambil sampel dari jumlah total konsumen yang berbelanja di sebuah toko. Jika total konsumen yang berbelanja adalah 1000 orang dan peneliti ingin mengambil sampel sebanyak 100 orang, maka peneliti dapat menentukan interval pemilihan sampel sebagai 10 ($1000/100 = 10$).

Stratified

Semisal ingin melakukan survey terhadap kepuasan suatu produk

- Strata pertama adalah pelanggan perempuan
- Strata kedua adalah pelanggan laki-laki
- Strata ketiga adalah pelanggan dengan pendapatan diatas rata-rata.

Sampling:

Pengantar Statistika dan Analisis Data

Karakter Distribusi



DISTRIBUTION PARAMETER

UKURAN PEMUSATAN	↔	Mean, Median, Modus, Kuartil Atas, Kuartil Bawah, dll
UKURAN PENYEBARAN	↔	Range, Std.Dev, Variansi, dll
KEMENCENGAN	↔	Skewness
KELANCIPAN	↔	Kurtosis



Definition of each Measure

- **Mean** adalah nilai rata-rata dari sekelompok data.
- **Median** adalah nilai tengah dari sekelompok data yang telah diurutkan.
- **Modus** adalah nilai yang paling sering muncul dalam sekelompok data.

- **Kuartil atas (Q3, P75)** adalah nilai yang membagi data menjadi empat bagian dengan 75% data di bawahnya.
- **Kuartil bawah (Q1, P25)** adalah nilai yang membagi data menjadi empat bagian dengan 25% data di bawahnya.

Definition of each Measure

- **Range** adalah perbedaan antara nilai terbesar dan terkecil dalam sekelompok data.
- **Standar deviasi** (std dev) mengukur seberapa jauh rata-rata dari sekelompok data dari nilai-nilai dalam kelompok tersebut.
- **Variansi** adalah nilai rata-rata dari perbedaan dari setiap data dari mean.



UKURAN PEMUSATAN DATA

Statistik yang memberikan informasi **dimana data terkumpul** dengan ukuran/jumlah tertentu.

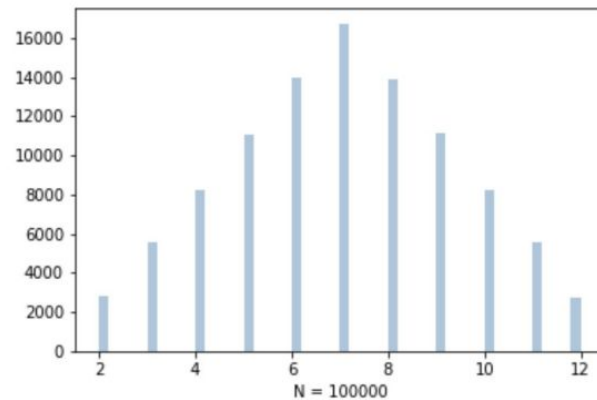
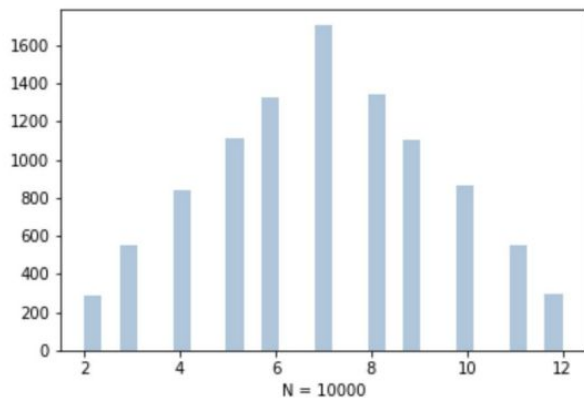
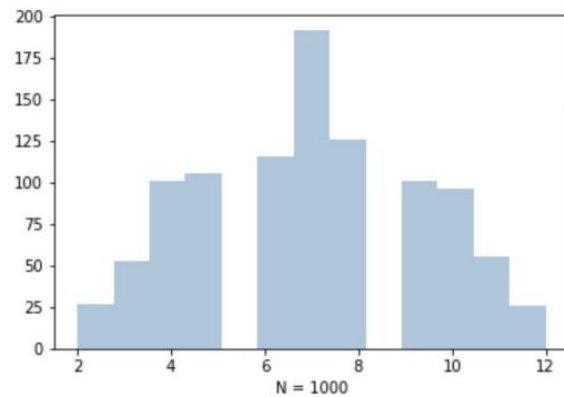
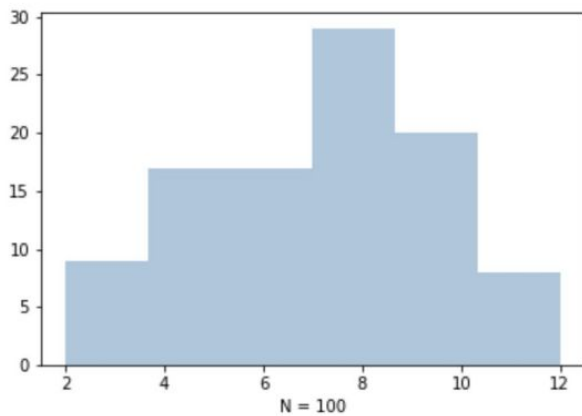
Misal: Mean, Kuartil Bawah, Kuartil tengah (Median), Kuartil Atas, dll.



UKURAN PENYEBARAN DATA

Statistik yang memberikan informasi **bagaimana data menyebar** di sekitar pusat data.

Misal: Range, Variansi, Standar Deviasi, dll.



Karakter Distribusi

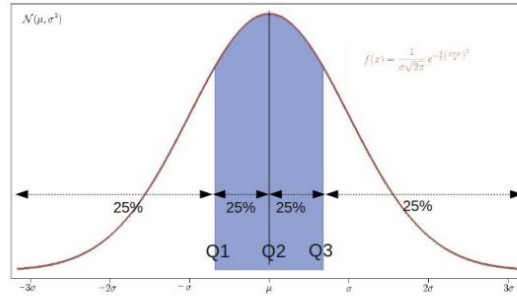


BENTUK DISTRIBUSI

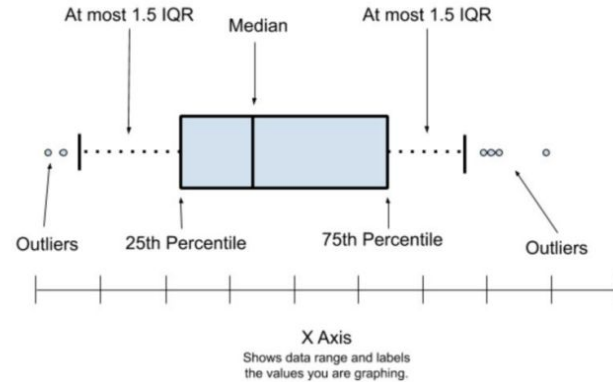
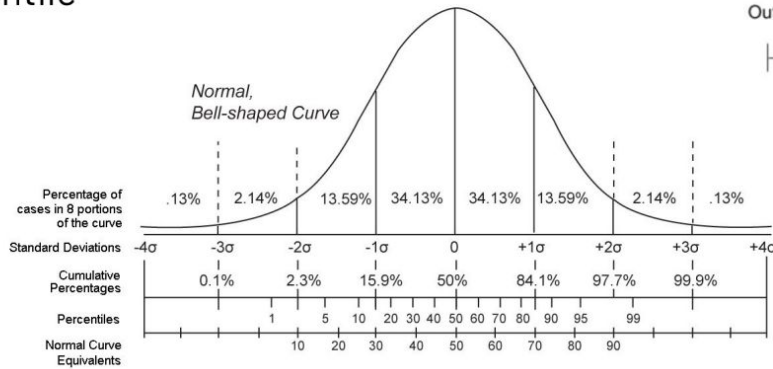
SIMETRIS	↔	Mean = Median
RIGHT SKEWED/SKEW POSITIF	↔	Mean > Median
LEFT SKEWED/SKEW NEGATIF	↔	Mean < Median
PUNCAK	↔	Berpuncak Tunggal (Modus = 1), Berpuncak Jamak (Modus > 1)



Quantile

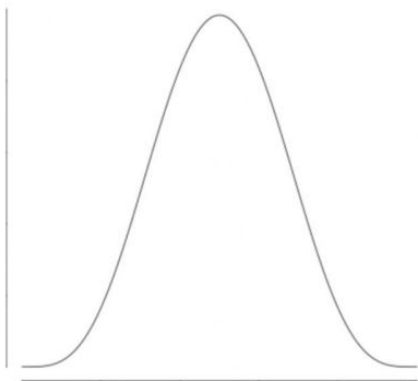


Percentile

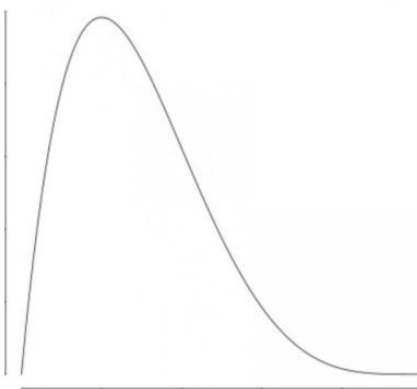




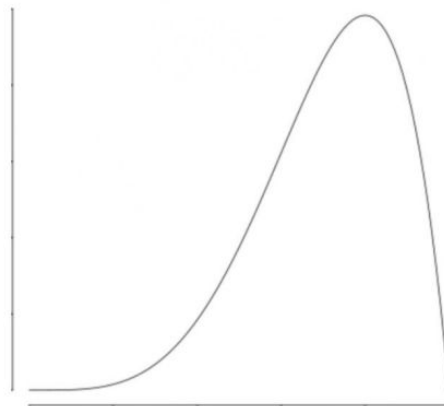
No Skew

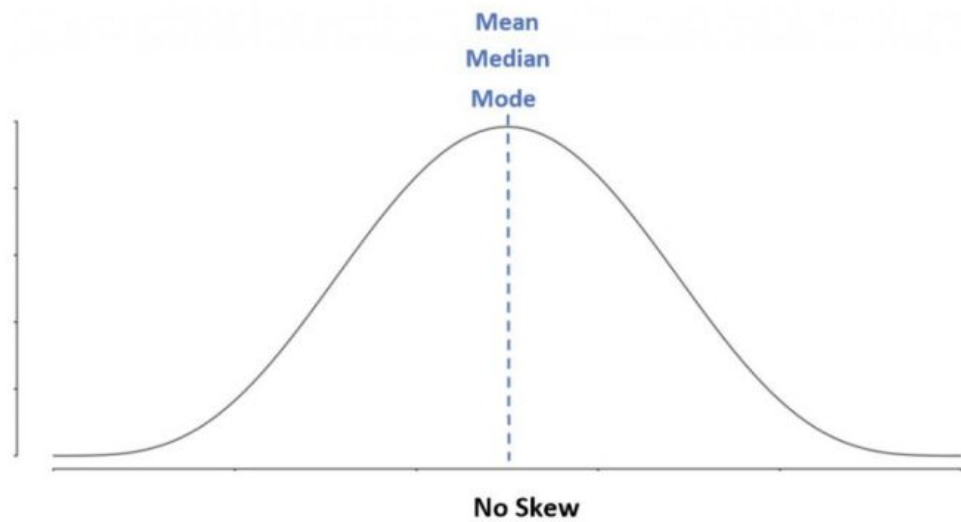


Right Skewed Distribution



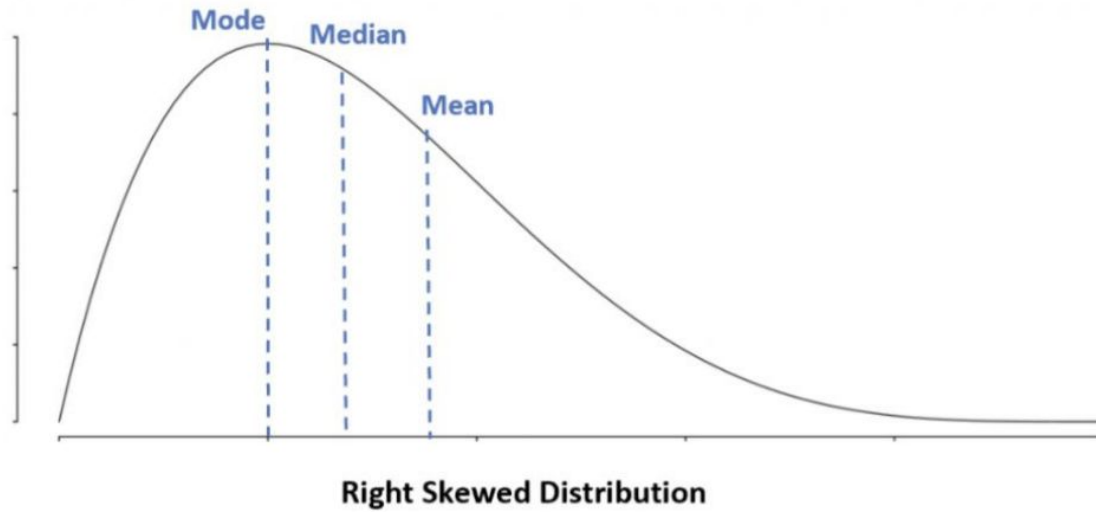
Left Skewed Distribution



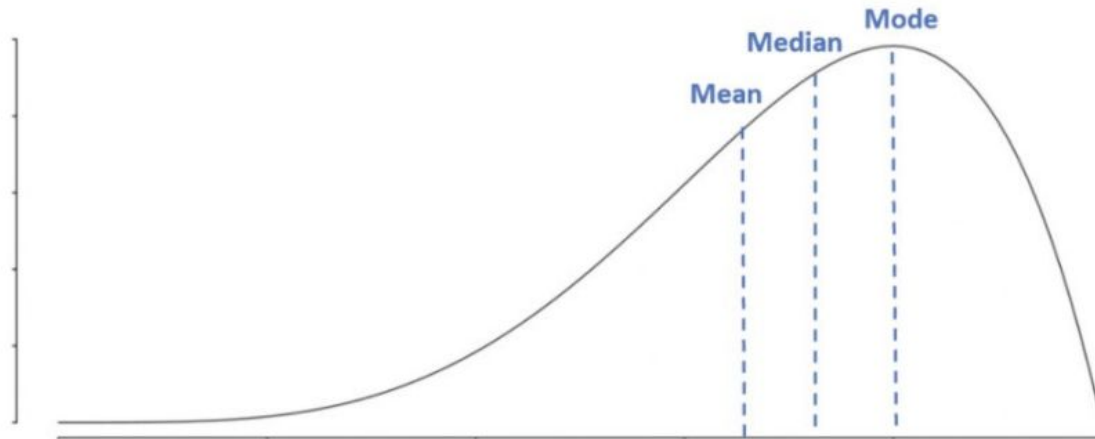


Distribusi simetris: mean = median = mode





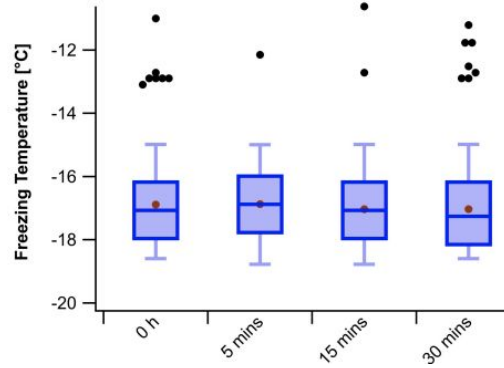
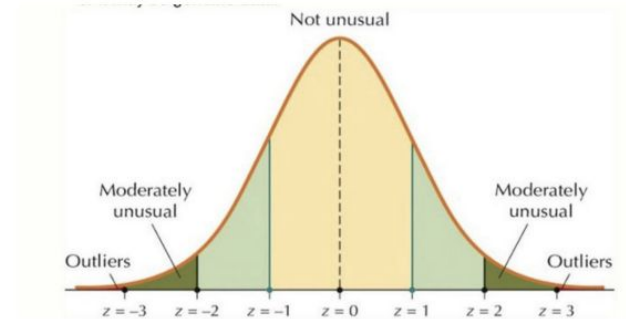
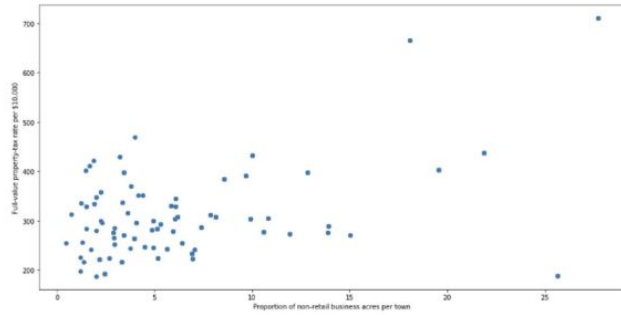
Distribusi menceng kanan: $\text{mode} < \text{median} < \text{mean}$



Left Skewed Distribution

Distribusi menceng kiri: $\text{mean} < \text{median} < \text{mode}$

Outlier/Pencilan



Outlier/Pencilan

Penyebab paling umum dari outlier pada kumpulan data:

- Kesalahan entri data (kesalahan manusia)
- Kesalahan pengukuran (kesalahan instrumen)
- Kesalahan eksperimental (ekstraksi data atau kesalahan perencanaan/pelaksanaan percobaan)
- Disengaja (pencilan tiruan dibuat untuk menguji metode deteksi)
- Kesalahan pemrosesan data (manipulasi data atau mutasi kumpulan data yang tidak diinginkan)
- Kesalahan pengambilan sampel (mengambil atau mencampur data dari sumber yang salah atau beragam)
- Alami (bukan kesalahan, hal baru dalam data)

Contoh

Data banyaknya pelanggan yang beli gorengan di warung A di 15 hari terakhir pada bulan September 2020:

26 37 76 49 95 69 83 87 39 95 59 83 83 87 46

Setelah diurutkan,

26 37 39 46 49 59 69 76 83 83 83 87 87 95 95

1. MEAN

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 67.60$$

2. MEDIAN

(Nilai tengah yang membagi dua kelompok data sama banyak)

$$\text{Med} = x(8) = 76$$

3. MODUS

(Nilai yang paling sering muncul)

$$\text{Mod} = 83$$

4. KUARTIL

$$\text{Kuartil bawah (q1)} = x((n+1)/4) = x((15+1)/4) = x(4) = 46$$

$$\text{Kuartil tengah (q2)} = x(2*(n+1)/4) = x(2*(15+1)/4) = x(8) = 76$$

$$\text{Kuartil atas (q3)} = x(3*(n+1)/4) = x(3*(15+1)/4) = x(12) = 87$$

Contoh

Data banyaknya pelanggan yang beli gorengan di warung A di 15 hari terakhir pada bulan September 2020:

26 37 76 49 95 69 83 87 39 95 59 83 83 87 46

Setelah diurutkan,

26 37 39 46 49 59 69 76 83 83 83 87 87 95 95

5. RANGE (JANGKAUAN DATA)

$$R = \text{data max} - \text{data min} = 95 - 26 = 69$$

7. STANDAR DEVIASI (SIMPANGAN BAKU)

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = 23,01$$

6. VARIANSI

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = 529,2571$$

8. JANGKAUAN ANTAR KUARTIL

$$dq = q_3 - q_1 = 87 - 46 = 41$$

Contoh

Data banyaknya pelanggan yang beli gorengan di warung A di 15 hari terakhir pada bulan September 2020:

26 37 76 49 95 69 83 87 39 95 59 83 83 87 46

Setelah diurutkan,

26 37 39 46 49 59 69 76 83 83 83 87 87 95 95

9. DATA PENCILAN

Data yang nilainya berbeda jauh dari kelompok data yang lain.

Langkah-langkah:

1. Hitung $dq \rightarrow dq = 41$
2. Hitung Batas Bawah Pencilan (BBP) = $q1 - k.dq = 46 - (1,5)(41) = -15,5$.
Pilih $k=1,5$.
3. Hitung Batas Atas Pencilan (BAP) = $q3 + k.dq = 87 + (1,5)(41) = 148,5$
4. Pencilan Bawah < BBP -> **Tidak ada pencilan bawah**
5. Pencilan atas > BAP -> **Tidak ada pencilan atas**

Uji Hipotesis dan Regresi

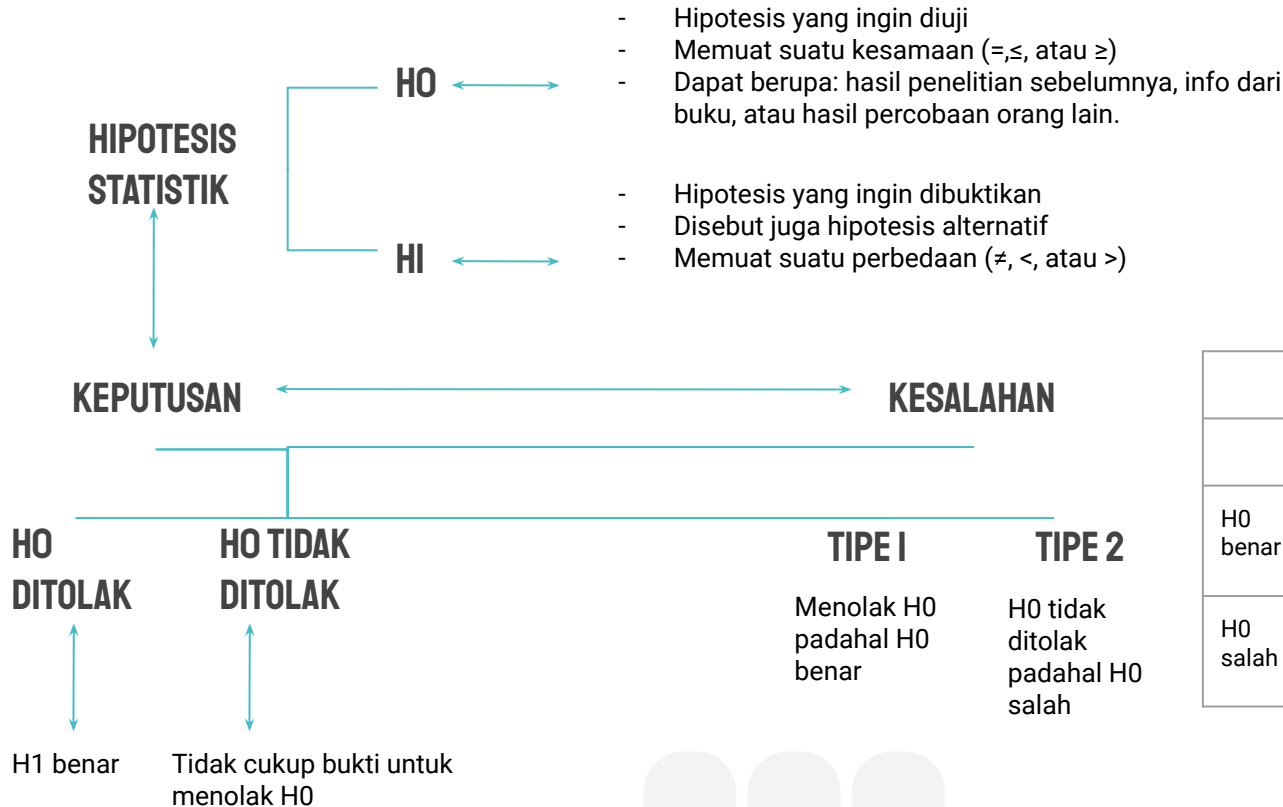
Definisi Uji Hipotesis

- Uji hipotesis adalah metode statistik **yang digunakan dalam pengambilan keputusan statistik** dengan menggunakan data eksperimen.
- pada dasarnya, uji hipotesis merupakan asumsi yang kita buat tentang parameter populasi.

Misal: rata-rata nilai Kalkulus di Kelas X adalah 80.

Untuk membuktikannya, kita butuh metode statistika. Maka, jika kalian sering mendengar “statistically significant”, itu adalah berkat uji hipotesis.

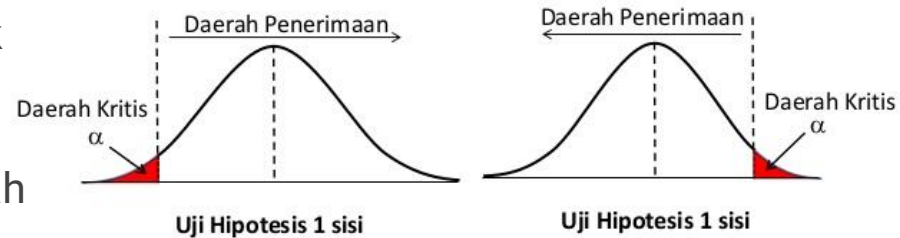
Uji Hipotesis



	Keputusan	
	H ₀ ditolak	H ₀ tidak ditolak
H ₀ benar	Kesalahan tipe I (False Positive)	Benar
H ₀ salah	Benar	Kesalahan Tipe II (False Negative)

Statistik Uji dan Titik Kritis

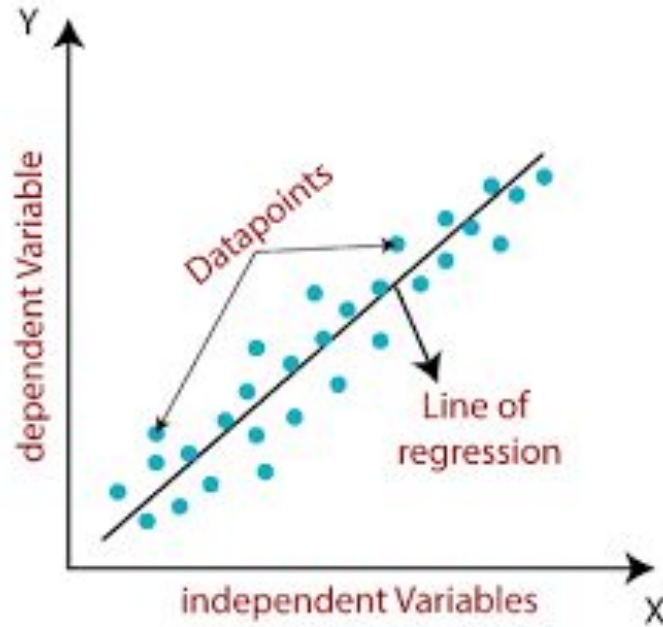
- Statistik uji digunakan untuk menguji hipotesis statistik yang telah dirumuskan. Notasinya berpadanan dengan jenis distribusi yang digunakan.
- Titik kritis membatasi daerah penolakan dan penerimaan H_0 . Diperoleh dari tabel statistik yang bersangkutan.
- H_0 ditolak jika nilai statistik uji jatuh di daerah kritis.



Tujuan Regresi

- Menentukan/menaksir parameter-parameter yang terlibat dalam suatu model matematik yang linear terhadap parameter-parameter tersebut.
- Melakukan prediksi terhadap nilai suatu variabel, misalkan Y , berdasarkan nilai variabel yang lain, misalkan X , dengan menggunakan model regresi linier (interpolasi)

Regression



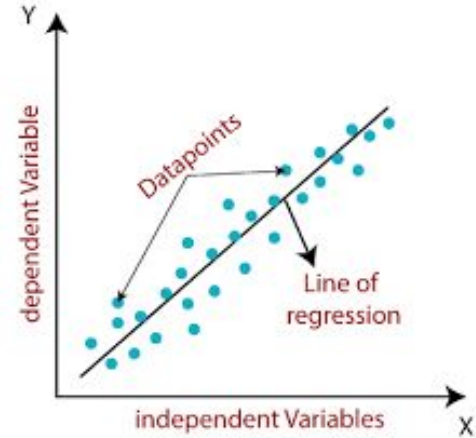
Model Regresi

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

- β_1 dan β_0 merupakan parameter-parameter model yang akan ditaksir
- e_i adalah galat pada observasi ke- i (acak)

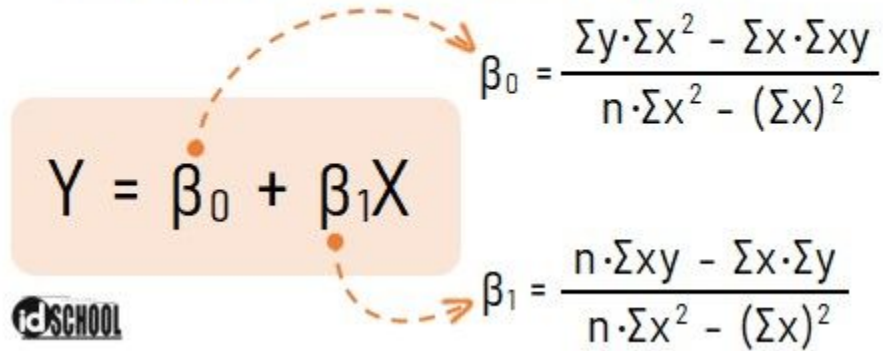
Sumber galat:

- Ketidakmampuan model regresi dalam memodelkan hubungan prediktor dan respon dengan tepat.
- Ketidakmampuan peneliti dalam melakukan pengukuran dengan tepat
- Ketidakmampuan model untuk melibatkan semua variabel prediktor.



Formula Regresi

Rumus Hitung
Persamaan Regresi Linear Sederhana



The diagram illustrates the relationship between the simple linear regression equation and the formulas for its coefficients. On the left, the equation $Y = \beta_0 + \beta_1 X$ is enclosed in an orange rounded rectangle. A dashed orange arrow originates from the β_0 term and points to its corresponding formula on the right. Another dashed orange arrow originates from the $\beta_1 X$ term and points to its corresponding formula. Below the equation box is the 'idSCHOOL' logo.

$$Y = \beta_0 + \beta_1 X$$
$$\beta_0 = \frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma xy}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$
$$\beta_1 = \frac{n \cdot \Sigma xy - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$

Study Case

<https://docs.google.com/spreadsheets/d/1lvKJIFFxR6xaQeQL8nVhOeeSq5tl-DvIGTcgWR-GA60/edit#gid=2117411496>

MySkill

#RintisKarirImpian

Thank you!