Rakamin Academy

Data Scientist
at Telecommunication Company

Halo!

Perkenalkan saya **Abdul**lah Ghifari.

Sebelumnya

bukalapak

tiket.com

Institut Pertanian Bogor

Abdullah Ghifari
https://www.linkedin.com/in/abdullah-ghifari/

Rakamin Academy

Statistic

Correlation & Distribution

Objektif

Memahami dan dapat melakukan pencarian Correlation, Causality, metode sampling, Probability dan Distribution dan kapan menerapkannya,

Expected Output

Memahami apa itu **Correlation dan Causality**, dan apa perbedaan antara keduanya
Mampu menerapkan **Correlation dan Causality**, dan kapan menerapkannya
Memahami apa itu **Probability dan Distribution**, beserta prinsipnya
Mampu menerapkan **Probability dan Distribution**, dan kapan menerapkannya
Memahami jenis jenis **Metode Sampling** (Probability dan Non-Probability Sampling)
Mampu menerapkan berbagai **Metode Sampling** (Probability dan Non-Probability Sampling), dan kapan menerapkannya

Hands-On Required :

Hands - On :

1. Statistics II - Correlation and Distribution.ipynb
2. Statistics II - Sampling.ipynb

Dataset :

1. HR_comma_sep.csv
2. Iris.csv
3. ab_data.csv

Klik disini untuk mengakses folder Hands-On dan Dataset

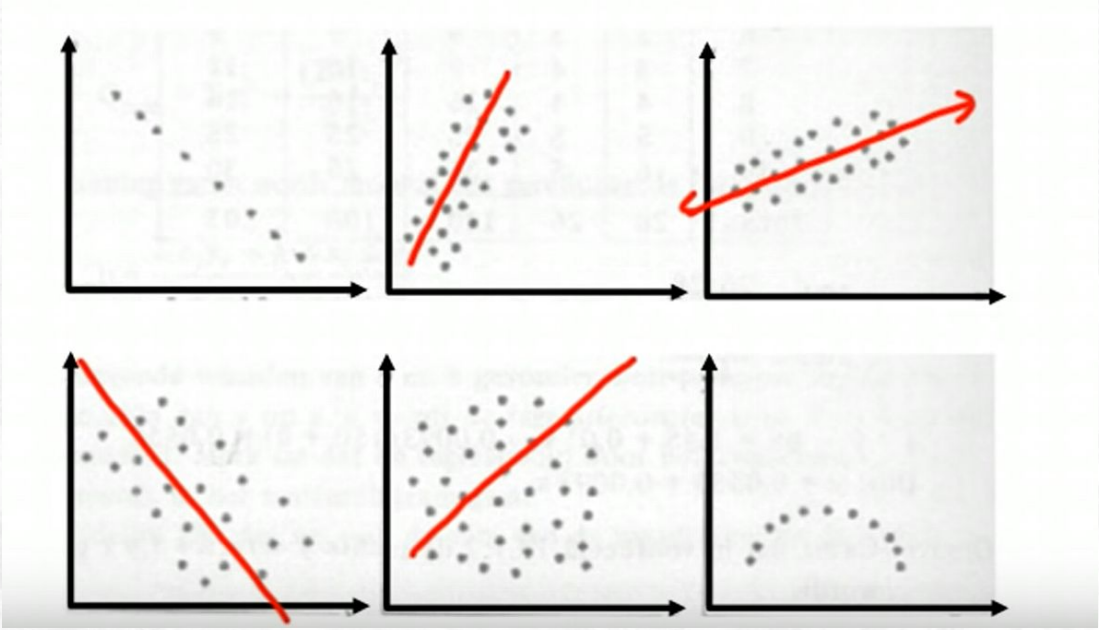
Outline Pembelajaran

Rakamin Academy

Correlation & Distribution

- Correlation & Causality
- Probability & Distribution
- Sampling Method
- Hands-On

Hubungan 2 Variabel



Rakamin Academy

Korelasi

Korelasi adalah ukuran statistik (dinyatakan sebagai angka) yang menggambarkan **ukuran dan arah hubungan antara dua atau lebih variabel**.

Namun, korelasi antar variabel **tidak** secara otomatis berarti bahwa perubahan dalam satu variabel adalah **penyebab** dari perubahan nilai-nilai variabel lain.

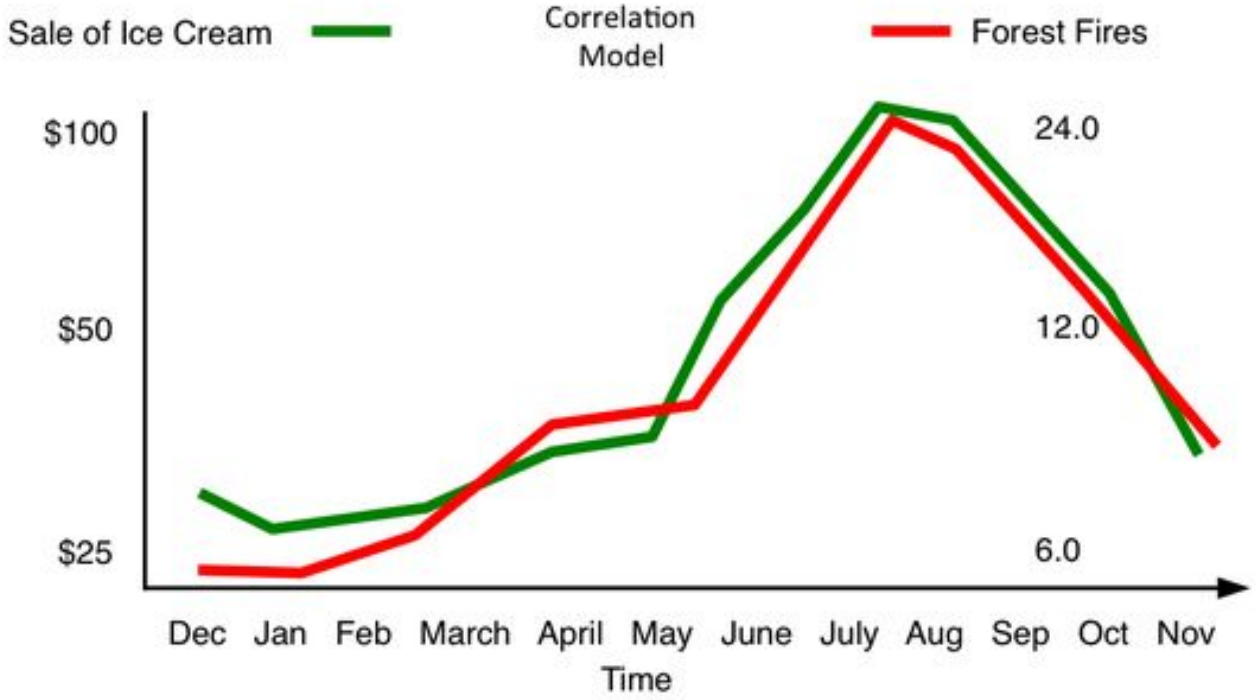
Kausalitas

Kausalitas menunjukkan bahwa satu peristiwa adalah **hasil dari** terjadinya peristiwa lainnya;

Yaitu ada hubungan **sebab akibat** antara kedua peristiwa tersebut. Ini juga disebut sebagai sebab dan akibat.

Korelasi != Kausalitas

Korelasi vs Kausalitas



https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect

Mengapa Mengetahui Korelasi dan Kausalitas itu penting?

Untuk mengidentifikasi sejauh mana **satu variabel terkait dengan variabel lain**.
Sebagai contoh:

1. Apakah ada hubungan antara tingkat pendidikan seseorang dan kesehatannya?
2. Apakah memelihara hewan peliharaan menyebabkan hidup lebih lama?
3. Apakah marketing campaign sebuah perusahaan mampu meningkatkan penjualan produk mereka?

Mengapa Mengetahui Korelasi dan Kausalitas itu penting?

Jika ada korelasi, maka ini dapat menjadi **panduan penelitian lebih lanjut** untuk menyelidiki apakah satu tindakan menyebabkan yang lain.

Dengan memahami korelasi dan kausalitas, memungkinkan kebijakan dan program yang bertujuan untuk mencapai hasil yang diinginkan agar **lebih tepat sasaran.**

Bagaimana Menghitung Korelasi?

Korelasi Pearson

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

Bagaimana Menghitung Korelasi?

Untuk dua variabel, korelasi statistik diukur dengan menggunakan Koefisien Korelasi, diwakili oleh simbol (r)

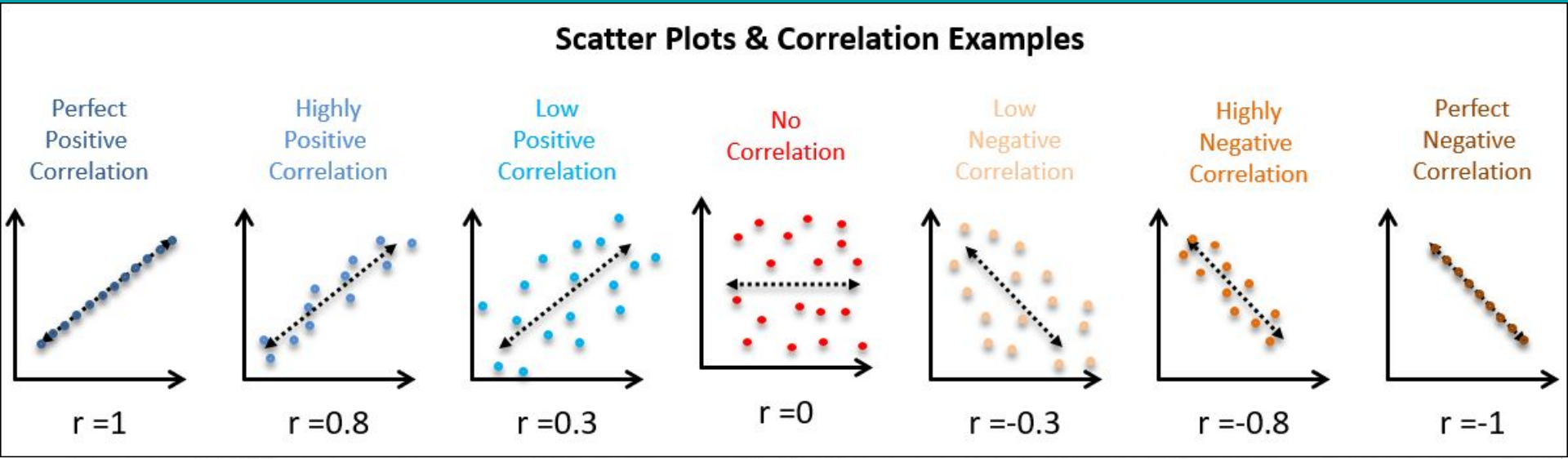
Nilai numerik koefisien berkisar dari +1.0 hingga -1.0, yang memberikan indikasi kekuatan dan arah hubungan.

Bagaimana Menghitung Korelasi?

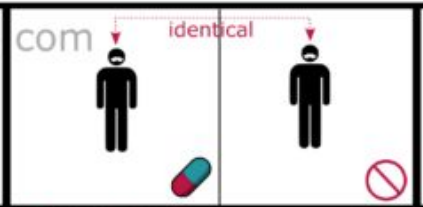
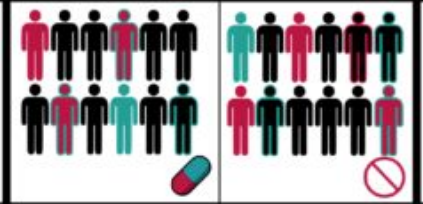
Jika koefisien korelasi memiliki nilai negatif (di bawah 0) itu menunjukkan hubungan negatif antara variabel. Semakin besar variabel A maka variabel B akan semakin kecil.

Jika koefisien korelasi memiliki nilai positif (di atas 0) itu menunjukkan hubungan positif antara variabel. Semakin besar variabel A maka variabel B akan semakin besar.

Gambaran Mudah terkait Korelasi



Bagaimana Mengukur Kausalitas?

Experiment		Control and treatment are **identical** and their behavior is **deterministic**. Causal effect of treatment is directly the difference between observations for the two groups. *Physics, Biology, ~~Social sciences~~*
Statistical Experiment		Control and treatment are not identical but divided at random. This makes it possible to build a precise estimate of the causal effect of treatment. *A/B testing, Central Limit Theorem, Bayesian Statistics*
Quasi-experiment		Control and treatment are not identical and divided by a "natural" criterion. Depending on "internal" and "external" quality of the criterion, it is possible to build a good estimate of the causal effect of treatment. *Differences-in-differences, Regression Discontinuity, Instrumental variables, Matching, Controlled Regression*
Counterfactuals		Control group does not exist, instead its behaviour is estimated with a predictive model of what would have happened without the treatment (= counterfactual). *Synthetic Differences-in-Differences, Athey & Imbens, CausalImpact*

Stronger evidence

Outline Pembelajaran

Rakamin Academy

Correlation & Distribution

- Correlation & Causality
- Probability & Distribution
- Sampling Method
- Hands-On

Probabilitas

Probabilitas mempelajari **keacakan**

Kita tahu semua kemungkinan yang akan terjadi,
tetapi kami tidak tahu hasil apa yang akan terjadi

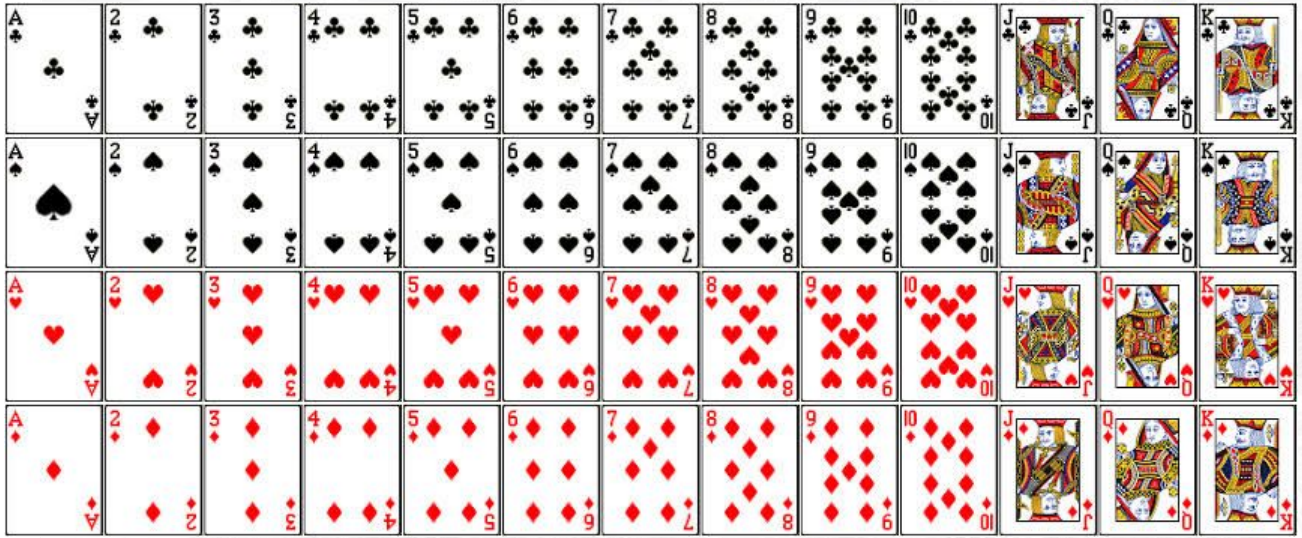
Probabilitas

Probabilitas adalah deskripsi numerik tentang seberapa besar kemungkinan suatu peristiwa terjadi atau seberapa besar kemungkinan proposisi itu benar.

Biasanya dilambangkan dengan P (event)
P (A): Probabilitas kejadian A

P(A) = kemungkinan kejadian A / semua kemungkinan

Probabilitas - Quiz 1

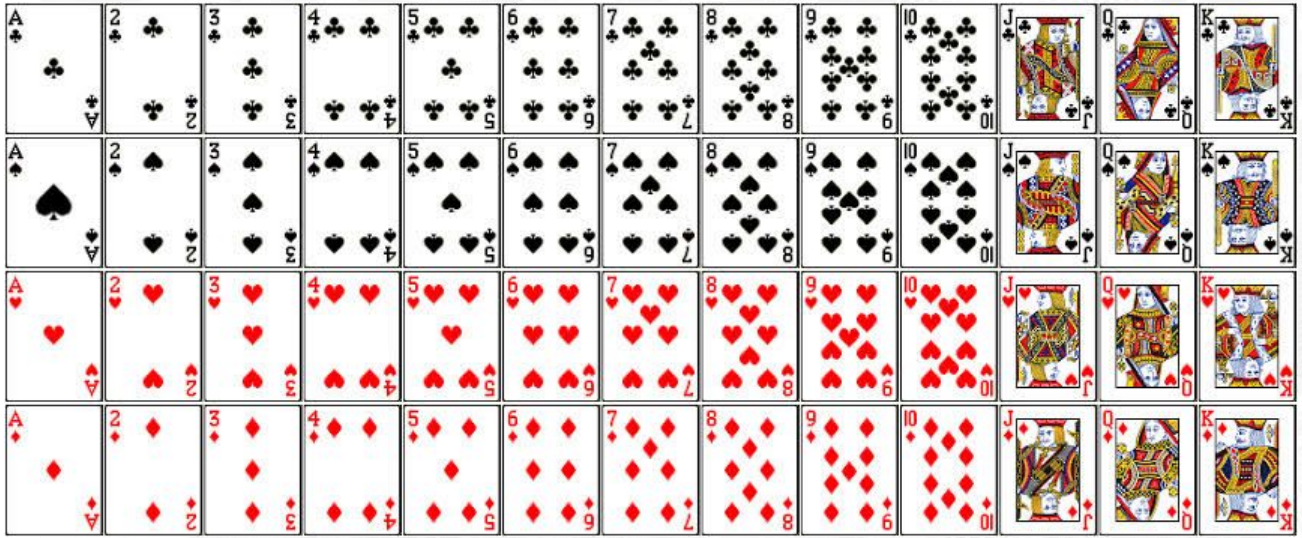


P(Queen) = ?

P(Red) = ?

P(3) = ?

Probabilitas - Quiz 1

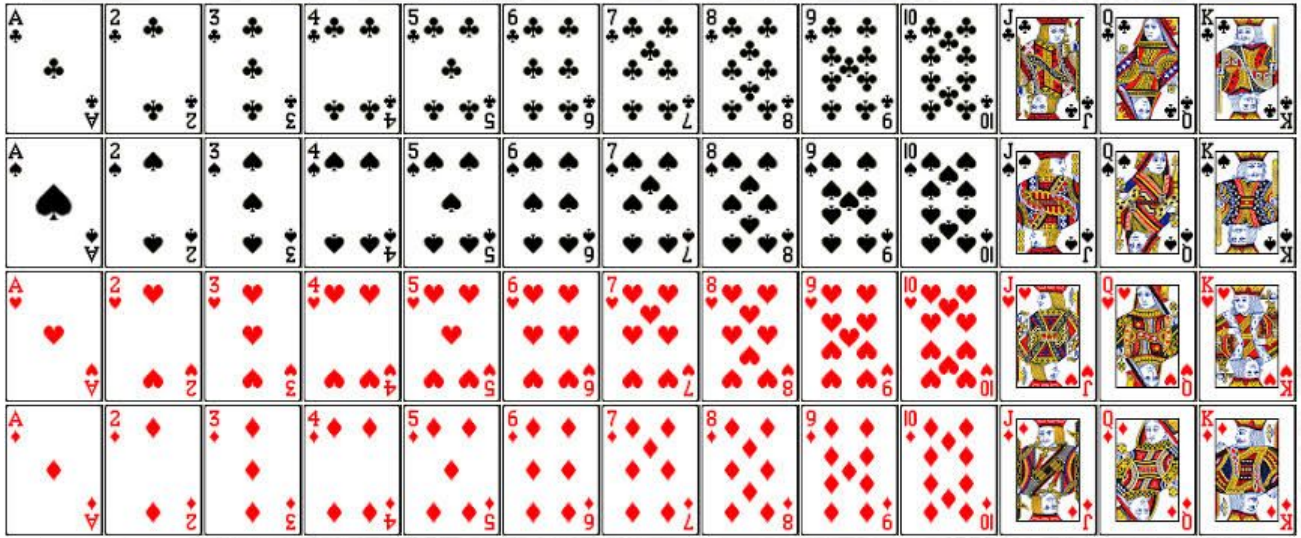


P(Queen) = 4/52

P(Red) = 26/52

P(3) = 4/52

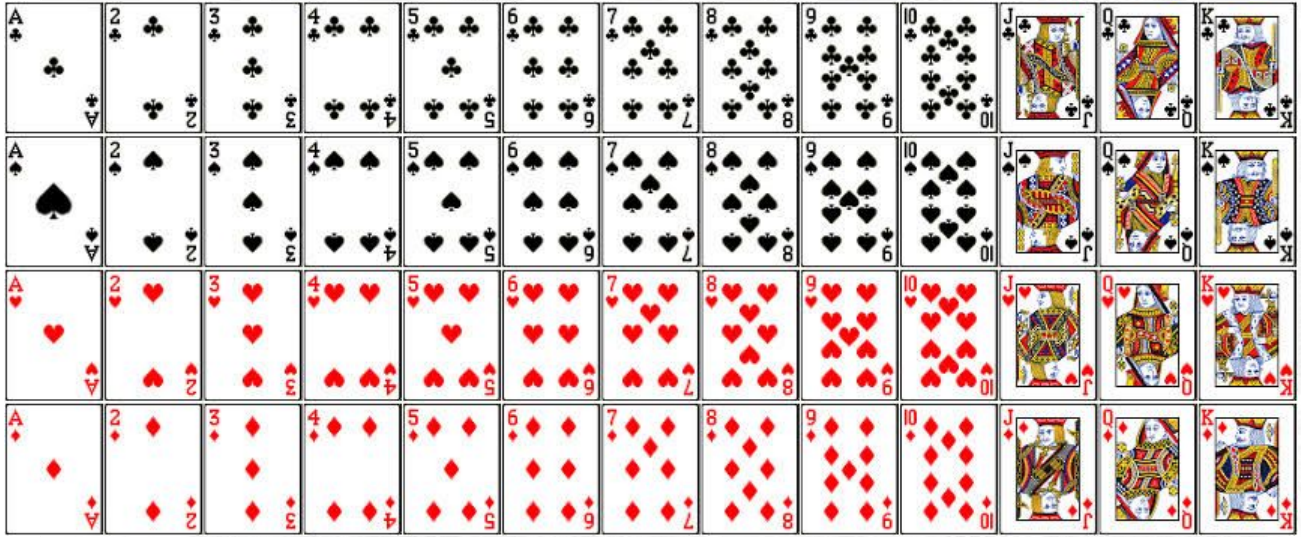
Probabilitas - Quiz 2



P(Queen or 3) = ?

P(Queen or Red) = ?

Probabilitas - Quiz 2



P(Queen or 3) = 8/52

P(Queen or Red) = 28/52

Probabilitas - Quiz 3

Rolling 1 dadu:

- Ahmad akan mendapatkan $ 1 jika # dari titik kurang dari atau sama dengan 3
- Billy akan mendapatkan $ 2 jika # titik kurang dari atau sama dengan 2

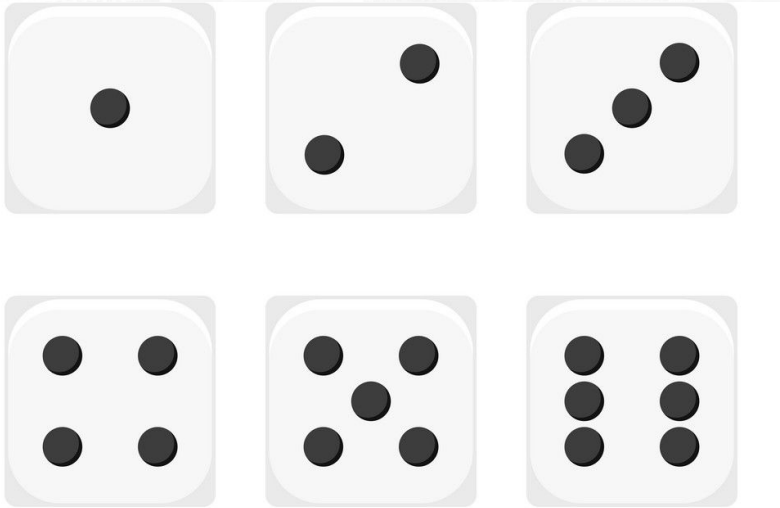
Anda ingin menjadi siapa: Ahmad atau Billy?
Jika kita melemparkan 100 kali, kamu ingin jadi siapa?

Probabilitas - Quiz 3

P(Ahmad) = 3/6 = ½ = 50%
P(Billy) = 2/6 = 1/3 = 33%

Ekspektasi yang didapat :
E(Ahmad) = 1/2 * 1 = 0.5 $
E(Billy) = 1/3 * 2 = 0.66 $

Ekspektasi 100 percobaan:
Ahmad = 100*0.5 = 50$
Billy = 100*0.66 = 66$



Sebaran Peluang

Semua kemungkinan hasil dan probabilitas yang terjadi

two-tosses	head-head	head-tail	tail-head	tail-tail
probability	0.25	0.25	0.25	0.25

Aturan:

1. Apabila semua ditambahkan hasilnya = 1
2. Harus disjoint (saling lepas)
3. Masing-masing harus lebih dari 0

Probabilitas - Quiz 4

Diketahui 5 pengguna yang sering menggunakan e-commerce apps.
Namun mereka menggunakan apps yang berbeda-beda:

1. Orang #1 : Toqpedia
2. Orang #2 : Toqpedia dan Syohee
3. Orang #3 : Toqpedia dan Syohee
4. Orang #4 : Syohee
5. Orang #5 : Syohee

Buatlah sebaran peluang (probability distribution) dari data berikut.

Probabilitas - Quiz 4

Cara 1

E-commerce	Total	Perc
Toqpedia	3	60%
Syohee	4	80%
Total	**7**	**140%**

Cara 2

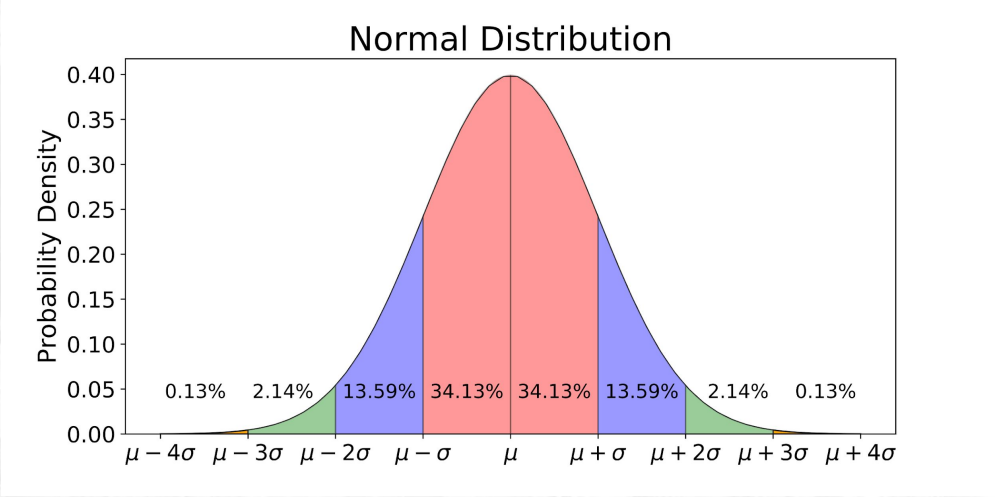
E-commerce	Total	Perc
Toqpedia	1	20%
Syohee	2	40%
Toqpedia & Syohee	2	40%
Total	**5**	**100%**

Dalam pembentukan probability jangan sampai terdapat elemen yang redundan sehingga lebih baik dipisah untuk dualist dan solus.

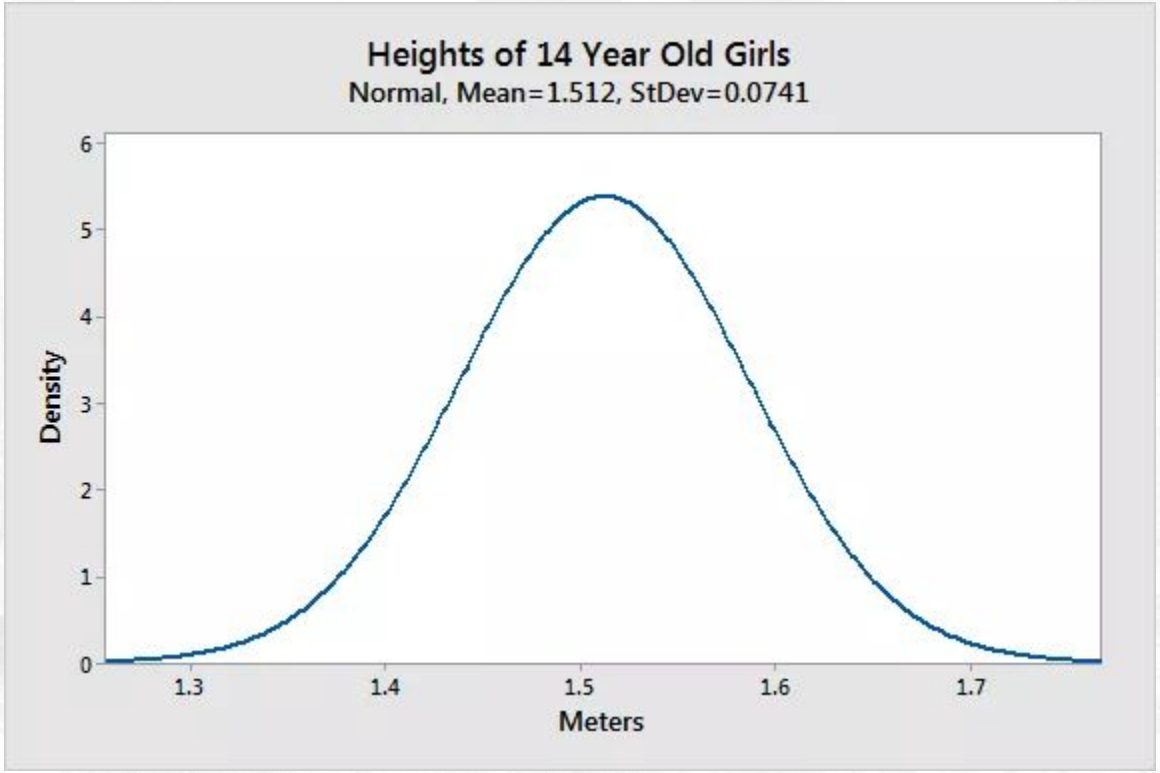
Beberapa Sebaran yang Umum Diketahui

Normal - Sebaran Kontinu

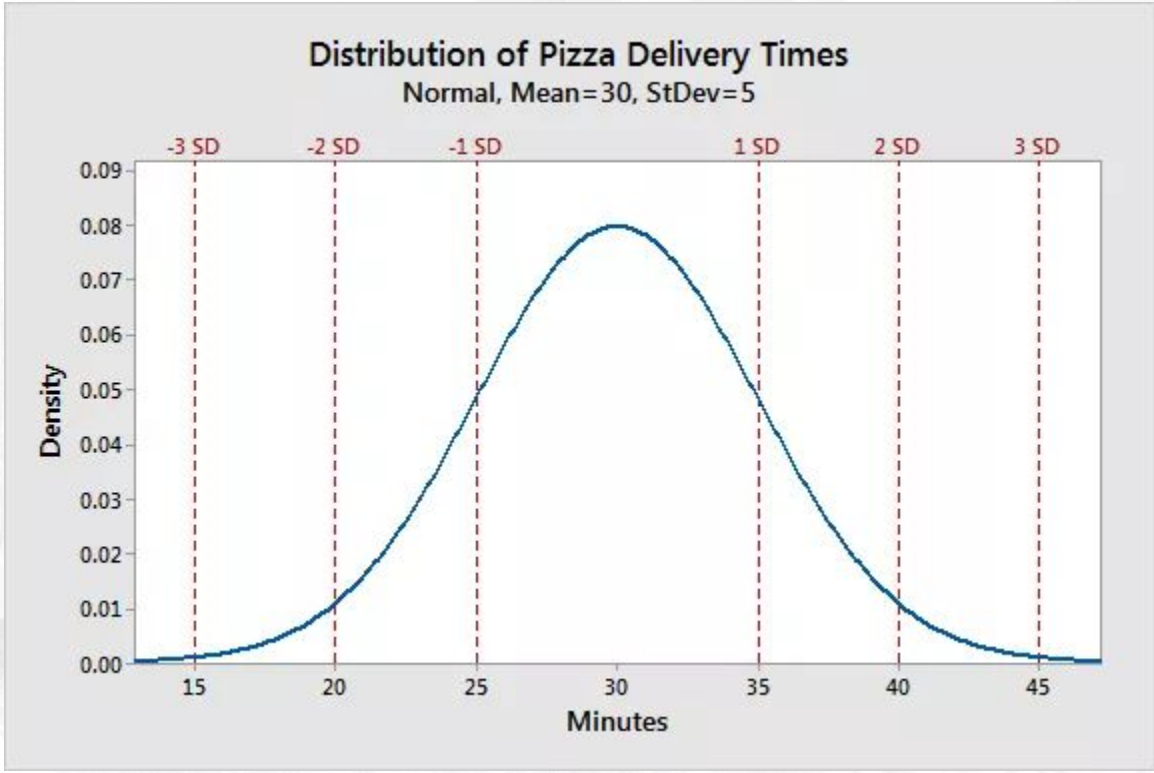
Distribusi normal adalah distribusi peluang kontinu yang paling penting dalam statistika



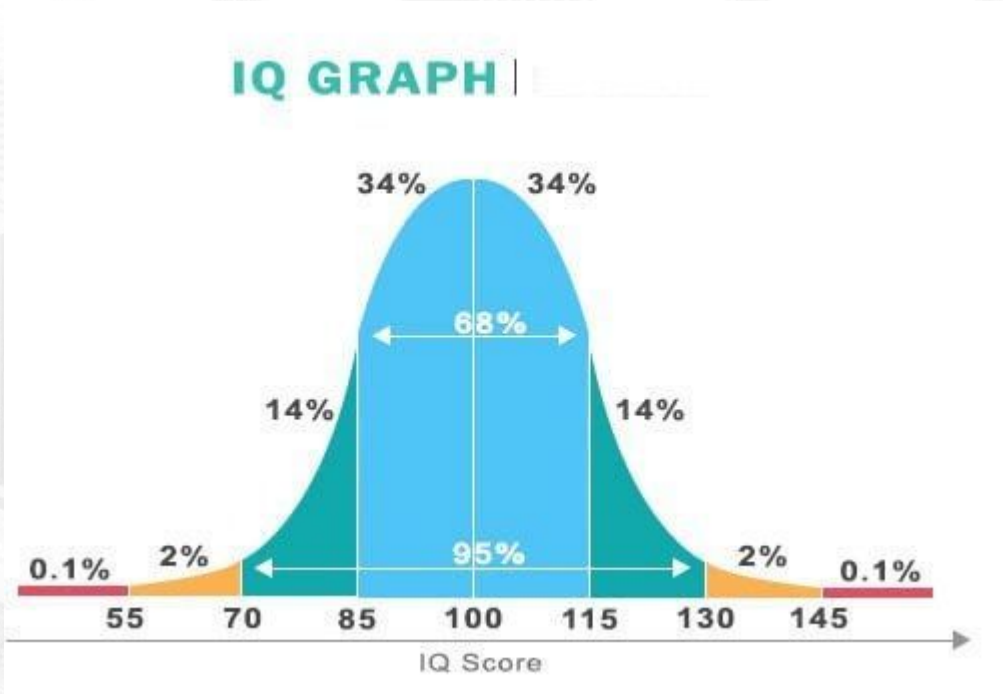
Contoh Sebaran Normal



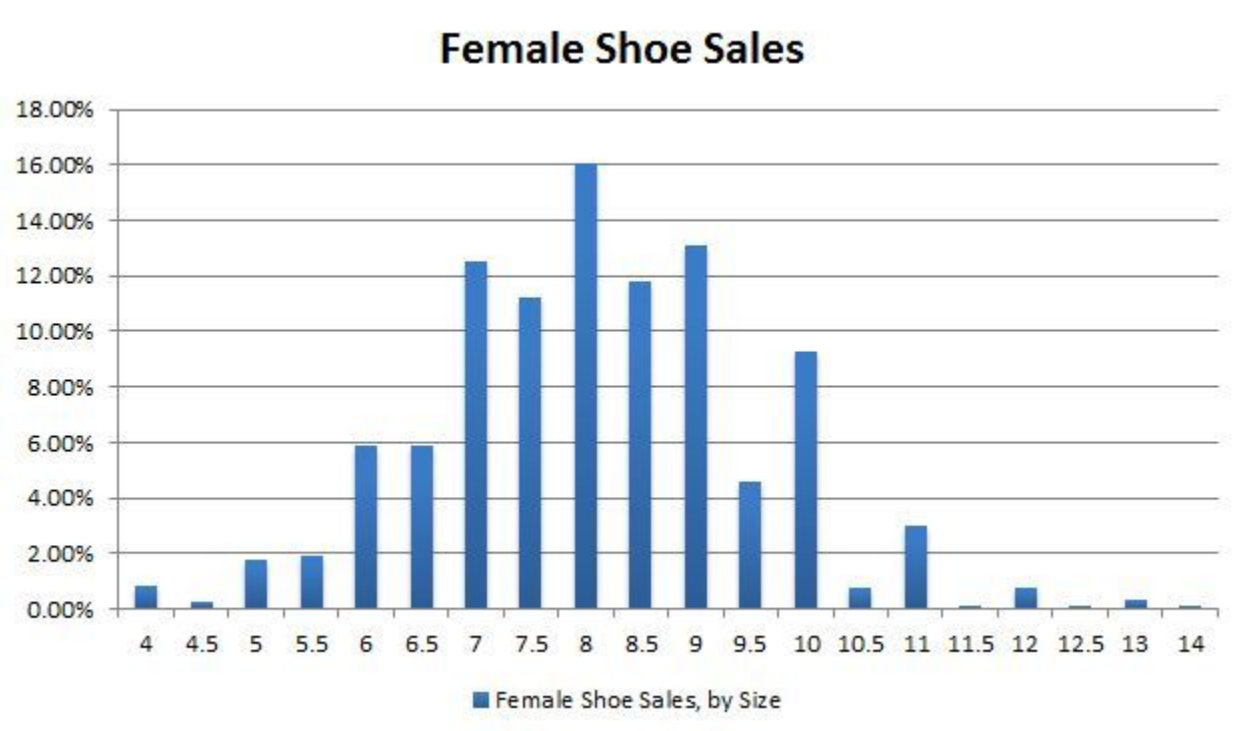
Contoh Sebaran Normal



Contoh Sebaran Normal

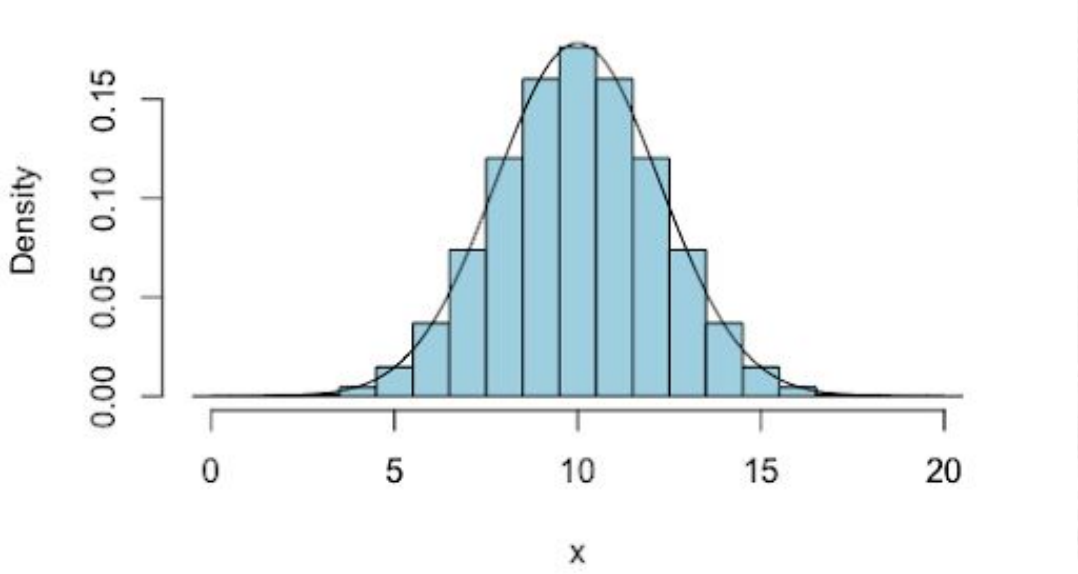


Contoh Sebaran Normal



Asal Muasal Sebaran Normal

Distribusi binomial yang sangat banyak akan membentuk sebaran kontinu yaitu sebaran normal



Normal - Sebaran Kontinu

Suatu Variabel Random berdistribusi normal ditulis **$X \sim N(\mu,\sigma^2)$**

Distribusi ini bergantung pada dua parameter yaitu *mean* **(μ)** dan *variance* **(σ^2)**

Fungsi kepadatan peluang

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Kondisi :

1. Data yang kita punya: $-\infty < x < \infty$
2. Rata-rata: $-\infty < \mu < \infty$
3. Variance: $\sigma^2 \geq 0$

Karakteristik Penting dari Sebaran Normal

1. Bentuknya simetris (symmetric)
2. Grafiknya menyerupai bentuk lonceng
3. Syarat asumsi dari mayoritas ilmu statistika
4. Pada case tertentu mampu meningkatkan performa model machine learning.

Outline Pembelajaran

Rakamin Academy

Correlation & Distribution

- Correlation & Causality
- Probability & Distribution
- Sampling Method
- Hands-On

Mengapa sih kita perlu sampling?

1. Data tidak tersedia
2. **Cost** yang dibutuhkan sangat besar jika mengambil semua populasi
3. Mengambil data dari semua populasi akan memakan **waktu** yang dibutuhkan sangat lama
4. **Biaya komputasi** akan sangat besar jika seluruh data populasi sangat banyak
5. Dengan mengambil sampling kita bisa menggunakan data yang sedikit dan dapat **mengestimasi nilai parameter**.

Tipe Metode Sampling

1. Probability Sampling
2. Non-Probability Sampling

Apa itu Probability Sampling?

Probability sampling adalah teknik pengambilan sampel dengan menggunakan **peluang** untuk setiap sampelnya.

Sehingga setiap sampel **memiliki kesempatan yang sama** untuk terambil.

Apa itu Non-Probability Sampling?

Probability sampling adalah teknik pengambilan sampel tidak menggunakan metode peluang.

Bisa disederhanakan sebagai sampel diambil secara **kebetulan**.

Ilustrasi Sampling Method

Di suatu negara, ada negara bagian yang ingin **memisahkan diri** dari negara utamanya.

Salah satu acara talkshow di negara tersebut mencoba untuk **mengumpulkan suara** dari warga di wilayah tersebut. Acara tersebut **menggunakan 2 metode**,

Random: mengambil sample dengan memilih secara random dari buku telepon
Voluntary: dicantumkan nomor telepon televisi dan siapapun bisa menyuarakan suaranya.

Kemudian diperoleh data sebagai berikut:

Technique	Participant	Choose	Percentage
Random	5,000	1,500	30%
Voluntary	30,000	21,000	70%

Tipe-Tipe Kesalahan pada Sampling

Sampling Error

kesalahan yang ditimbulkan karena kita hanya mengamati sebagian saja (contoh), tidak semuanya (populasi). Dapat dikontrol dengan secara hati-hati mendesain penarikan contohnya.

Misalkan: Dalam penarikan sample terdapat ***margin of error*** yang menandakan kesalahan yang diprediksi

Non-Sampling Error

kesalahan yang ditimbulkan karena sebab lain pada proses survei, dan lebih sulit dikendalikan.

Misalkan: Ketika ingin mengambil **sample ke sebuah rumah**, namun ternyata tidak ada orang di dalam rumah tersebut.

Ukuran Sampel

Slovin :

$$n = \frac{N}{1 + Ne^2}$$

N : Number of Population
e : Margin of Error

*margin of error yang biasa digunakan adalah 5%

Contoh : Quick Count

Lembaga Survey yang ada di Indonesia ingin melakukan survey terkait Pemilu Tahun 2019. Diketahui jumlah total populasi di Indonesia sebanyak 200 juta orang. Dengan margin of error sebesar 0.5%, berapakah jumlah sampel yang dibutuhkan?

Answer: Quick Count

Lembaga Survey yang ada di Indonesia ingin melakukan survey terkait Pemilu Tahun 2019. Diketahui jumlah total populasi di Indonesia sebanyak 200 juta orang. Dengan margin of error sebesar 0.5%, berapakah jumlah sampel yang dibutuhkan?

Outline Pembelajaran

Rakamin Academy

Correlation & Distribution

- Correlation & Causality
- Probability & Distribution
- Sampling Method
- Hands-On

Rakamin Academy

Terima Kasih!