Capaian Pembelajaran

Topik yang akan disampaikan adalah beberapa konsep dasar Statistika:

- Peluang (probability)
- Variabel acak (random variable)
- Populasi dan sampel
- Jenis fitur data
- Statistika deskriptif
- Statistika infererens

Pengertian Peluang (*Probability*)

- Teori peluang (*probability theory*) adalah cabang ilmu statistika untuk melakukan penalaran mengenai peluang (*probability*) sebuah kejadian yang terjadi secara acak (random).
- Eksperimen adalah sebuah proses yang menghasilkan satu dari sejumlah kemungkinan keluaran yang bersifat acak (random).
  - Melemparkan sebuah mata uang logam. Keluaran adalah sisi H atau T.
  - Melemparkan sebuah dadu bersisi enam. Keluaran adalah salah satu dari sisi 1, 2, 3, 4, 5, 6
- Kejadian dari eksperimen adalah himpunan bagian dari seluruh himpunan keluaran.
  - A: kejadian keluaran dari eksperimen melempar sebuah dadu yang merupakan bilangan genap atau A = $\{2, 4, 6\}$.
  - B: kejadian keluaran dari eksperimen melempar sebuah dadu yang merupakan bilangan ≤ 4 atau B = $\{1, 2, 3, 4\}$.

Ketidakpastian dan Metode Statistika

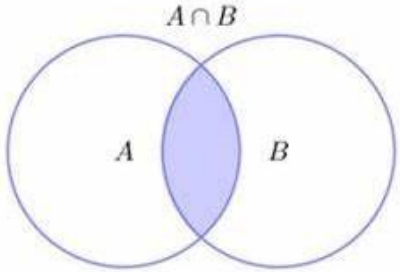
- Data input dari proses data science dapat mengandung ketidakpastian/error karena berbagai faktor:
  - Kesalahan dalam pembacaan alat ukur/sensor.
  - Kesalahan dari alat ukur/sensor
  - Kesalahan di dalam transportasi/transmisi data dari sensor ke komputer pengumpulan data atau diantara computer, dll.
- Statistika adalah cabang ilmu Matematika untuk menangani ketidakpastian di dalam data.

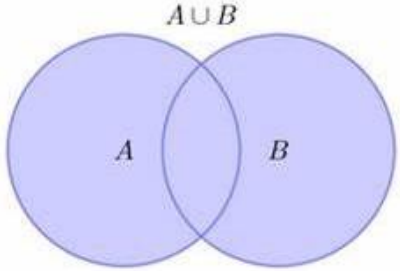
Pengertian Peluang (*Probability*)

(1) Pendekatan teori peluang: klasikal

- Peluang sebuah kejadian ditetapkan berdasarkan asumsi bahwa setiap keluaran memiliki peluang yang sama.
- Jika sebuah eksperimen dapat menghasilkan satu dari n-buah keluaran maka peluang setiap keluaran adalah $\frac{1}{n}$.
- Setiap keluaran diasumsikan mempunyai peluang yang sama.
- Jika A dan B adalah dua buah kejadian maka:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$





Pengertian Peluang (*Probability*)

(2) Pendekatan teori peluang: frekwensi-relatif

- Peluang sebuah kejadian ditetapkan berdasarkan frekwensi keluaran yang telah diperoleh sebelumnya.
- Jika A adalah sebuah kejadian yang menghasilkan keluaran tertentu (misalnya: x) dan diasumsikan bahwa telah dilakukan eksperimen sebanyak n-buah ulangan dimana kejadian A telah terjadi $n(A)$ kali maka peluang kejadian A adalah:

$$\mathrm{P(A)} = \lim_{n\to\infty} \frac{n(A)}{n}$$

- Jika A dan B adalah dua buah kejadian maka:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Pengertian Peluang (*Probability*)

(3) Pendekatan teori peluang: aksiomatik

- Peluang sebuah kejadian dipandang sebagai sebuah fungsi P yang memetakan setiap kemungkinan kejadian terhadap bilangan positif dalam selang $[0, 1]$
- Jika A adalah sebuah kejadian maka:

$$0 \leq P(A) \leq 1.$$

- Jika S adalah himpunan semesta dari seluruh kemungkinan kejadian A_i maka:

$$P(S) = P(A_1 \cup A_2 \cup \cdots) = 1.$$

- Jika A dan B adalah dua buah kejadian maka:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

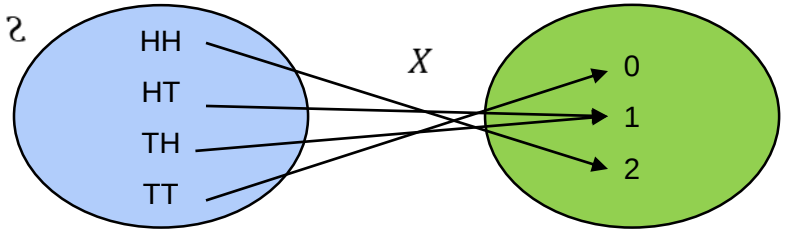
Peluang Kejadian Bebas dan Bersyarat

- Jika A dan B adalah dua buah kejadian yang bersifat bebas maka:
  - $P(A \cap B) = P(A)P(B)$
- Jika A dan B adalah dua buah kejadian bersyarat maka:
  - $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
  - $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B \cap A)}{P(A)}$
- Teorema Bayes:
  - $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B \mid A)}{P(B)}$
- Join Probability:
  - $P(A \cap B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$

Variabel Acak (*Random Variable, Variable*)

- Jika S adalah himpunan semua keluaran maka sebuah *random variable* X adalah sebuah fungsi yang memetakan himpunan S kedalam himpunan bilangan bulat.
  - Misalnya sebuah eksperimen melempar dua buah mata uang logam dimana setiap uang logam memiliki sisi: H dan T.
  - S adalah himpunan keluaran berupa sisi mata uang logam pertama dan kedua maka $S = \{HH, HT, TH, TT\}$.
  - Jika *random variable* X didefinisikan sebagai jumlah sisi H dari setiap keluaran didalam himpunan S maka:

$$X = \{0, 1, 2\}$$



Kategori variabel acak (variabel)

- Variabel diskrit (*discrete random variable*):
  - Nilai variabel bersifat diskrit
- Variabel kontinyu (*continuous random variable*):
  - Nilai variabel merupakan bilangan riil

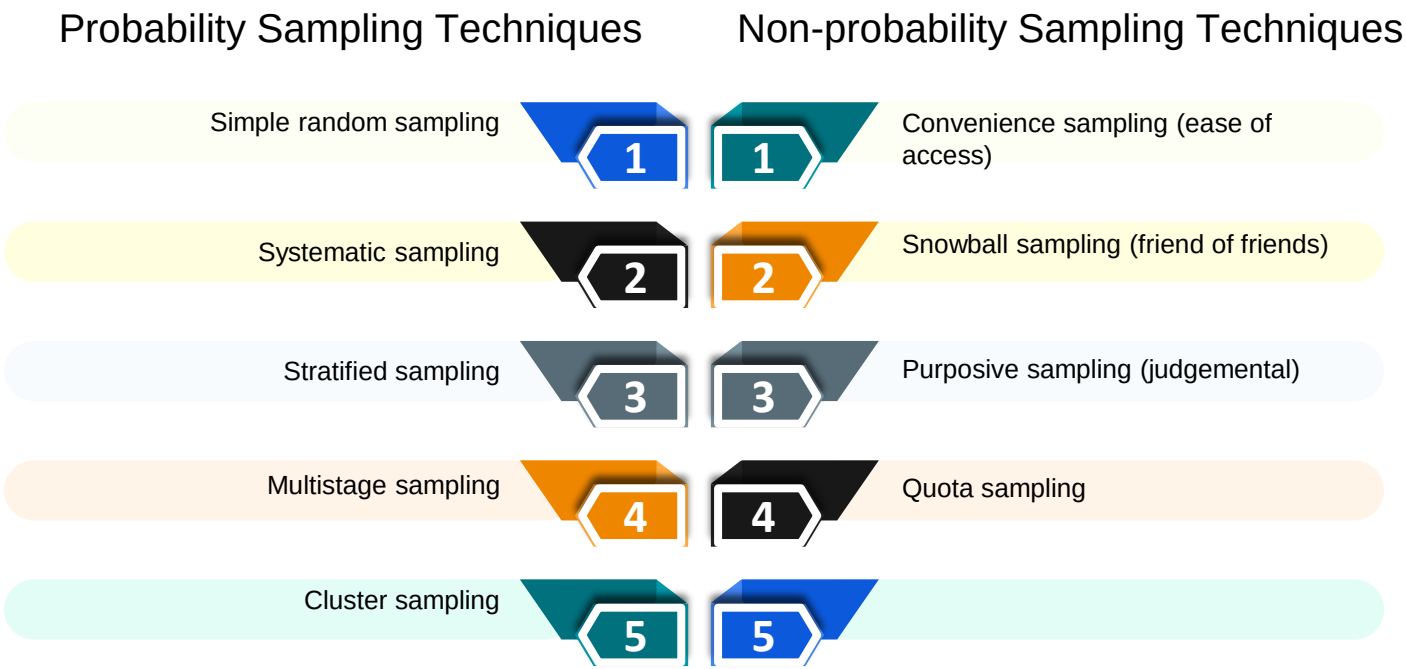
Populasi dan Sampel

- Populasi Target (*Target Population*):
  - Himpunan dari keseluruhan objek studi yang menjadi target pembuatan generalisasi.
  - Contoh: seluruh penduduk yang terinfeksi virus C19, seluruh penderita diabetes, dll.
- Populasi yang dapat diakses (*Accessible Population*):
  - Himpunan bagian dari populasi target yang secara aktual dapat diakses.
  - Contoh: seluruh penduduk yang terinfeksi virus C19 di DKI Jakarta, seluruh penderita diabetes di Kota Bogor Barat.
- Sampel:
  - Pada umumnya menganalisis seluruh anggota populasi tidak dapat/praktis dilakukan.
  - Anggota populasi yang terpilih menggunakan suatu teknik pengambilan sampel.
  - Contoh: sejumlah citra dari dataset Imaget

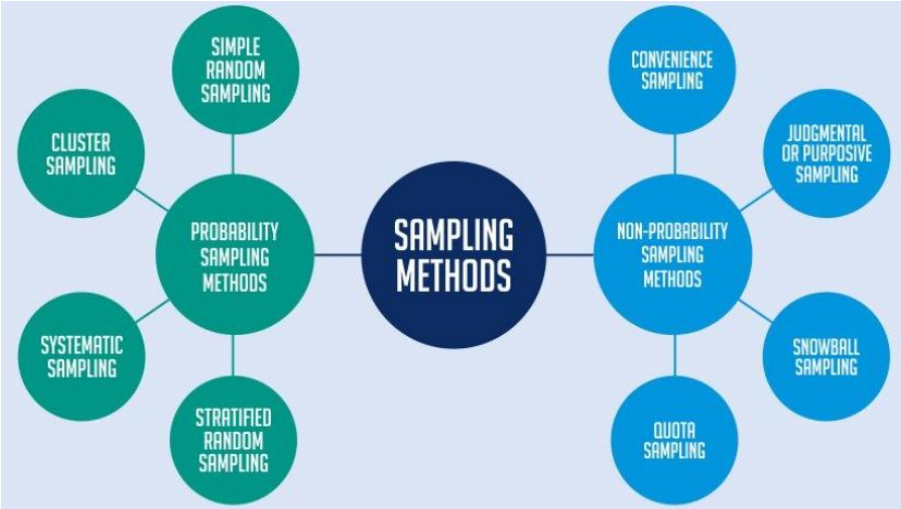
Teknik pemilihan sampel

- Teknik untuk memilih sampel dari populasi dengan tujuan agar sampel tersebut dapat “mewakili” seluruh anggota populasi.
- Teknik random sampling (*probability sampling*):
  - Teknik pengambilan sampel dimana semua individu dalam populasi baik secara sendiri-sendiri atau bersama-sama memiliki kesempatan yang sama untuk dipilih menjadi sampel
- Teknik *non-random sampling* (*non-probability sampling*):
  - Teknik pengambilan sampel dimana tidak semua anggota populasi memiliki kesempatan yang sama untuk dipilih menjadi sampel.
  - Penggunaan teknik non-probability sampling ini sering digunakan dengan mempertimbangkan sejumlah faktor: ketersediaan data, risiko pengambilan sampel, dll

Teknik pemilihan sampel



Teknik pemilihan sampel



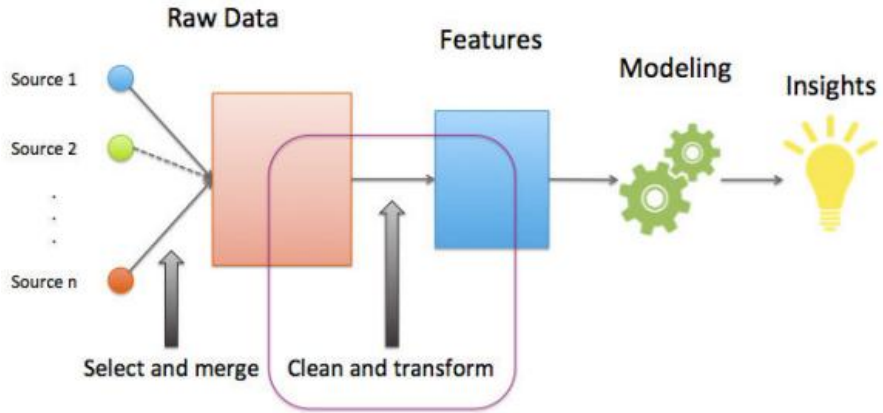
- ***Probability Sampling***:
  - Populasi diketahui
  - Randomisasi/keteracakan: Ya
  - Conclusiver
  - Hasil: Unbiased
  - Kesimpulan: Statistik
- ***Non-Probability Sampling***:
  - Populasi tidak diketahui
  - Keterbatasan penelitian
  - Randomisasi/keteracakan: Tidak
  - Exploratory
  - Hasil: Biased
  - Kesimpulan: Analitik

Data dan Informasi

- Data adalah
  - Fakta (factual information) hasil pengukuran atau pengamatan yang digunakan sebagai dasar penalaran atau perhitungan (Merriam-Webster, 2024).
  - Secara umum data adalah: sejumlah fakta mengenai objek hasil pengukuran atau hasil pengamatan berupa uraian/deskripsi atau bilangan.
  - Data mentah adalah sejumlah data yang belum “dibersihkan”.
  - Contoh: data seorang pegawai
- Variabel (atribut, fitur) data:
  - Atribut dari objek.
  - Contoh fitur dari data pegawai: Nama, Jenis Kelamin, Tempat Lahir, Tanggal Lahir, Status Perkawinan,
- Informasi
  - Informasi adalah: data yang telah diklasifikasikan atau disusun sehingga memberikan nilai bagi penggunanya.
  - Contoh: daftar pegawai yang berjenis kelamin: Laki-laki dan berusia diatas 40 tahun.

Fitur dari Data

- Fitur (variabel, atribut) adalah representasi dari data.
  - Setiap jenis data memiliki fitur sendiri.
  - Fitur objek: attribute/variabel.
  - Contoh fitur dari data pegawai: Nama, Jenis Kelamin, Tempat Lahir, Tanggal Lahir, Status Perkawinan,
- Rekayasa Fitur (feature engineering) adalah proses memformulasikan fitur yang paling tepat untuk merepresentasikan data.



Pengertian jenis data sebuah fitur

- Skala pengukuran sebuah fitur dikelompokkan kedalam nominal, ordinal, interval, dan rasio (Stanley Smith Stevens, 1940).
- Tujuan pengelompokan skala pengukuran fitur adalah untuk:
  - Menjelaskan karakteristik sebuah fitur.
  - Menetapkan analisis statistic yang tempat untuk fitur tersebut.

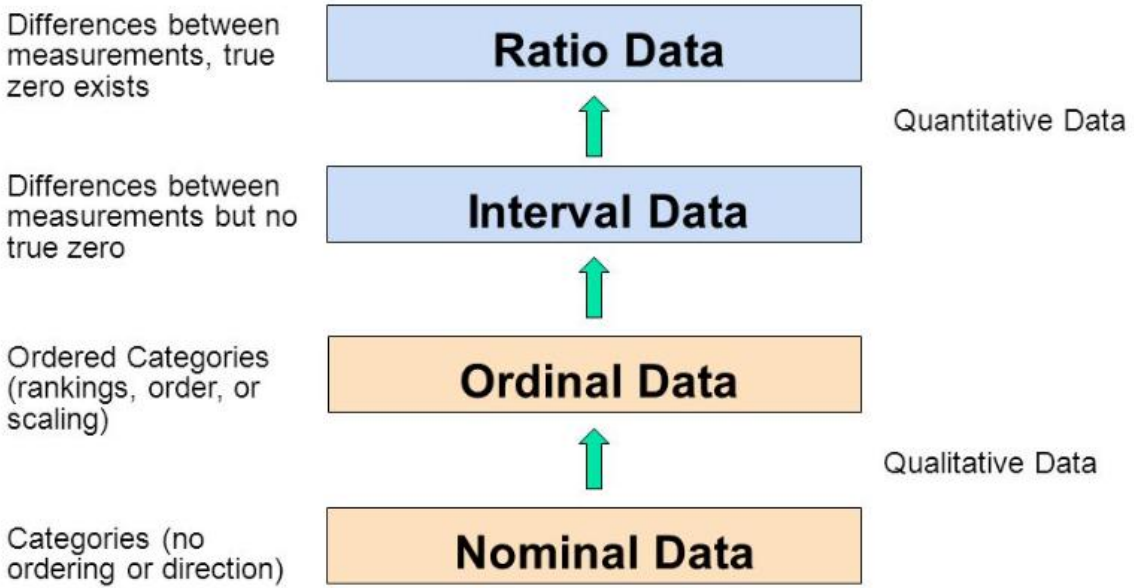
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Pentingnya jenis data sebuah fitur

Tipe data sebuah fitur menentukan statistik deskriptif yang tepat untuk penelaahan fitur data, contoh:

- Jenis Kelamin atau Agama memiliki tipe data nominal sehingga tidak dapat dihitung Mean (rata-rata) fitur tersebut.
- Status Sosial Ekonomi (*Social Economy Status*) memiliki tipe data ordinal sehingga tidak dapat dihitung Mean (rata-rata) fitur tersebut.
- Warna benda dapat dipandang sebagai tipe data nominal tetapi dari perspektif ilmu Fisika warna berkaitan dengan Panjang gelombang sinar sehingga merupakan tipe data rasio.
- Suhu sebuah benda dalam skala ℃ atau ℉ merupakan data interval, tetapi dalam skala Kelvin merupakan tipe data rasio.
- Tingkat penghasilan seseorang merupakan data ordinal tetapi Besar Penghasilan memiliki tipe data rasio.

Jenis data (tipe pengukuran)



Tinggi objek, Panjang benda, Berat objek

Suhu, Tahun, Skor test IQ, Waktu,

Tingkat pendidikan (D3, S1, S2, S3)
Tingkat kepuasan terhadap layanan (Tidak Puas, Puas, Sangat Puas)

Jenis Kelamin (Laki-laki, Perempuan),
Golongan darah (A,B,O),
Prodi (Manajemen, SI, TI, Teknik)

Jenis data (tipe pengukuran)

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Statistika Deskriptif

- Cabang dari ilmu statistika yang bertujuan untuk memberikan ringkasan mengenai fitur dari sampel data.

Jenis data	Ukuran Pemusatan	Ukuran Penyebaran	Ukuran Keeratan dua Variabel	Sebaran Data
Nominal	Modus	--	--	Frekwensi nilai, Proporsi
Ordinal	Median, Modus	--	--	Frekwensi nilai, Proporsi
Interval	Mean, Median, Modus	Varians, Kuartil, Persentil, Kisaran	Koefisien Korelasi	Frekwensi interval nilai
Ratio	Mean, Median, Modus	Varians, Kuartil, Persentil, Kisaran	Koefisien Korelasi	Frekwensi interval nilai

Statistika Deskriptif

- **Ukuran Pemusatan:** ukuran yang menjelaskan titik pusat fitur data
  - Mean $(\dot{x}) = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$
  - Kuartil ke-1 (Q_1) adalah nilai fitur data dimana 25 % dari jumlah sampel nilainya < Q_1
  - Kuartil ke-2 (Q_2) atau Median adalah nilai data dimana 25 % dari jumlah sampel nilainya < Q_2.
  - Kuartil ke-3 (Q_3) adalah nilai data dimana 75 % dari jumlah sampel nilainya < Q_3.
  - Modus adalah nilai yang paling sering muncul (frekwensi paling tinggi) dari sampel data
- **Ukuran Variabilitas: ukuran variabilitas data**
  - Varians atau variance $(s^2) = \frac{1}{n-1}\left[(x_1 - \dot{x})^2 + (x_2 - \dot{x})^2 + \cdots + (x_n - \dot{x})^2\right]$ dimana n: jumlah sampel data.
  - Standar deviasi $(s) = \sqrt{s^2}$
  - Kisaran atau range = $x_{max} - x_{min}$

Statistika Deskriptif

- **Ukuran Keeratan dua Variabel:**
  - Koefisien Korelasi Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

  - Dimana: x_i adalah nilai variabel x ke-i, y_i adalah nilai variabel y ke-i, n adalah jumlah sampel data.
  - Kisaran nilai r: $-1 \leq r \leq 1$

r	Interpretasi
0	Tidak berkorelasi
0,01-0,20	Korelasi Sangat rendah
0,21-0,40	Rendah
0,41-0,60	Agak rendah
0,61-0,80	Cukup
0,81-0,99	Tinggi
1	Sangat tinggi

Statistika Inferens

- Cabang dari ilmu statistika yang bertujuan untuk memprediksi (generalisasi) karakteristik sebuah populasi berdasarkan sampel data.
- Metode statistika inferens:
  - Hypothesis testing
  - Regression analysis.

Ringkasan

- Pengolahan data menggunakan teknik statistika dapat dilakukan dengan berbagai kakas, termasuk penggunaan Rapidminer untuk kalangan non-programmer yang bersifat *open source*.
- Rapidminer merupakan kakas non pemrograman visual yang memanfaatkan metode *drag and drop,* sehingga lebih *user-friendly* dan mudah digunakan.

Referensi

- Dennis Aprilla C, Donny Aji Baskoro, Lia Ambarwati, I Wayan Simri Wicaksana. *Belajar Data Mining dengan RapidMiner*. 2013. Universitas Dian Nuswantoro. https://repository.dinus.ac.id/docs/ajar/Belajar_Data_Mining_dengan_RapidMiner.pdf
- Rapidminer Studio Manual, 2014. https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih