KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 13:
EDA pada Real World Data

Outline

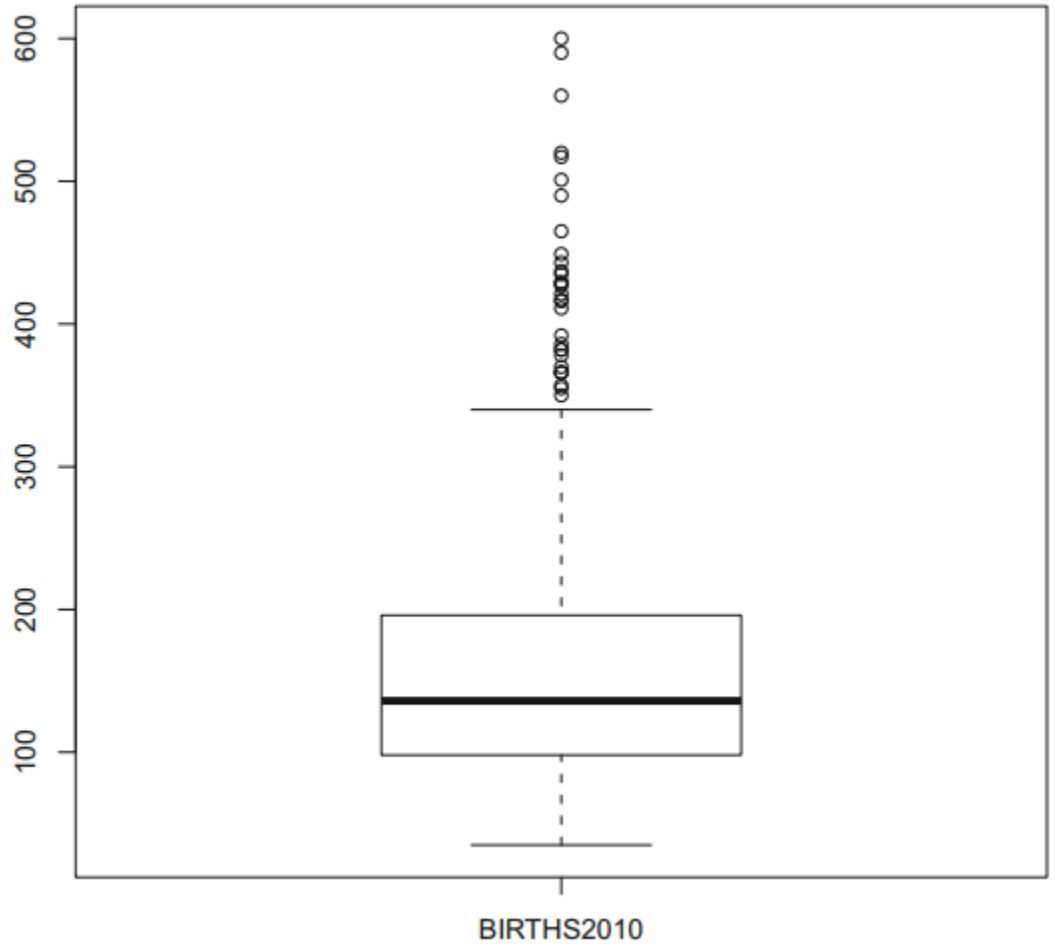
- Jenis distribusi data
- Identifikasi distribusi dan pola data
- Interpretasi distribusi dan pola data

Distribusi Data

- Sebuah langkah yang penting dalam EDA adalah memahami distribusi data
- Ringkasnya, distribusi data adalah sebuah pemetaan terhadap data dimana tiap poin pada data tersebut berada dalam area atau region tertentu
- Untuk memvisualisasikan distribusi data, boxplot, histogram, dan density plot sering digunakan
- Selain itu, terdapat angka skewness dan kurtosis untuk mengukur kecondongan data secara numerik

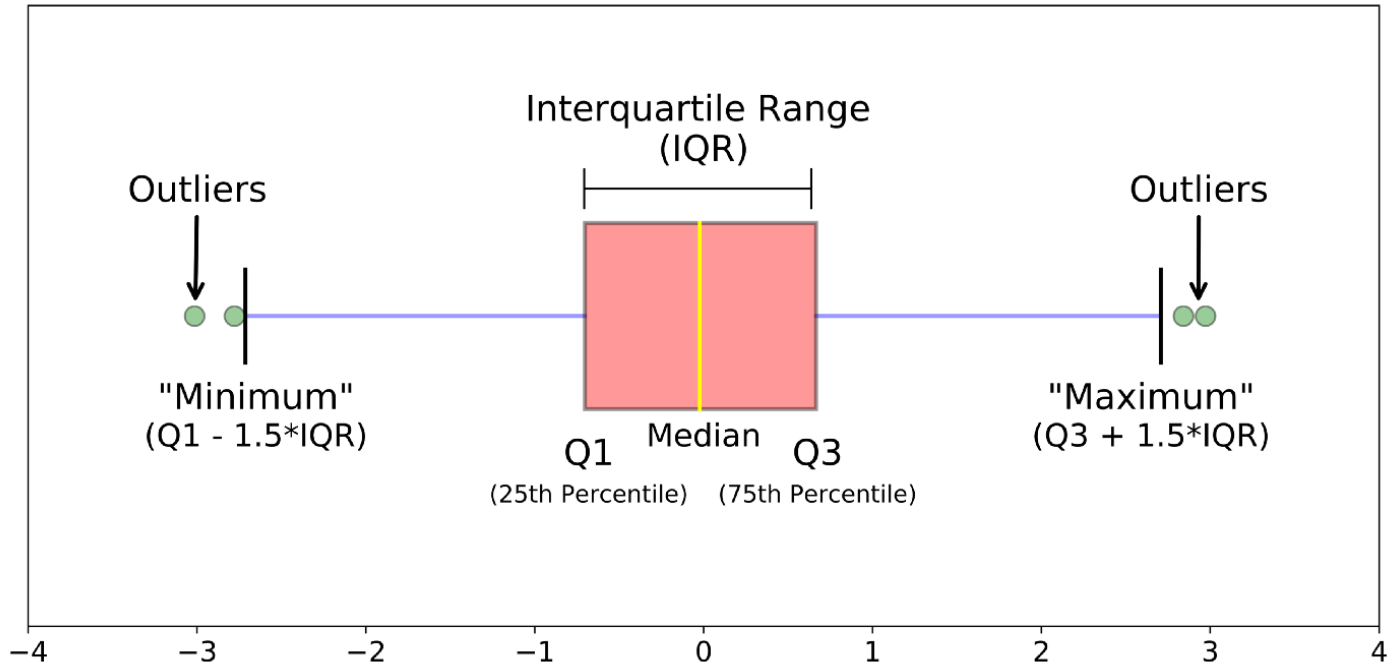
Box and Whisker Plot / Box Plot

- Sebuah plot yang dapat merepresentasikan nilai minimum, Q1, median, Q3, dan maximum suatu atribut dalam data numerik.
- Sebuah boxplot dapat memberikan gambaran distribusi data
- Terdapat 3 elemen:
  - Box: A rectangular box with a solid horizontal line in bold.
  - Whiskers: A pair of vertical dotted lines, with solid horizontal end lines which are called notches.
  - Extremities: Unfilled circles or bubbles located above or below the whiskers.

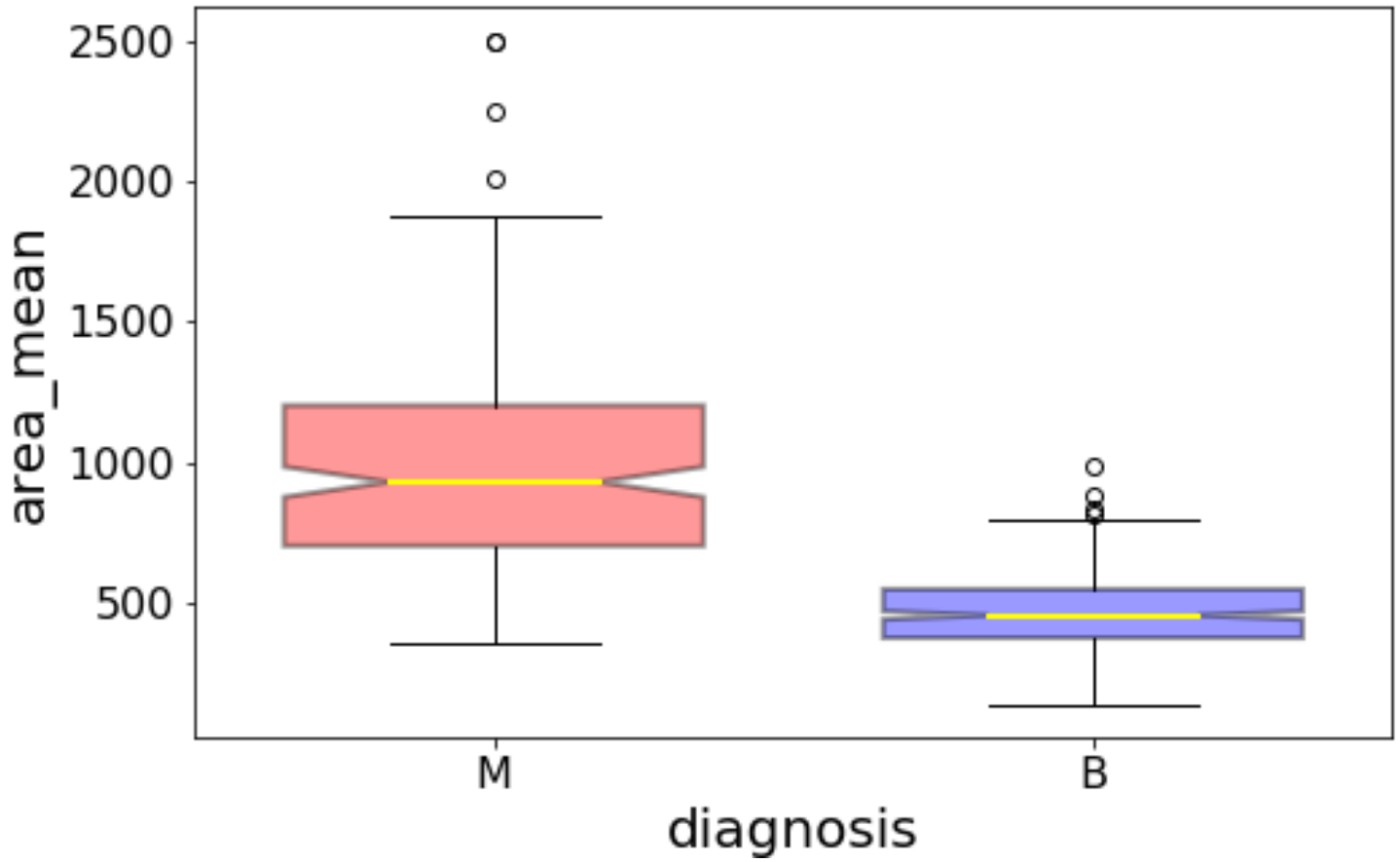


Komponen Box Plot

- Garis tebal didalam box dapat diinterpretasikan sebagai Median
- 2 garis pada luar box dapat diinterpretasikan sebagai Quartil 1 dan Quartil 3
- Sedangkan 2 garis diujung whisker merupakan nilai minimum dan maksimum
- Besarnya ukuran box, merepresentasikan centrality atau kedekatan sebuah data terhadap titik tengahnya
- Sedangkan titik bulat di luar whisker merepresentasikan data yang tidak terjangkau oleh 1.5 kali dari Inter Quartile Range, atau disebut outlier.

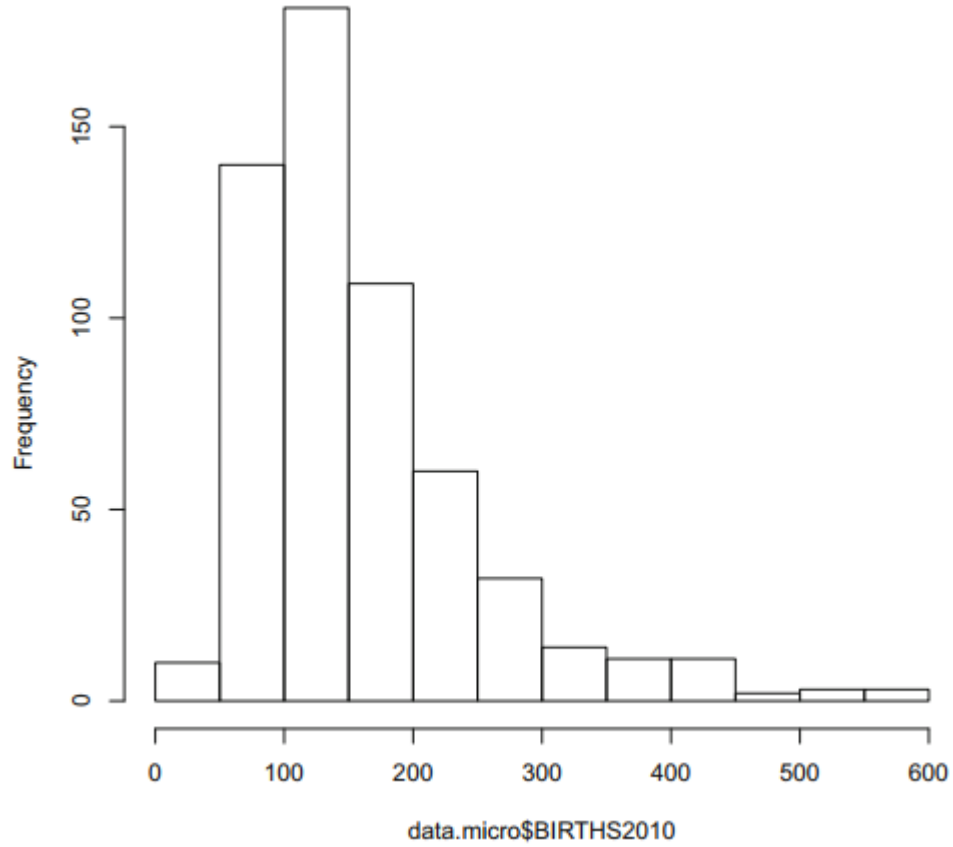


Boxplot untuk 2 atau lebih Atribut



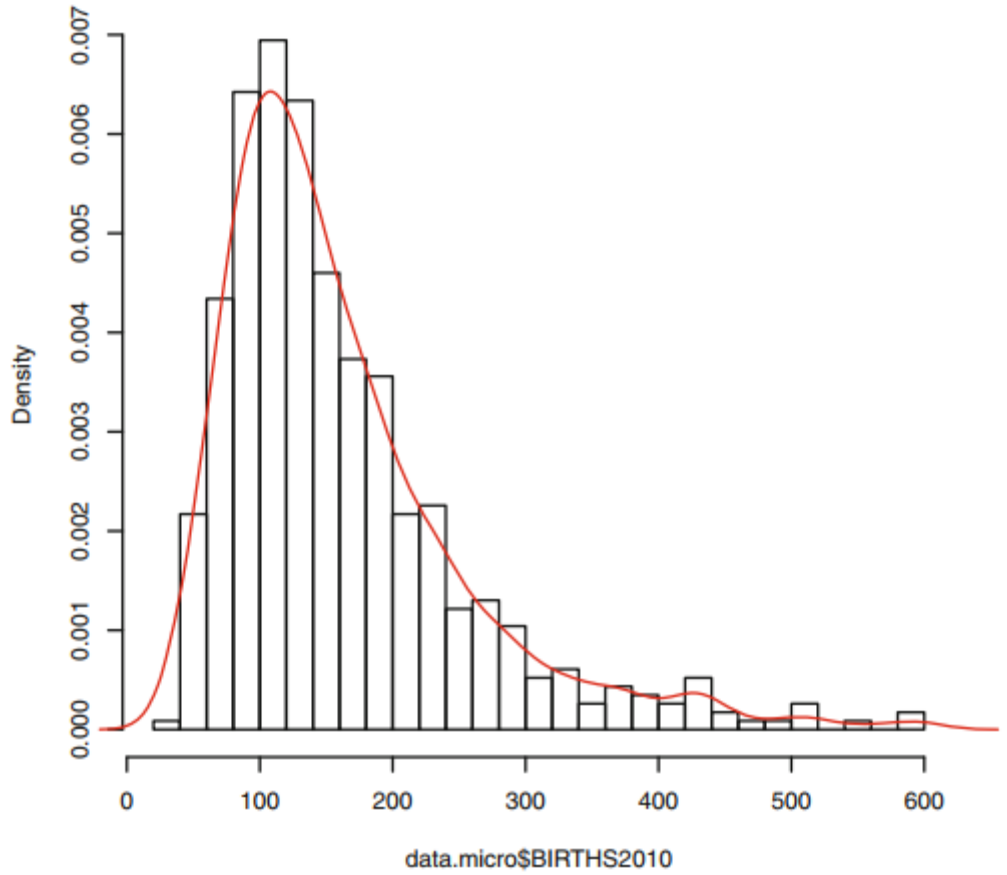
Histogram

- Selain Boxplot, Histogram atau Bar Chart juga dapat merepresentasikan distribusi data
- Mudah untuk melihat kecondongan data dan apakah data tersebut memenuhi distribusi normal (more on this later..)



Histogram + Density Plot

- Untuk melihat distribusi kontinu
- Dapat mendeteksi skewness secara visual (serta centrality tendency)



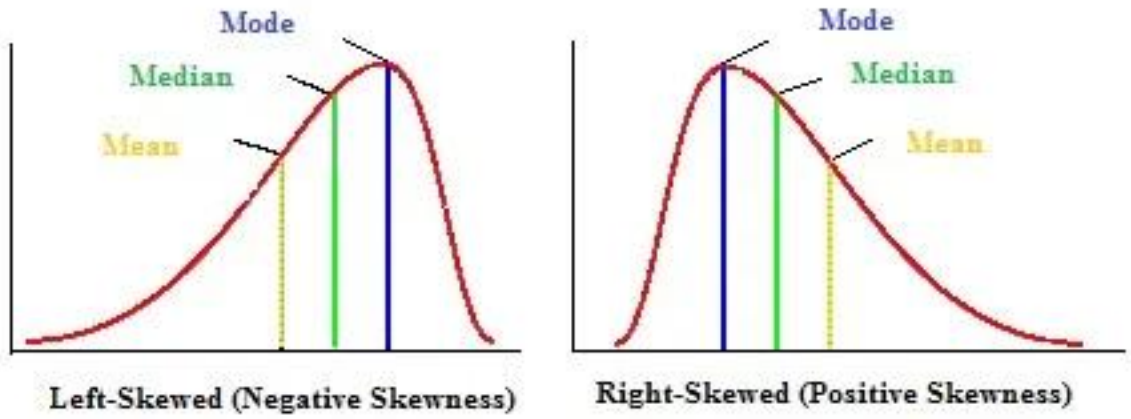
Mengukur Kesimetrisan Data dengan Skewness dan Kurtosis

- Histogram dan Boxplot menunjukkan secara visual karakteristik sebuah atribut
- Serta menunjukkan bagaimana data tersebar
- Skewness adalah sebuah pengukuran yang menunjukkan ke-tidak-simetrisan suatu variabel numerik
- Sebuah variabel dengan skewness = 0 berarti data tersebut terdistribusi secara simetris

$$skewness = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_i (x_i - \bar{x})^2\right)^{\frac{3}{2}}},$$

Skewness

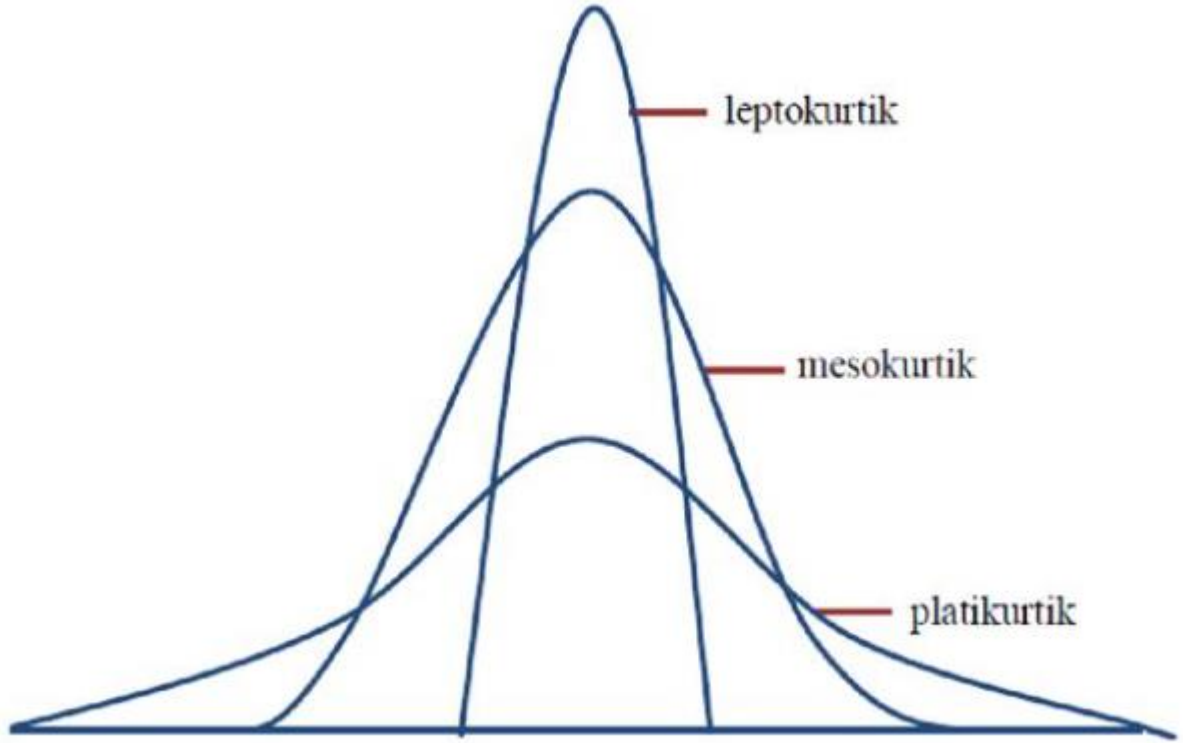
- Menggunakan skewness() pada R, kita dapat melihat nilai skewness dari sebuah variabel dalam data
- Jika nilai skewness positif, berarti variabel tersebut disebut right-skewed. Sedangkan jika nilainya negatif, berarti left-skewed



Kurtosis

- Disebut juga keruncingan suatu variabel atau kepuncakan dari sebuah distribusi variabel
- Dalam distribusi normal (Gaussian) ni kurtosis adalah 3.
- Jika nilai kurtosis kurang dari 3 disebu Platikurtik
- >3 disebut Leptokurtik
- =3 disebut Mesokurtik

$$\text{kurtosis} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_i (x_i - \bar{x})^2\right)^2}.$$

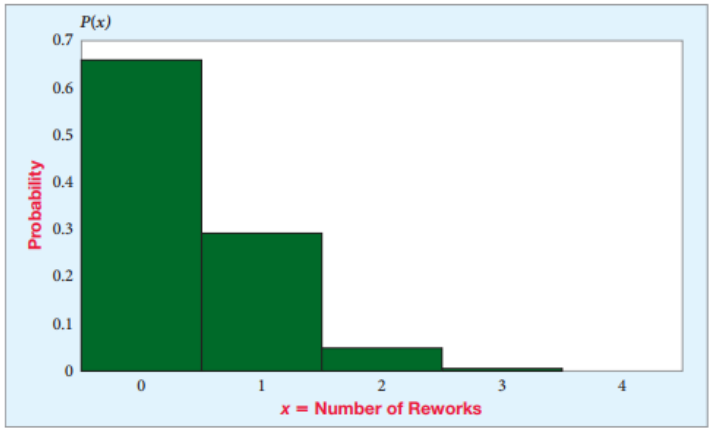


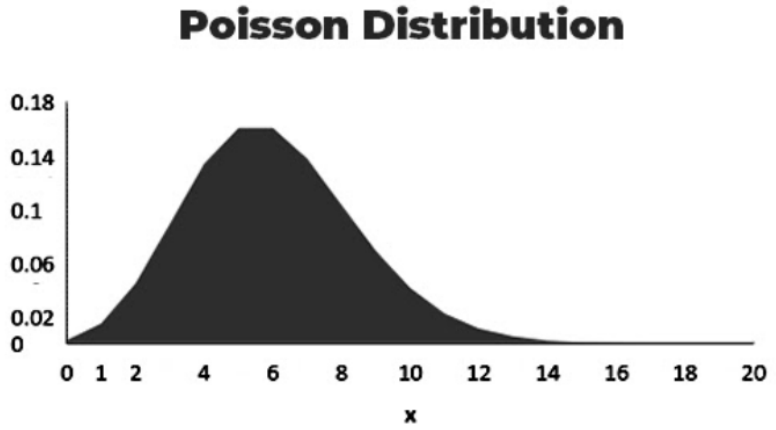
Jenis distribusi data

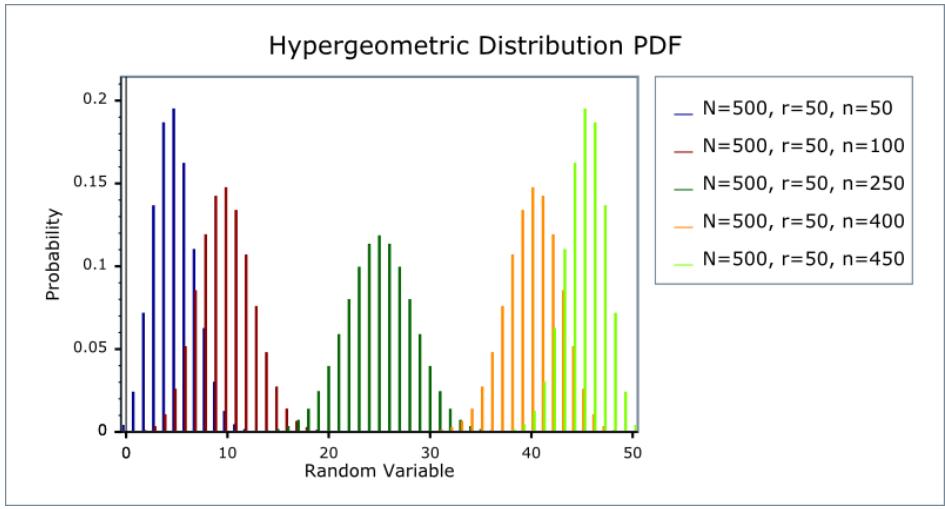
- Data yang dikumpulkan berdasarkan hasil observasi maupun bacaan otomatis sensor memiliki distribusi
- Distribusi ini melambangkan sebaran dan karakteristik dari suatu atribut dalam data tersebut
- Distribusi data dibagi dalam 2 bagian:
  - Distribusi Diskrit, untuk data yang bersifat kategorik
  - Distribusi Kontinyu, untuk data yang bersifat kontinyu / real number

Distribusi Data Diskrit

- Ada banyak jenis distribusi data diskrit, namun distribusi ini yang sering ditemui dalam real world data:
  - Distribusi Binomial
  - Distribusi Hypergeometric
  - Distribusi Poisson
- Bentuk visualisasi untuk jenis-jenis distribusi tersebut adalah

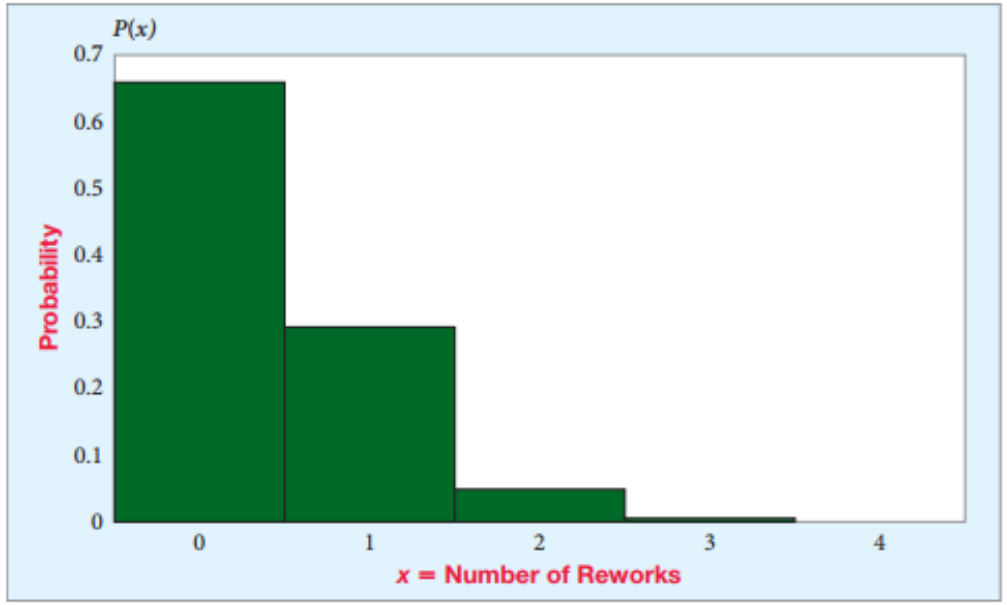






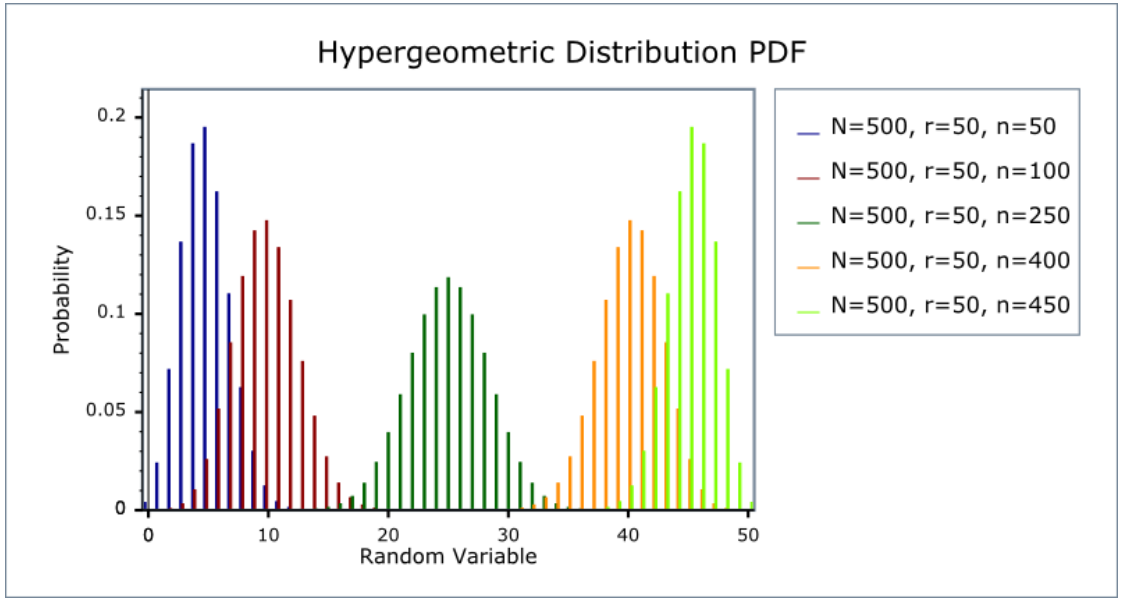
Distribusi Binomial

- Distribusi binomial memiliki kemungkinan 2 kesimpulan dalam 1 kali observasi.
- Karakteristik distribusi binomial adalah:
  - Jumlah percobaan harus ditentukan untuk setiap observasi.
  - Setiap percobaan hanya memiliki 2 kesimpulan, gagal atau sukses
  - Probabilitas keberhasilan untuk tiap percobaan harus sama
  - Setiap percobaan bersifat independen terhadap percobaan lainnya (tidak terpengaruh outcome)



Distribusi Hipergeometrik

- Dalam distribusi binomial diasumsikan bahwa peluang suatu kejadian tetap atau konstan atau antar-kejadian independen.
- Dalam dunia nyata, jarang terjadi hal demikian. Nilai setiap kejadian mungkin berbeda atau tidak konstan.
- Distribusi dengan probabilitas berbeda adalah Distribusi Hipergeometrik.



Distribusi Poisson

- Dalam distribusi Binomial, jumlah percobaan yang dilakukan cenderung kecil
- Poisson menemukan bahwa jika jumlah percobaan diatas 50, distribusi Binomial tidak efektif lagi
- Distribusi Poisson dipakai untuk menentukan peluang suatu kejadian yang jarang terjadi, tetapi mengenai populasi yang luas atau area yang luas dan juga berhubungan dengan waktu.

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$

λ adalah rata rata kejadian sukses di periode yang ditentukan

X adalah jumlah kejadian sukses di periode tersebut

Asumsi dan Karakteristik Distribusi Poisson

Asumsi:

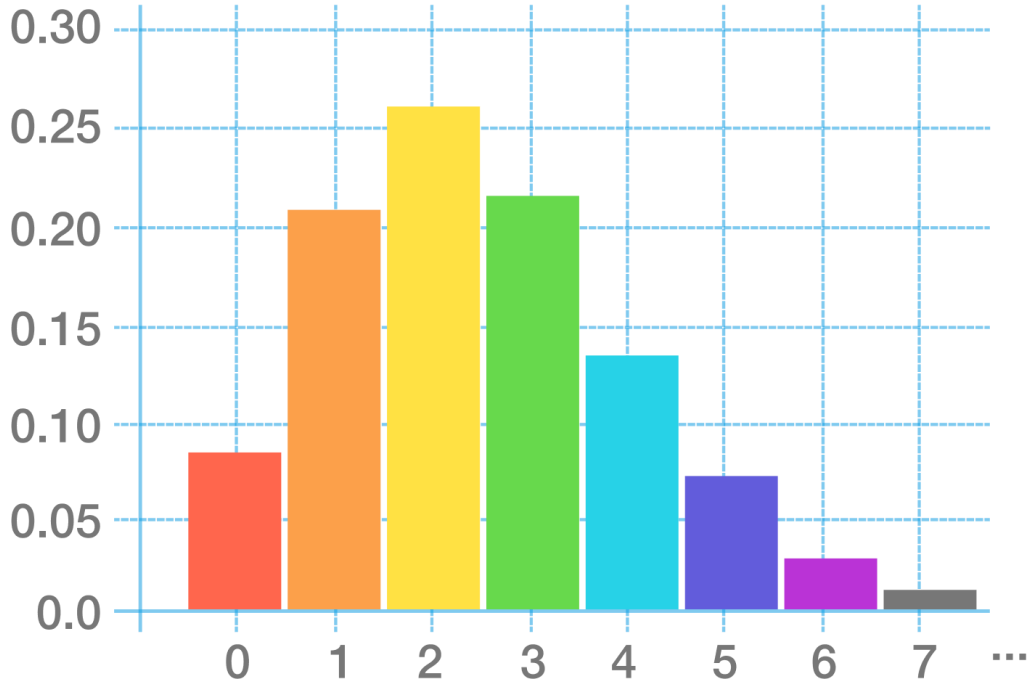
- Probabilitas kesuksesan dalam waktu singkat akan sama dengan probabilitas kesuksesan dalam waktu lama
- Probabilitas kesuksesan dalam satu waktu sama dengan nol sebanding dengan durasi yang mengecil
- Sebuah kejadian sukses tidak mempengaruhi kejadian sukses lainnya

Karakteristik:

- Kejadian bersifat independen (tidak mempengaruhi satu sama lain)
- Sebuah kejadian dapat terjadi beberapa kali dalam suatu periode yang ditentukan
- Dua kejadian tidak dapat muncul secara bersamaan
- Rata rata sebuah kejadian akan terjadi selalu konstan

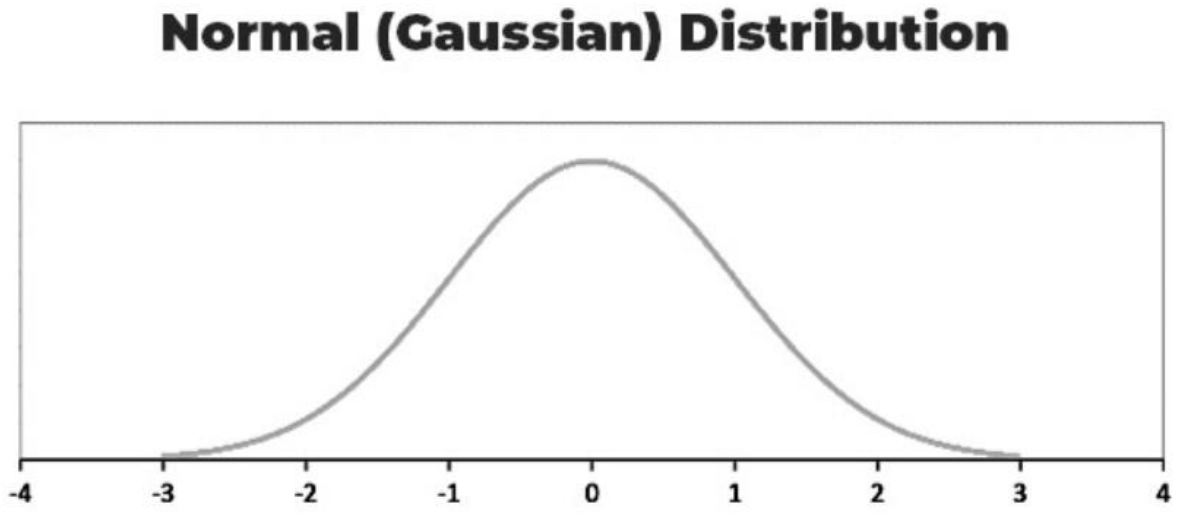
Contoh Penggunaan Distribusi Poisson

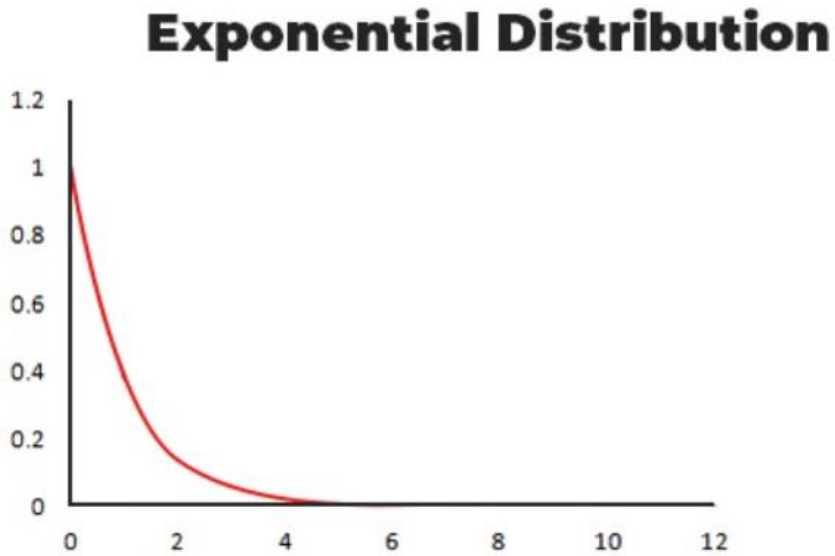
- Dalam World Cup, suatu klub dalam setiap pertandingan rata rata menghasilkan 2.5 gol, dengan distribusi Poisson, hal tersebut dapat divisualisasikan sebagai berikut



Distribusi Data Kontinyu

- Distribusi peluang kontinu adalah variabel acak / hasil observasi yang dapat memperoleh semua nilai pada skala kontinu / (bilangan real)
- Tipe distribusi kontinyu:
  - Distribusi Normal (Gaussian)
  - Distribusi Eksponensial
- Visualisasi untuk jenis-jenis distribusi tersebut adalah





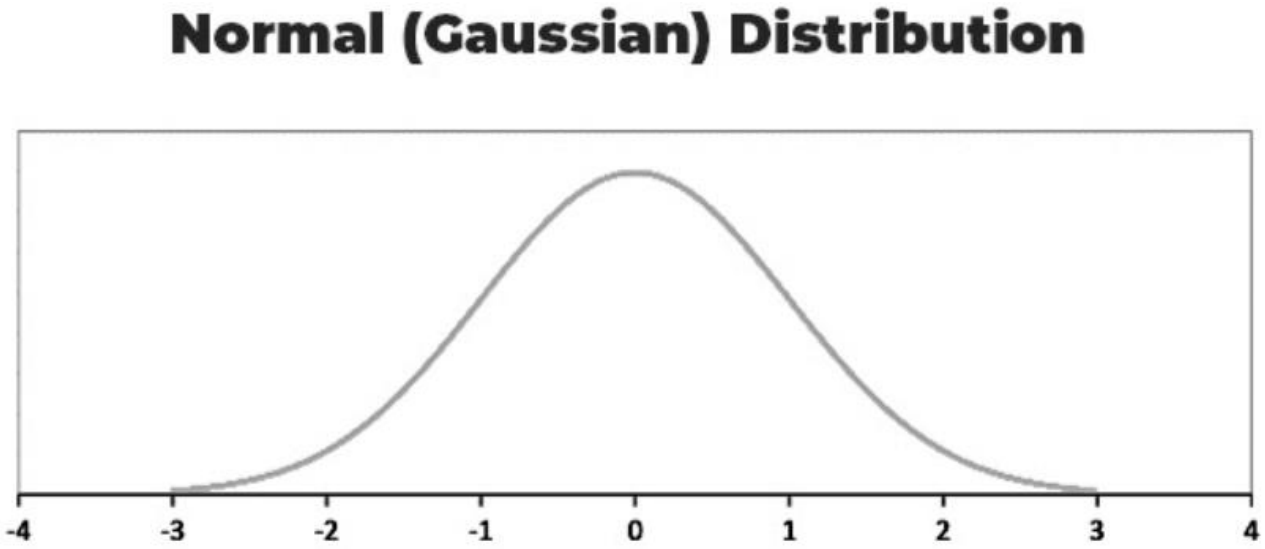
Distribusi Normal (Gaussian)

- Distribusi normal sangat sering sekali digunakan dan ditemukan dalam real world
- Proses umum dalam kehidupan seringkali dapat direpresentasikan dengan distribusi normal, sebagai contoh:
  - Distribusi jumlah pendapatan penduduk
  - Rata rata berat badan penduduk
  - Rata rata nilai mahasiswa
- Notasi matematis untuk distribusi ini adalah

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

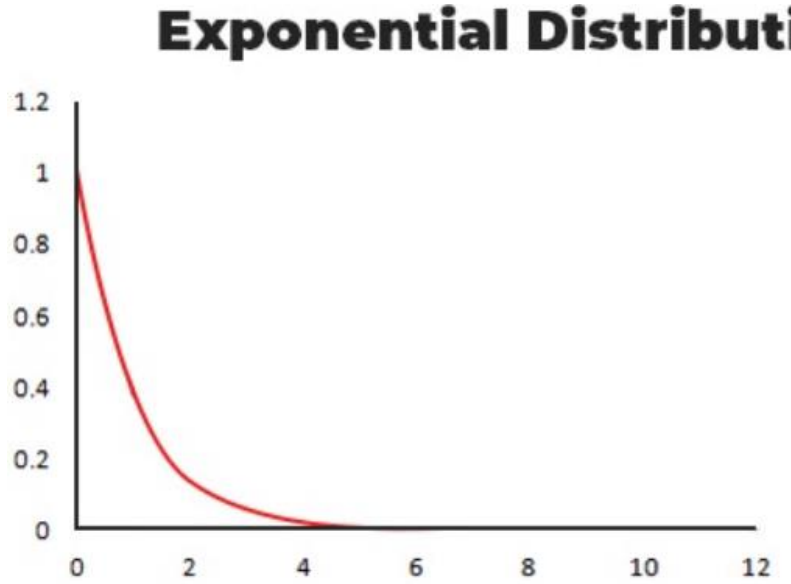
Karakteristik Distribusi Normal

- Dalam sebuah density plot, sebuah data dikatakan memiliki distribusi normal jika:
  - Kurvanya berbentuk garis lengkung yang halus dan berbentuk seperti Lonceng (Bell Shaped)
  - Simetris terhadap rata rata (mean)
  - Kedua ekor/ ujungnya semakin mendekati sumbunya tetapi tidak pernah memotong
  - Mean, modus, dan mediannya sama



Distribusi Eksponensial

- Distribusi eksponensial memiliki dimensi waktu, sama seperti Poisson
- Seringkali digunakan untuk Survival Analysis, contoh: Masa hidup AC, masa hidup laptop, dsb.
- Sebuah variabel dikatakan memiliki distribusi eksponensial jika



$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$

Tugas

- Membuat plot: histogram dan boxplot dari data berikut;
  a) data oecd_le (package socviz): atribut life expectancy at birth, di negara USA
  b) data oecd_le (package socviz): atribut life expectancy at birth, di negara Belgium
  c) data oecd_le (package socviz): atribut life expectancy at birth, di negara Canada
- Menyimpulkan distribusi dan/atau pola data dari plot tersebut
- Laporkan dalam bentuk laporan praktikum dengan menyertakan langkah langkah pengerjaan berupa narasi dan screenshot R serta hasil analisis dari setiap langkah.

- Tugas dikerjakan berkelompok.
- Tugas dikumpulkan paling lambat pukul 23.59 WIB di LMS.
- Beri nama file tugas: Tugas 13_Kelompok XX. (Contoh: Tugas 13_Kelompok 01)

FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN

PERTUKARAN MAHASISWA merdeka

Kampus Merdeka INDONESIA JAYA

Terima Kasih

SMART UNIVERSITY

TOP #465 WORLD UNIVERSITY RANKINGS UNAIR UNGGUL

ZONA INTEGRITAS menuju wilayah bebas korupsi