**Pengantar Penerapan Dasar-Dasar Descriptive Statistics**

Setelah memahami berbagai konsep dasar dalam analisis data, sekarang mari kita telaah salah satu komponen penting dalam analisis data. Apakah Anda bisa menebaknya?

*Yup*, betul sekali, pada beberapa materi ke depan, kita akan mendalami berbagai konsep dasar statistik yang sering digunakan untuk menganalisis data beserta penerapannya. Bagi Anda yang belum familier dengan statistik, jangan khawatir karena pada materi ini, kita akan mulai dengan konsep paling dasar terlebih dahulu.

*So*, apa yang ada dipikiran kalian ketika mendengar kata statistik? Mungkin banyak dari Anda membayangkan sekumpulan data dan rumus matematis yang disertai dengan istilah rumit dan sulit dipahami. Sebenarnya statistik tidaklah semenakutkan itu.

*By definition*, statistik merupakan ilmu yang menerapkan berbagai metode saintifik dalam proses pengumpulan, analisis, interpretasi, serta penyajian data. Berdasarkan definisi tersebut, Anda mungkin sudah bisa menebak bahwa statistik merupakan konsep yang sangat penting dalam proses analisis data. Oleh karena itu, untuk menjadi seorang praktisi data yang andal, Anda perlu berteman baik dengan statistik.

Terdapat beberapa konsep statistik yang umum digunakan untuk mendeskripsikan sebuah data yang dikenal sebagai *descriptive statistics*. Beberapa konsep tersebut bisa Anda lihat di bawah ini.

* *Measuring central tendency*untuk mengidentifikasi nilai khas atau sentral dalam sebuah data.
* *Measuring dispersion* untuk menilai variabilitas atau sebaran nilai dalam sebuah data.
* *Measuring asymmetric* untuk menilai ketidaksimetrisan dalam sebuah distribusi data.
* *Data relationship* untuk mengidentifikasi hubungan dua atau lebih *feature*/*column*/*variable* dalam sebuah data.

Nah, materi ini akan mengajak Anda untuk berkenalan dan berteman baik dengan berbagai konsep dalam statistik. Namun, sebelum membahas seluruh konsep tersebut, Anda perlu memahami berbagai tipe data terlebih dahulu.

**Tipe Data**

Pada proses analisis, tipe data akan menentukan metode analisis dan bentuk visualisasi data yang tepat. Oleh karena itu, pemahaman terhadap tipe data merupakan pengetahuan yang krusial bagi seorang praktisi data. Secara garis besar, data dapat dibagi menjadi dua tipe, yaitu kuantitatif (numerik) dan kategoris [4].

* **Data kuantitatif**  
  Data kuantitatif merupakan tipe yang direpresentasikan dalam skala numerik sehingga memungkinkan kita untuk menjalankan operasi matematis.



Bayangkan jika Anda merupakan seorang pencinta kucing yang memelihara banyak kucing. Data kuantitatif akan merepresentasikan jumlah kucing yang Anda miliki.  
  
Tipe data kuantitatif dapat dibagi ke dalam dua kategori seperti berikut.

* + **Data kontinu**  
    Data kontinu dapat direpresentasikan dalam berbagai nilai numerik seperti bilangan desimal, bulat, dll. Umumnya data kontinu memiliki nilai yang bisa dibagi atau diubah ke dalam unit yang lebih kecil. Salah satu contoh data kontinu adalah umur kucing dalam satuan tahun. Data umur kucing ini masih bisa diubah ke dalam unit yang lebih kecil, seperti bulan, hari, dst.
  + **Data diskret**  
    Data diskret merupakan data numerik yang hanya bisa direpresentasikan dalam bilangan bulat dan tidak dapat dibagi ke dalam unit yang lebih kecil. Jumlah kucing merupakan salah satu contoh data diskret.
* **Data kategoris**  
  Data kategoris merupakan tipe data yang menggunakan kelas atau label untuk merepresentasikan kelompok dari suatu informasi/data.



Bayangkan jika Anda memiliki banyak kucing. Tentunya setiap kucing tersebut memiliki ciri khas masing-masing, seperti warna bulu, warna mata, atau bahkan jenis spesies dari kucing tersebut. Informasi terkait ciri khas dari suatu kucing ini merupakan contoh dari tipe data kategoris.

Tipe data kategoris dapat dibagi lagi ke dalam dua tipe seperti berikut.

* **Data ordinal**  
  Data ordinal merupakan tipe data kategoris yang dapat diurutkan berdasarkan peringkat. Sebagai contoh, Anda memberikan rating 1 sampai 5 terhadap seluruh kucing yang Anda miliki untuk membedakan keaktifan dari kucing tersebut. Kucing yang memiliki rating 5 merupakan kucing yang sangat aktif, sedangkan kucing yang memiliki rating 1 merupakan kucing yang sangat pasif. Data seperti ini merupakan contoh data ordinal.
* **Data nominal**  
  Data nominal ialah tipe data kategoris yang tidak memiliki urutan atau peringkat. Warna bulu atau jenis spesies kucing merupakan contoh dari data nominal. Nilai dari kedua data tersebut tidak dapat kita urutkan. Hal inilah yang menjadi perbedaan mendasar antara data nominal dan ordinal.

**Data kuantitatif**

Data kontinu : Direpresentasikan dalam berbagai nilai numerik dan bisa dibagi atau diubah ke dalam unit yang lebih kecil (Umur, tinggi badan, berat badan, dll.)

Data diskret : Direpresentasikan dalam bilangan bulat dan tidak dapat dibagi ke dalam unit yang lebih kecil (Jumlah kucing peliharaan, jumlah produk yang terjual, dll.)

**Data kategoris**

Data ordinal : Tipe data kategoris yang dapat diurutkan berdasarkan rangking (Rating, kategori pelanggan berdasarkan *income*, dll.)

Data nominal : Tipe data kategoris yang tidak memiliki urutan atau rangking (Jenis kelamin, warna bulu kucing, dll.)

**Measuring Central Tendency**

Pada materi sebelumnya kita telah melihat beberapa tipe data. Tipe data tersebut akan menentukan metode statistik yang digunakan untuk menganalisis data tersebut. Sebagai contoh, pada data kuantitatif, kita dapat menggunakan beberapa parameter statistik untuk mendeskripsikannya, salah satunya adalah *measuring central tendency*.

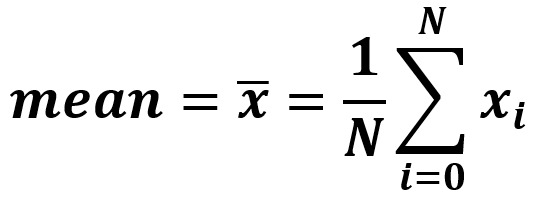
Nah, pada materi kali ini, kita akan mengupas secara tuntas berbagai konsep dalam measuring central tendency. Selain itu, kita juga akan melihat cara menghitungnya.

Seperti biasa ketika mempelajari suatu hal baru, kita harus mulai dari konsep dasar terlebih dahulu. Mari kita mulai dari definisi measuring central tendency.

Measuring central tendency merupakan sekumpulan parameter statistik untuk menggambarkan nilai khas atau sentral yang mewakili keseluruhan observasi atau data. Nilai khas atau sentral tersebut direpresentasikan menggunakan tiga parameter statistik yaitu *mean*, *median*, dan *mode*[5]. Mari kita melihat ketiga parameter statistik tersebut secara detail.

**Mean**

Mean atau sering juga disebut sebagai *expected value* merupakan salah satu parameter yang paling sering digunakan untuk menggambarkan nilai khas atau sentral dari suatu data. Ia diperoleh dari menghitung jumlah keseluruhan data dan dibagi dengan banyaknya data yang dimiliki. nilai mean yang diperoleh dari contoh di atas berbentuk bilangan desimal. Hal ini tentunya kurang masuk akal untuk merepresentasikan nilai khas dari data ini. Selain itu, parameter ini sangat sensitif terhadap *noise*dan *outlier*yang terdapat dalam sebuah data.



**Median**

Median merupakan parameter yang merepresentasikan nilai tengah atau persentil ke-50 dari keseluruhan observasi atau data. Parameter ini diperoleh dengan mengambil nilai persentil ke-50 dari data yang telah diurutkan.

Untuk data yang berjumlah ganjil, nilai median diperoleh dengan mengambil langsung nilai tengah dari data yang telah diurutkan. Namun, hal ini berbeda jika data yang dimiliki berjumlah genap. Pada kasus seperti ini, kita perlu mengurutkan dan menghitung nilai rata-rata dari dua data yang berada di tengah. nilai mean yang diperoleh dari contoh di atas berbentuk bilangan desimal. Hal ini tentunya kurang masuk akal untuk merepresentasikan nilai khas dari data ini.

Median dikenal sebagai salah satu parameter yang sangat toleran terhadap outlier. Apabila Anda mengubah salah satu nilai dari data tersebut, parameter median yang diperoleh tidak akan berubah secara signifikan.

**Mode**

Parameter ketiga dalam measuring central tendency ialah mode. Parameter ini akan memberikan kita gambaran atau informasi terkait nilai yang paling sering muncul dalam suatu data.

Untuk menghitung parameter mode, kita perlu mengurutkan nilai dari suatu data terlebih dahulu. Selanjutnya memilih nilai yang paling sering muncul. Sebagai contoh, pada data jumlah kucing peliharaan yang dimiliki pegawai Dicoding, nilai yang paling sering muncul adalah 2. Nilai ini merupakan mode dari data tersebut.

**Measuring Dispersion**

Sebelumnya kita telah membahas tentang *measuring central tendency* untuk mendeskripsikan nilai khas atau sentral dari keseluruhan observasi atau data. Namun, tahukah Anda bahwa hal itu masih belum cukup untuk menggambarkan keseluruhan cerita yang dimiliki sebuah data? Oleh karena itu, selain melihat nilai sentral, kita juga perlu mengetahui sebaran dari data tersebut.

Nah, pada materi kali ini, kita akan fokus membahas sebuah konsep statistik untuk menilai sebaran dari suatu data yaitu *measuring dispersion* atau sering juga disebut *measures of spread*. Ia akan memberikan informasi terkait seberapa jauh suatu titik data dengan titik data lainnya. Hal ini mampu memberikan kita gambaran variasi nilai dari data tersebut. Untuk menggambarkan measuring dispersion, kita dapat menggunakan empat parameter, yaitu *range*, *interquartile range*, *variance*, dan *standard deviation* [5].

**Range**

Parameter pertama yang akan kita bahas adalah range. Ia merupakan parameter yang digunakan untuk melihat perbedaan antara nilai maksimum dan minimum dari suatu data.

Untuk menghitung parameter ini, kita perlu mengidentifikasi nilai maksimum dan minimum dari suatu data terlebih dahulu. Selanjutnya, kita mengurangi nilai maksimum dan nilai minimum. Selisih inilah yang menjadi parameter range.

**Interquartile Range**

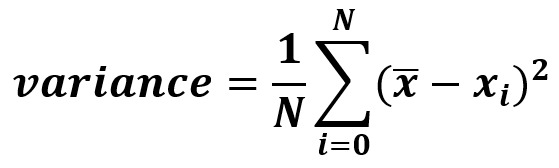
Interquartile Range atau sering disingkat IQR merupakan parameter statistik yang menggambarkan selisih antara kuartil ketiga (Q3) dan kuartil pertama (Q1). Parameter ini mewakili nilai range atau rentang di mana sebagian besar titik data berada.

Untuk menghitung IQR, kita perlu menentukan nilai Q1 dan Q3 terlebih dahulu. Selanjutnya, kita dapat menghitung selisih keduanya. Selisih inilah yang menjadi nilai dari parameter IQR.

**Variance**

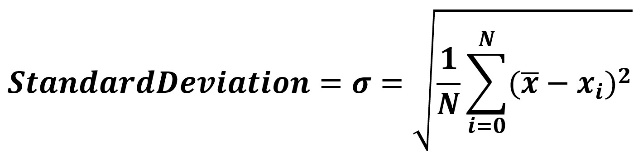
Selain range dan IQR, sebaran dari suatu data juga dapat direpresentasikan menggunakan parameter variance. Ia merupakan parameter yang digunakan untuk menggambarkan besar simpangan suatu titik data dari nilai mean-nya.

Sebelum menghitung nilai variance, kita membutuhkan nilai mean terlebih dahulu. Berikutnya kita perlu menghitung total selisih kuadrat dari seluruh titik data dengan nilai mean-nya. Nilai total tersebut selanjutnya akan dibagi dengan banyak titik data yang dimiliki. Berikut merupakan rumus untuk menghitung variance.



**Standard Deviation**

Jika Anda perhatikan, pada perhitungan variance kita sebenarnya telah menghitung rata-rata selisih kuadrat antar titik data dan nilai mean-nya. Penggunaan selisih kuadrat ini membuat unit dari parameter variance berbeda dengan unit data sehingga cukup sulit untuk diinterpretasikan. Oleh karena itu, kita membutuhkan satu parameter lain yang akan menghitung akar kuadrat dari variance. Parameter tersebut bernama standard deviation.



Standard deviation merupakan parameter yang paling sering digunakan para praktisi data untuk menilai sebaran atau simpangan dalam sebuah data. Semakin kecil nilai standard deviation, semakin kecil pula jarak antar titik data dengan nilai mean-nya sehingga bisa disimpulkan data tersebut memiliki sebaran yang sempit. Di sisi lain, semakin besar nilai standard deviation, semakin luas pula sebaran datanya. Selain itu, standard deviation juga sering digunakan untuk menilai seberapa baik parameter mean dalam merepresentasikan suatu data.

**Measuring Asymmetric**

Pada dua materi sebelumnya, kita telah membahas berbagai parameter statistik yang digunakan untuk mengidentifikasi nilai khas dan sebaran nilai dalam sebuah data. Untuk melengkapi pengetahuan Anda, sekarang kita membahas konsep lain dalam *descriptive statistics*yaitu *measuring asymmetric*.

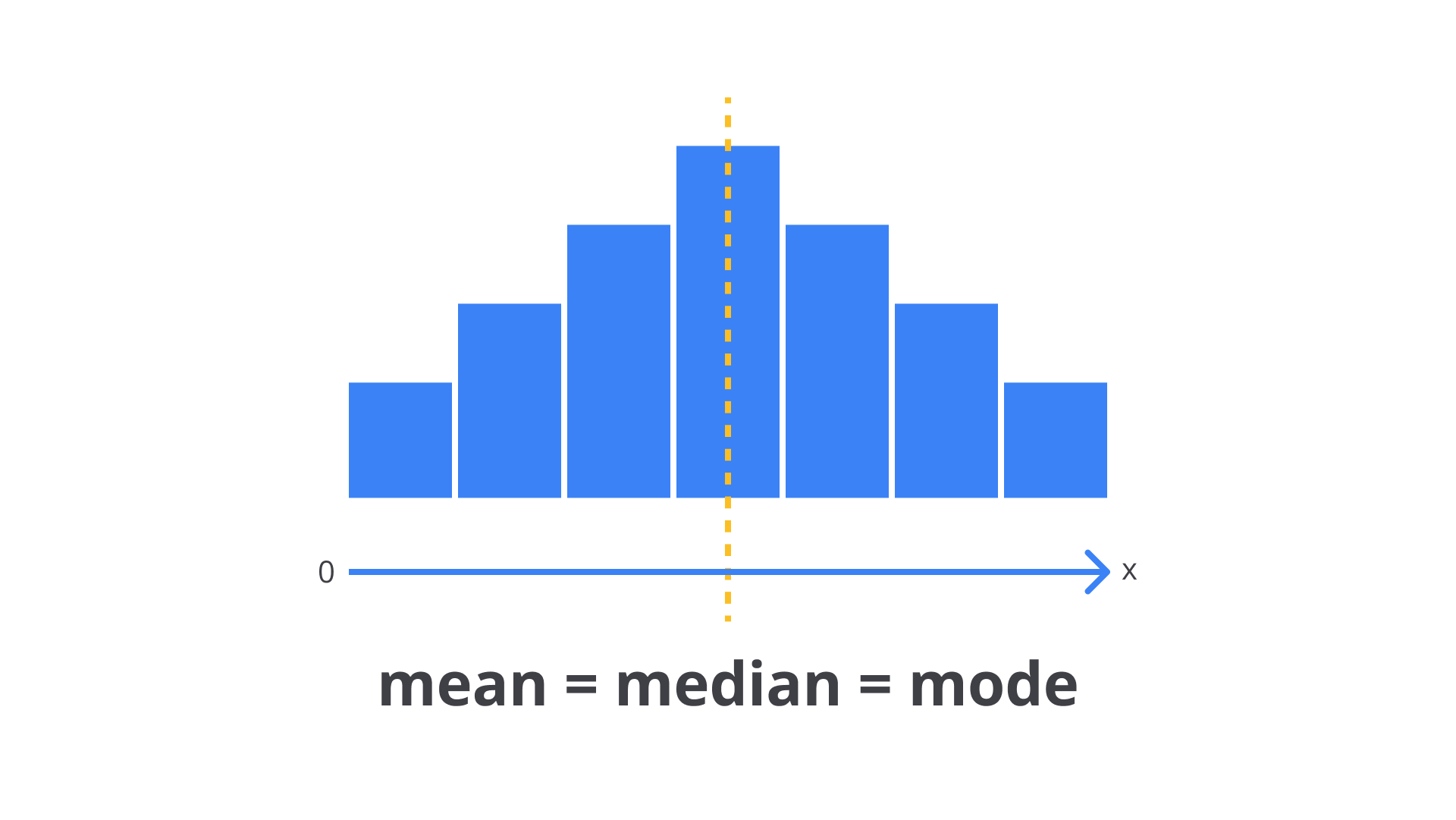
Measuring asymmetric merupakan konsep statistik yang digunakan untuk mengidentifikasi ketidaksimetrisan dalam sebuah distribusi data numerik. Parameter statistik yang umum digunakan dalam measuring asymmetric ialah *skewness*. Namun, sebelum kita membahas lebih jauh terkait parameter skewness, alangkah baiknya kita mulai dari konsep dasar terlebih dahulu yaitu definisi distribusi data.

**Distribusi Data**

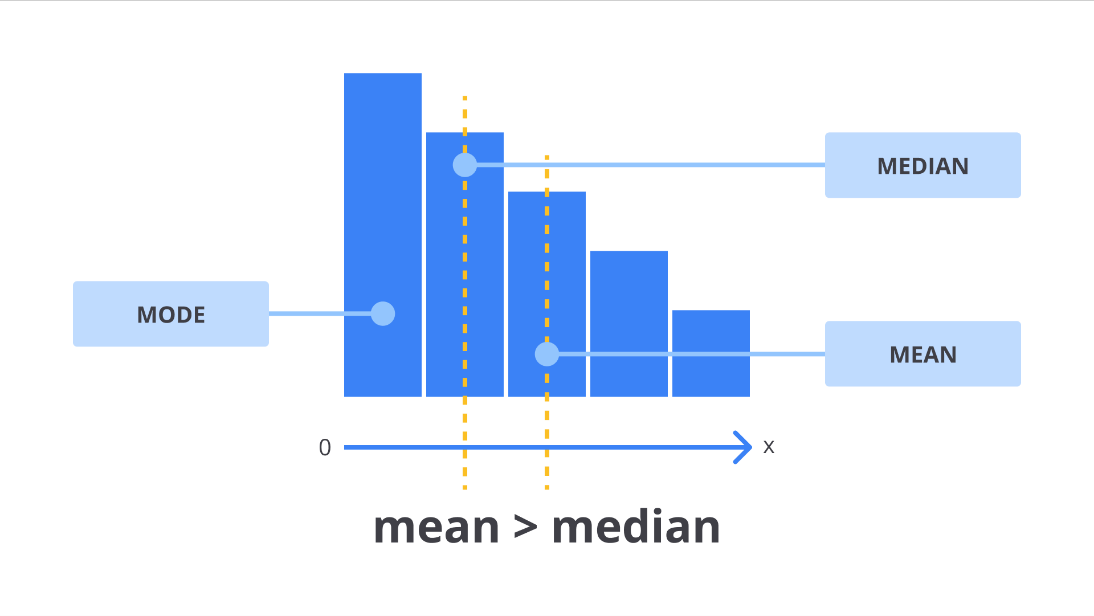
Distribusi data atau *data distribution* merupakan sebuah konsep statistik yang digunakan untuk menunjukkan frekuensi suatu nilai muncul dalam sebuah data. Umumnya kita menggunakan grafik histogram untuk melihat distribusi suatu data. Grafik tersebut akan menunjukkan frekuensi kemunculan setiap nilai dalam sebuah data.

Secara umum, berdasarkan bentuk grafik histogram yang terbentuk, kita bisa membagi distribusi data dalam tiga bentuk yaitu *symmetric distribution*, *right-skewed distribution*, dan *left-skewed distribution*[4].

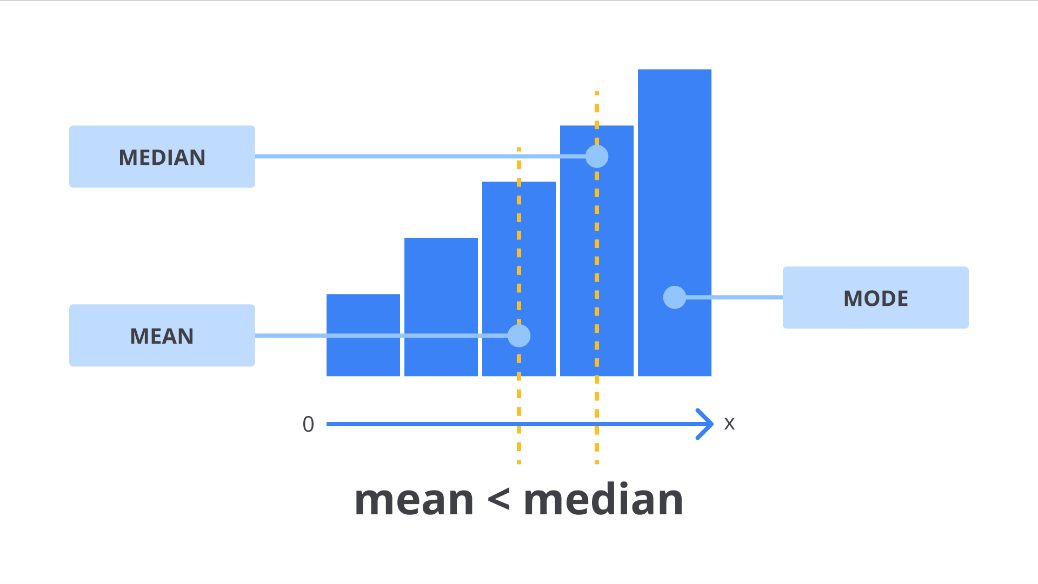
* **Symmetric distribution**  
  Symmetric distribution (*normal distribution* atau *bell curve distribution*) merupakan suatu distribusi data yang memungkinkan kita untuk membagi data menjadi dua bagian secara simetris. Distribusi data ini memiliki nilai mean, median, serta mode yang sama dan berada di sentral distribusi data.



* **Right-skewed distribution**  
  Right-skewed distribution merupakan distribusi data yang memiliki sebagian besar populasi data yang terkonsentrasi pada bagian kiri. Distribusi data ini memiliki nilai mean lebih besar dari nilai median dan juga mode.



* **Left-skewed distribution**  
  Left-skewed distribution merupakan distribusi data yang terjadi ketika sebagian besar populasi data berada pada bagian kanan. Umumnya distribusi ini memiliki nilai median dan mode yang lebih besar dari nilai mean.



Nah, itulah ketiga bentuk distribusi data yang umum dijumpai berdasarkan bentuk grafik histogramnya. Untuk membuat grafik histogram, tentunya kita membutuhkan sebuah tool. Pada bahasa pemrograman Python, terdapat banyak tool yang bisa digunakan untuk membuat plot histogram salah satunya adalah *library* matplotlib.



Grafik histogram di atas memiliki bentuk left-skewed distribution yang kurang sempurna. Bentuk yang tidak sempurna inilah yang paling sering kita temui di industri. Hal ini karena data yang umum dijumpai di industri sangat berantakan dan kotor.

**Skewness**

Oke, kita telah membahas tiga bentuk distribusi data yang dilihat berdasarkan grafik histogram yaitu symmetric distribution, right-skewed distribution, dan left-skewed distribution. Selain melihatnya menggunakan grafik histogram, kita juga bisa mengukur ketidaksimetrisan dalam distribusi data menggunakan parameter skewness.

Skewness merupakan parameter statistik yang digunakan untuk mengukur kesimetrisan sebuah distribusi data. Ia mampu memberikan kita gambaran tentang banyaknya data yang menyimpang dari symmetric distribution. Parameter ini memiliki jangkauan nilai mulai dari negatif hingga positif [5].

* Nilai skewness positif menggambarkan data yang memiliki distribusi yang cenderung right-skewed.
* Nilai skewness nol menggambarkan data yang memiliki distribusi simetris sempurna.
* Nilai skewness negatif merepresentasikan data dengan distribusi yang cenderung left-skewed.

**Data Relationship**

Pada dua materi sebelumnya, kita telah membahas berbagai parameter statistik yang digunakan untuk mengidentifikasi nilai khas dan sebaran nilai dalam sebuah data. Selain itu, kita juga telah membahas parameter statistik untuk menilai ketidaksimetrisan dalam sebuah distribusi data numerik. Parameter tersebut tentunya sangat membantu kita dalam memahami data. Namun, apakah itu sudah cukup? Tentu saja jawabannya tidak.

Semua parameter yang telah kita bahas sebelumnya hanya menggambarkan keadaan satu *feature*/*column*/*variable* dari suatu data. Namun, pada kenyataannya kita sering kali bekerja dengan sebuah data yang memiliki lebih dari satu feature. Oleh karena itu, kita juga perlu mengetahui hubungan antar feature dalam satu data. Nah, untuk melakukannya kita perlu berkenalan dengan konsep lain dalam statistik yaitu *data relationship*.

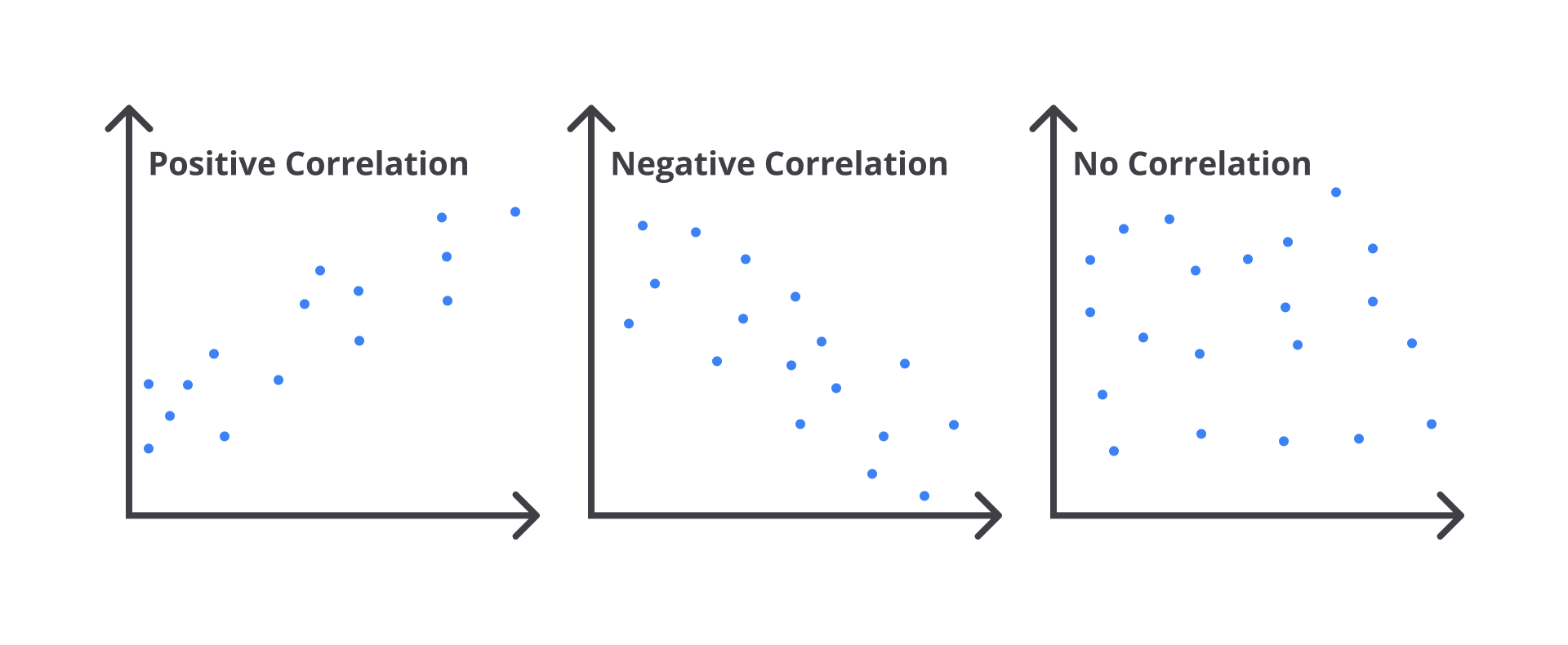
*By definition*, data relationship merupakan konsep statistik untuk mengidentifikasi hubungan antar feature bertipe numerik dalam sebuah data. Untuk mengidentifikasi hubungan antar feature bertipe numerik, kita bisa menggunakan dua parameter yaitu *correlatio*n dan *covariance*.

Bagi Anda yang belum familier dengan kedua parameter tersebut, jangan khawatir karena pada materi ini kita akan membahas keduanya secara tuntas. *So*, tanpa membuang waktu lagi. Yuk, kita simak penjelasan berikut!

**Correlation**

Parameter pertama yang akan kita bahas untuk mengidentifikasi data relationship ialah correlation. Parameter ini digunakan untuk mengidentifikasi korelasi atau hubungan dari dua feature numerik dalam sebuah data. Korelasi ini digambarkan menggunakan nilai dengan rentang -1 hingga 1 [5].

Pada parameter correlation, nilai negatif menggambarkan korelasi berlawanan (*negative correlation*), sedangkan nilai positif merepresentasikan korelasi bersesuaian (*positive correlation*). Jika correlation dari dua feature bernilai nol, keduanya dinyatakan tidak memiliki korelasi (*no correlation*).



**Covariance**

Selain menggunakan correlation, kita juga bisa menggunakan parameter covariance untuk mengidentifikasi hubungan antar dua feature dalam sebuah dataset. Berdasarkan nilainya, covariance terbagi menjadi tiga bagian seperti berikut.

* Positive covariance yang menggambarkan dua feature yang berkorelasi positif atau bersesuaian.
* Negative covariance yang merepresentasikan dua feature yang berkorelasi negatif atau berlawanan.
* Zero covariance yang menandakan dua feature yang tidak berkorelasi satu sama lain.