**Apa itu Statistik Deskriptif ?**

Statistik deskriptif/Statistik deskriptif melibatkan meringkas/meringkas dan mengatur data sehingga dapat dengan mudah dijangkau. Tidak seperti statistik inferensial, statistik deskriptif tidak mencoba mendapatkan kesimpulan dari sampel ke seluruh populasi, tidak juga menggunakan probabilitas.

Dibagi menjadi dua kategori:

* Measures of central tendency
* Measures of variability (spread)

**Sampel dan Populasi**

Populasi adalah sekumpulan dari semua elemen yang kita minati. Jumlahnya bisa sangat banyak sehingga tidak cocok untuk *collecting* dan *analyzing* data. Nah, kesimpulan biasanya didapatkan dengan memilih dan memeriksa subkumpulan perwakilan dari populasi.Subset populasi ini disebut sampel. Idealnya, sampel harus mempertahankan fitur statistik esensial dari populasi hingga tingkat yang memuaskan.

**Outliers**

Titik data yang berbeda secara signifikan dari sebagian besar data yang diambil (entah dari sampel atau populasi).

Beberapa penyebab outliers

* Natural variation in data
* Change in the behaviour of observed system
* Error in data collection (common, misalnya keterbatasan instrumen atau prosedur pengukuran)

Tidak ada definisi matematis yang tepat tentang outlier, kita harus mengandalkan pengalaman, pengetahuan tentang subjek yang diminati, common sense untuk mengetahui data itu pencilan dan cara menanganinya.

**Measure of Central Tendency**

Mengacu pada gagasan bahwa ada satu bilangan yang paling baik untuk meringkas rangkaian, sebuah bilangan yang dalam beberapa hal merupakan “central” dari himpunan tersebut.

* Mean dianggap sebagai ukuran tendensi sentral yang paling andal untuk membuat asumsi tentang suatu populasi dari satu sampel.
* Median jauh lebih sedikit dipengaruhi oleh outlier dan skewed daripada rata-rata.
* Mode juga merupakan satu-satunya ukuran tendensi sentral yang dapat digunakan untuk variabel kategori karena kita tidak dapat menghitungnya, misalnya rata-rata untuk variabel "gender" atau "warna". Tetapi mungkin ada kumpulan data di mana tidak ada mode sama sekali karena semua nilai muncul dalam jumlah yang sama. Jika dua nilai muncul pada waktu yang sama dan lebih dari nilai lainnya maka kumpulan datanya adalah bimodal. Jika tiga nilai muncul pada waktu yang sama dan lebih dari nilai lainnya maka kumpulan datanya adalah trimodal dan untuk n mode, kumpulan data tersebut adalah multimodal.

**Measure of Spread / Dispersion**

Gagasan variabilitas dalam data kita. Juga digunakan untuk mengukur seberapa “bagus” mean merepresentasikan kumpulan data.

* Standard Deviation

Bagaimana data disebarkan/seberapa tersebar data dari mean. Standar deviasi rendah 🡪 titik data cenderung mendekati rata-rata kumpulan data. Sebaliknya. Ada situasi kita harus memilih sample atau population Standard Deviation. Meskipun sampel adalah bagian dari populasi, rumus SD mereka seharusnya sama, padahal sebenarnya tidak. Karena ini statistik deskriptif maka kita akan berusan dengan sample, jadi gunakan rumus Sample SD.

* Mean Deviation

Rata-rata perbedaan absolut antara setiap nilai dengan rata-rata semua nilai.

* Variance

Kuadrat jarak rata-rata antara setiap kuantitas dan mean. Kuadrat dari standar deviasi.

* Range

Perbedaan nilai terendah dan tertinggi.

* Persentil

Cara untuk merepresentasikan posisi suatu nilai dalam kumpulan data. Sebelumnya nilai harus diurut naik.

* Quartiles

Nilai yang membagi data kita menjadi beberapa quarters jika data diurutkan dalam urutan naik. Ada 3 quartil, jika dalam persentil, 25%, 50%, 75%

* Skewness

Ukuran asimetri distribusi probabilitas dari real-valued random variabel tentang meannya. Ada positive skew, ketika distribusi miring ke kanan, dapat diintreprestasikan Mean lebih besar daripada mode. Sebaliknya. Dan kalau mode, median, mean sama artinya distribusinya normal. Bisa dihitung dengan mencari skewness coefficient. Ada dua cara mengitung skewness coefficient : Pearson First dan Perason Second. Nilai negatif berarti distribusinya condong negatif. Nilai positif berarti distribusinya condong positif. Koefisien skewness pertama Pearson menggunakan mode. Oleh karena itu, jika frekuensi sangat rendah maka tidak akan memberikan ukuran yang stabil dari tendensi sentral.

* Kurtosis

Apakah data bersifat heavy-tailed (banyak outlier) atau light-tailed (kurang outlier) relatif terhadap distribusi normal. 3 tipe kurtosis adalah Mesokurtik (serupa dengan distribusi normal), Leptokurtik (lebih besar daripada Mesokurtik), Platykurtik,lebih rendah. Perbedaan utama antara skewness dan kurtosis adalah skewness mengacu pada tingkat simetri, sedangkan kurtosis mengacu pada tingkat keberadaan pensilan dalam distribusi.

* Correlation

Teknik statistik yang menunjukkan apakah dan seberapa kuat variabel berkaitan. Hasil utama korelasi disebut koefisien korelasi (atau “r”). Koefisien korelasi berkisar dari -1,0 hingga +1,0. Semakin dekat r ke +1 atau -1, semakin erat hubungan kedua variabel tersebut. Jika r mendekati 0, artinya tidak ada hubungan antar variabel. Jika r positif, berarti semakin besar satu variabel, variabel lainnya semakin besar. Jika r negatif berarti semakin besar satu variabel, variabel lain akan semakin kecil (sering disebut korelasi “invers”).

**Empirical Rules**

The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by σ) of the mean or average (denoted by µ).

In particular, the empirical rule predicts that 68% of observations falls within the first standard deviation (µ ± σ), 95% within the first two standard deviations (µ ± 2σ), and 99.7% within the first three standard deviations (µ ± 3σ).

Implementasi

* Mean : Weighted Mean, Harmonic Mean,