

Nama : Anindita Khusnul Oktavia
Program : Introduction to Python for Data Science

SUMMARY

DESCRIPTIVE STATISTICS

Measure of central tendency : berfungsi untuk meringkas seluruh rangkaian pengukuran.

1. Mean

Mean/average : Angka tunggal yang dapat memperkirakan nilai seluruh kumpulan data yang dianggap sebagai ukuran tendensi sentral yang paling andal untuk membuat asumsi tentang suatu populasi dari satu sampel.

Sample mean/sample arithmetic/average: rata-rata aritmatika dari semua item dalam kumpulan data.

Weighted mean/weighted arithmetic/weighted average : generalisasi dari rata-rata aritmatika yang memungkinkan untuk menentukan kontribusi relatif dari setiap titik data ke hasil.

Harmonic mean : reciprocal dari mean dari semua item dalam data set. Reciprocal artinya $1/\text{value}$.

Geometric mean : akar ke- n dari hasil kali semua n elemen x_i dalam data set. Ukuran ini berguna untuk membandingkan berbagai hal dengan properti yang sangat berbeda.

2. Median

Median : nilai yang membagi data dalam 2 bagian yang sama.

Sample median : elemen tengah dari kumpulan data yang diurutkan dalam urutan naik atau turun.

3. Modus

Modus : nilai yang mempunyai frekuensi tertinggi.

Sample mode : nilai dalam kumpulan data yang paling sering muncul. Jika tidak ada satu pun nilai, maka set tersebut multi modal karena memiliki beberapa nilai modal.

Measure of spread / Dispersion : merupakan variabeliats dalam data dan berfungsi untuk mengukur sebarapa “bagus” mean merepresentasikan data.

1. Standar Deviation

Standar deviation atau simpangan baku : pengukuran jarak rata-rata antara setiap besaran dengan mean.

Standar deviation merupakan positive square root dari sample variance.

Mean deviation atau mean absolute deviation : rata rata perbedaan absolute antara setiap nilai dalam satu set nilai dengan rata-rata semua nilai dari set itu.

2. Variance

Variance : kuadrat jarak rata-rata antara setiap kuantitas dan mean.

Sample variance : menunjukkan secara numerik seberapa jauh titik data dari mean.

3. Skawness

Skawness : ukuran asimetri distribusi probabilitas dari real valued random variabel tentang meannya à mengukur asimetri sampel data.

Negative skewness menunjukkan bahwa ada ekor dominan di sisi kiri, yang dapat kita lihat pada set pertama. Positive skewness values sesuai dengan ekor yang lebih panjang di sisi kanan, yang dapat kita lihat pada set kedua. Jika skewness mendekati 0 (misalnya antara -0.5 dan 0.5), maka dataset dianggap cukup simetris.

4. Pecentile

Sample p percentile adalah elemen dalam kumpulan data sehingga $p\%$ elemen dalam kumpulan data kurang dari atau sama dengan nilai tersebut. Juga, $(100 - p)\%$ elemen lebih besar dari atau sama dengan nilai tersebut.

5. Kuartiles

Quartiles : nilai yang membagi data menjadi beberapa quarters jika data diurutkan dalam urutan ascending atau menaik.

- **First quartile** adalah sampel persentil ke-25. Kuartil pertama membagi sekitar 25% item terkecil dari kumpulan data lainnya.
- **Second quartile** adalah sampel persentil ke-50 atau **median**. Kira-kira 25% item terletak di antara kuartil pertama dan kedua dan 25% lainnya antara kuartil kedua dan ketiga.
- **Third quartile** adalah sampel persentil ke-75. Kuartil ketiga membagi sekitar 25% item terbesar dari sisa kumpulan data.

6. Ranges

Rentang data/range data adalah perbedaan antara elemen maksimum dan minimum dalam kumpulan data à perbedaan antara nilai terendah dan tertinggi.

Interquartile range adalah perbedaan antara kuartil pertama dan ketiga. Setelah kita menghitung kuartil, kita dapat mengambil selisihnya:

7. Kurtosis

Kurtosis : ukuran apakah data bersifat heavy-tailed (banyak outlier) atau light-tailed (kurang outlier) relatif terhadap distribusi normal.

Measures of Correlation Between Pairs of Data

- Positive correlation exists when larger values of x correspond to larger values of y and vice versa.
- Negative correlation exists when larger values of x correspond to smaller values of y and vice versa.
- Weak or no correlation exists if there is no such apparent relationship.

1. Covariance

Sample covariance adalah ukuran yang mengukur kekuatan dan arah hubungan antara sepasang variabel:

- If the correlation is positive, then the covariance is positive, as well. A stronger relationship corresponds to a higher value of the covariance.
- If the correlation is negative, then the covariance is negative, as well. A stronger relationship corresponds to a lower (or higher absolute) value of the covariance.
- If the correlation is weak, then the covariance is close to zero.

2. Correlation Coefficient

Coefficient adalah ukuran lain dari korelasi antar data, ukuran ini juga sering disebut dengan standardized covariance.

- The value $r > 0$ indicates positive correlation.
- The value $r < 0$ indicates negative correlation.
- The value $r = 1$ is the maximum possible value of r . It corresponds to a perfect positive linear relationship between variables.
- The value $r = -1$ is the minimum possible value of r . It corresponds to a perfect negative linear relationship between variables.
- The value $r \approx 0$, or when r is around zero, means that the correlation between variables is weak.

Untuk implementasinya ada pada file .ipynb