IT496: Introduction to Data Mining

Lecture - 03

Statistics for Data Mining - II

[Measures of Dispersion]

Arpit Rana

28th July 2023

Measures of Spread (Dispersion)

. . .about the variety of the values. . .

Disclaimer: All images incorporated within the presentation slides have been sourced from a diverse array of online platforms.

Measures of Spread (Dispersion)

A measure of spread is used to describe the *variability* in a sample or population.

Why is it important?

- Mostly used in conjunction with a measure of central tendency.
- It gives us an idea of how well the measure of central tendency represents the data.

Example

- If the spread of values in the dataset is large, the mean is not as representative of the data as if the spread of data is small.
- This is because a large spread indicates that there are probably large differences between individual scores.

Range

The range of the set is the *difference between the largest and smallest values*.

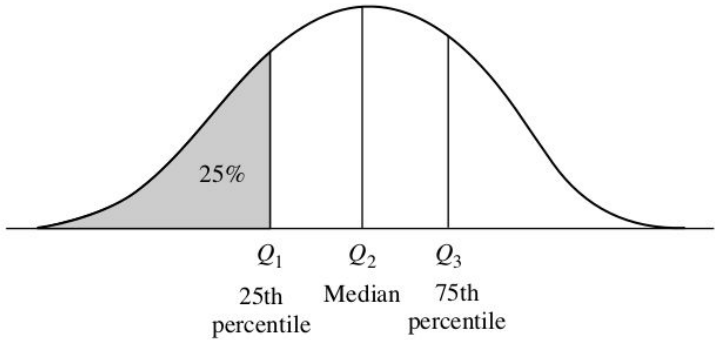
Let $x_1, x_2, \ldots, x_N$ be the set of N observed values or observations for a numeric attribute X.

$$\text{Range} = \max(X) - \min(X)$$

Quantiles

The data points that *split the data distribution into equal-size consecutive sets* are called ***quantiles***.

- The 2-*quantile* is the data point dividing the lower and upper halves of the data distribution. It corresponds to the *median*.
- The 4-*quantiles* are the three data points that split the data distribution into four equal parts. They are referred to as *quartiles*.
- The 100-*quantiles* divide the data distribution into 100 equal-sized consecutive sets. They are referred to as *percentiles*.



Quantiles

In general

- Suppose that $x_1, x_2, \ldots, x_N$ be the set of N observed values or observations for a numeric attribute X sorted in increasing order.
- The k^{th} q-**quantile** for a given data distribution is the value x such that at most k/q of the data values are less than x and at most $(q - k)/q$ of the data values are more than x.
- Here k is an integer such that $0 < k < q$. There are q–1 q-quantiles.

Interquartile Range (IQR)

The distance between the *first* and *third* quartiles gives the range covered by the middle half of the data.

This distance is called the interquartile range (IQR).

$$IQR = Q3 - Q1$$

- A common rule of thumb for identifying suspected **outliers** is to figure out values falling at least **1.5 × IQR** above the third quartile or below the first quartile.

Why is the scale **1.5** and not any other number?

Five-Number Summary

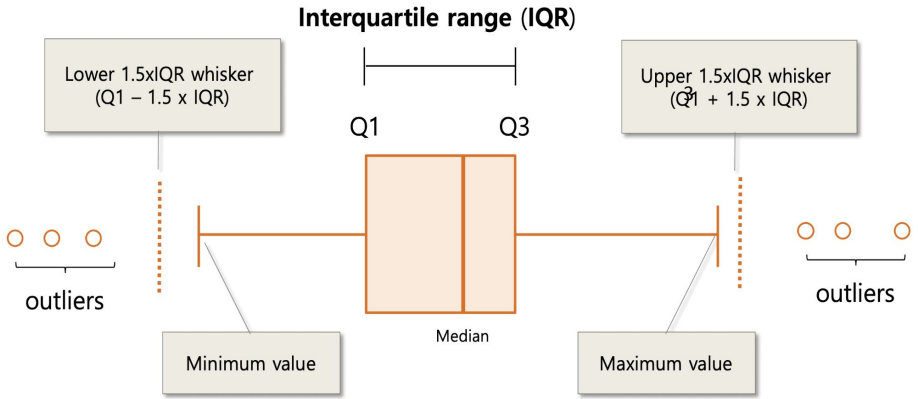
The median (Q2) and the quartiles Q1 and Q3 together contain no information about the tailing ends of the data.

The *five-number summary* of a distribution includes the *smallest* and *largest* individual observations, and is written in the order of *Minimum*, Q1 , *Median*, Q3 , *Maximum*.

Boxplot

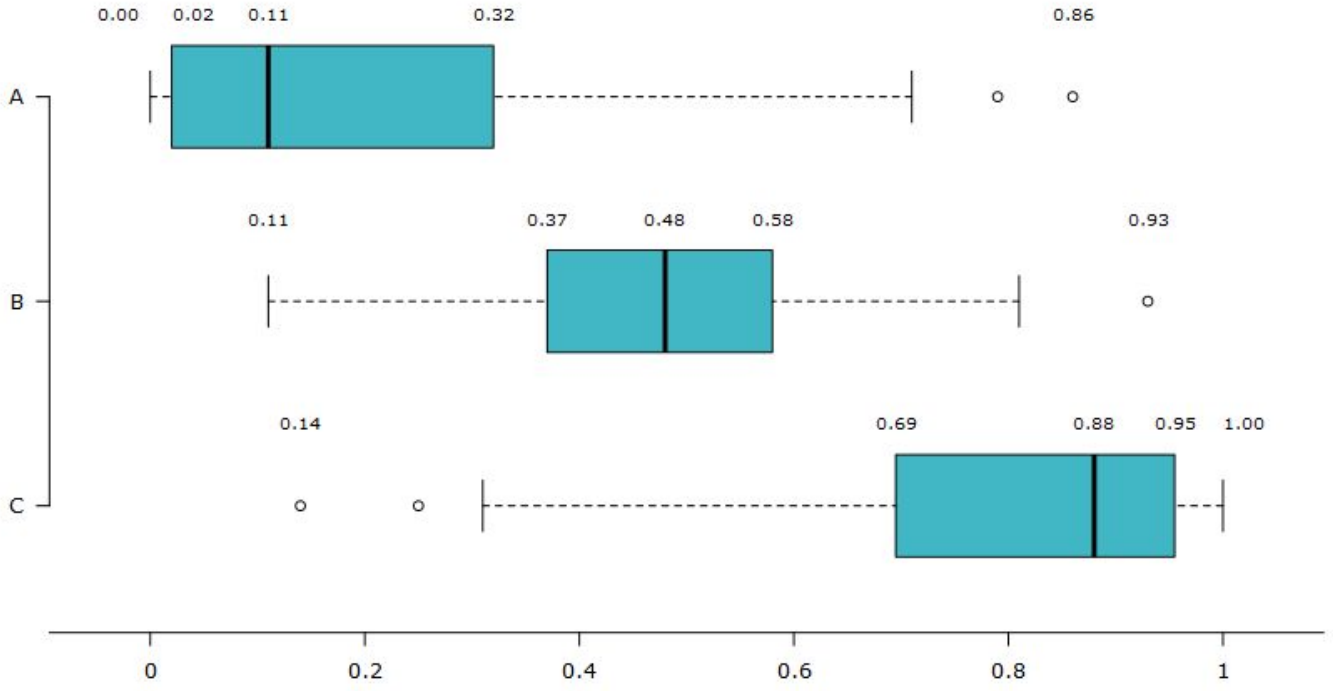
Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the *five-number summary* as follows:

- Typically, the ends of the box are at the quartiles so that the box length is the IQR.
- The median is marked by a line within the box.
- Two lines (called whiskers) outside the box extend to the smallest (Minimum) and the largest (Maximum) observations.



Boxplot

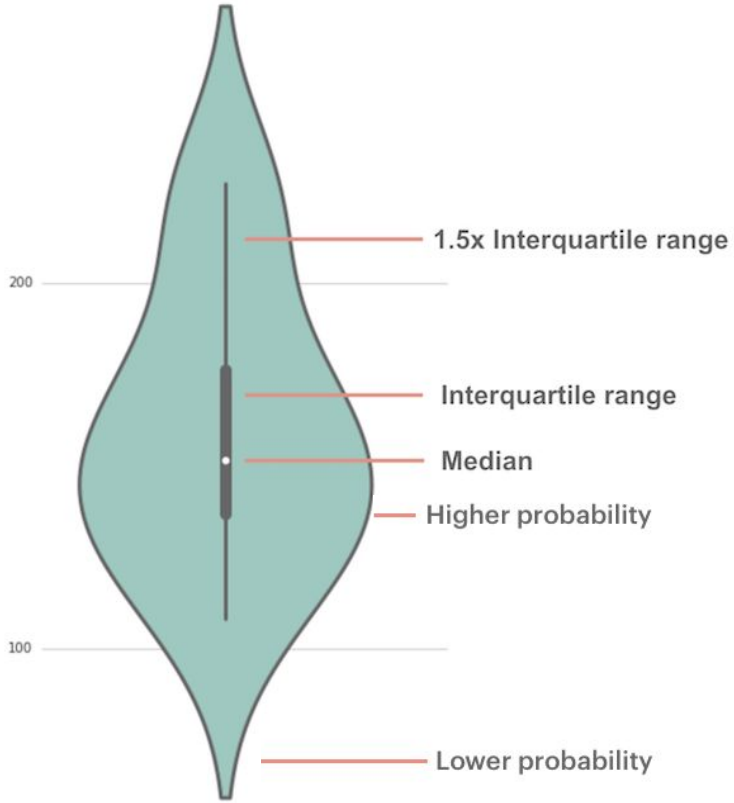
Skewed data using boxplot can be represented as below.



Violin Plot

Violin plots depict summary statistics and the density of each variable.

- the *white dot* represents the median
- the *thick gray bar* in the center represents the interquartile range
- the *thin gray line* represents the rest of the distribution, except for points that are determined to be “outliers”.



Variance and Standard Deviation

- Quantiles do not take into account every score in the data.
- To get a more representative idea of spread we need to take into account the actual values of each score in a dataset.

Variance and Standard Deviation

Variance and standard deviation indicate how spread out a data distribution is.

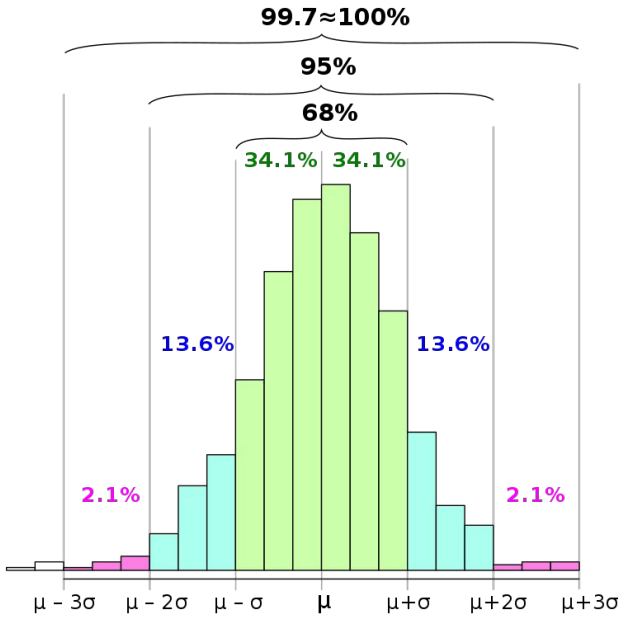
Let $x_1, x_2, \ldots, x_N$ be the set of N observed values or observations for a numeric attribute X.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The standard deviation (σ) of the observations is the square root of the variance (σ^2).

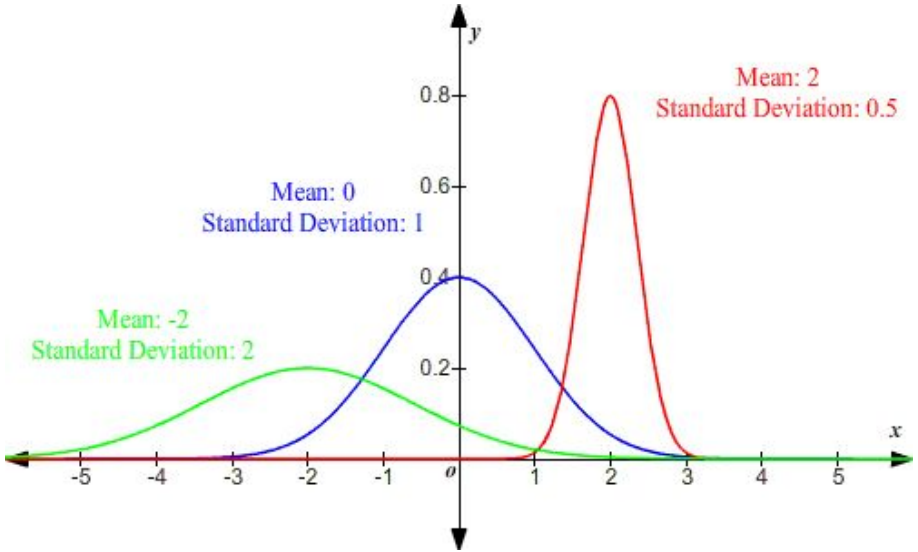
Properties on Standard Deviation

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of central tendency.
- σ = 0 only when there is no spread, that is, when all observations have the same value. Otherwise, σ > 0.



Most observations lie within a few standard deviations around the mean.

Properties on Standard Deviation



- A low standard deviation means that the data observations tend to be very close to the mean,
- while a high standard deviation indicates that the data are spread out over a large range of values.

Next lecture

Measures of Proximity

1st August 2023