Data Science Analysis Lab

Lecture Notes

Metin Oktay Yılmaz

Systems Design Master
FH Kärnten
Austria

Contents

1 Univariate Data

07.03.2023 - Lecture 1 - EDV1

Let's say we are measuring x. i substring denotes the measured value of x.
x_i: i^{th} measurement value in chronological order.
$x_{(i)}$: i^{th} measurement value in a ranked (or sorted) order as in:

$$x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$$

Mean value of a data set:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Property of the mean value:

$$x_1 - \bar{x}; x_2 - \bar{x}; ...; x_n - \bar{x}$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - \bar{x}n = 0$$

median, α - quantile
sorted dataset

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)}$$

Example: $n = 11$

The on that is right in the middle, the center of the sorted data set, is called the median. If the size of the data set is an even number, then the median is:

$$\tilde{x} = \frac{x_{0.5n} + x_{0.5n+1}}{2}$$

Measures of Dispersion or Variance

Range:

$$R = max - min = x_{(n)} - x_{(1)}$$

Dispersion: Is the deviation of a value within the data set, from the mean value of the data set.

$$\frac{1}{n}\sum_{u=1}^{u}(x_{(i)} - \bar{x})$$

Mean absolute deviation (MAD):

$$MAD = \frac{1}{u}\sum_{i=1}^{u} |x_{(i)} - \bar{x}|$$

Standard Deviation:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2}$$

$$\sum_{u=1}^{u}(x_{(i)} - \bar{x}) = 0$$

The number of independent terms is called "*degree of freedom*". Standard deviation is sensitive to outliers.

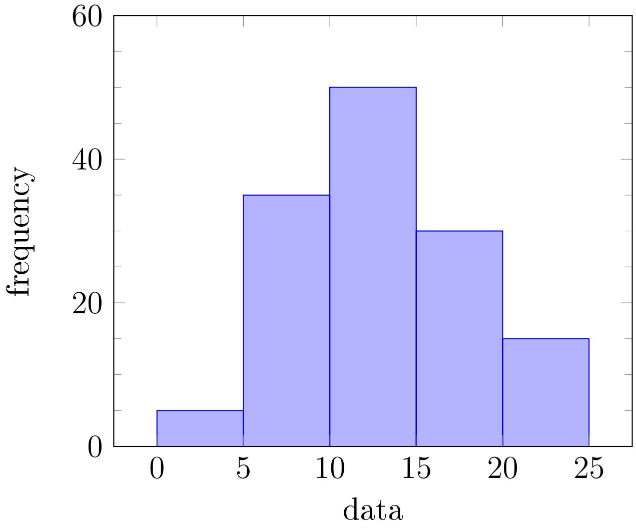
Interquantile Range (IQR)

$$IQR = Q(0.75) - Q(0.25)$$

Note: IQR contains %50 of all values.

Graphical Methods

Histogram:



Best class width for a histogram is:

$$\boxed{2 * IQR * n^{-1/3}}$$

Box and Whisker Plot:

Pattern Recognition