Introduction to Data Analytics

ITE 5201

Lecture6-Correlation Analysis

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

Pearson correlation coefficient

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson’s r.

The correlation coefficient is used to examine the relationship between two sets of data.

The value of the correlation coefficient tells us about the strength and the nature of the relationship.

Correlation coefficient values can range between +1.00 to -1.00.

The Pearson is a measure of the linear correlation between two variables X and Y.

Pearson Correlation Coefficient

- R=1   Strong Positive relationship
- R=0   Not linearly correlated
- R= -1 Strong negative relationship

The Pearson correlation assumes

Your data is normally distributed

You have continuous, numeric variables

Your variables are linearly related

Use Pearson variables

- To uncover (linear) relationships between variables
- Not to rule out possible (nonlinear) relationships between variables

Non-parametric correlation analysis

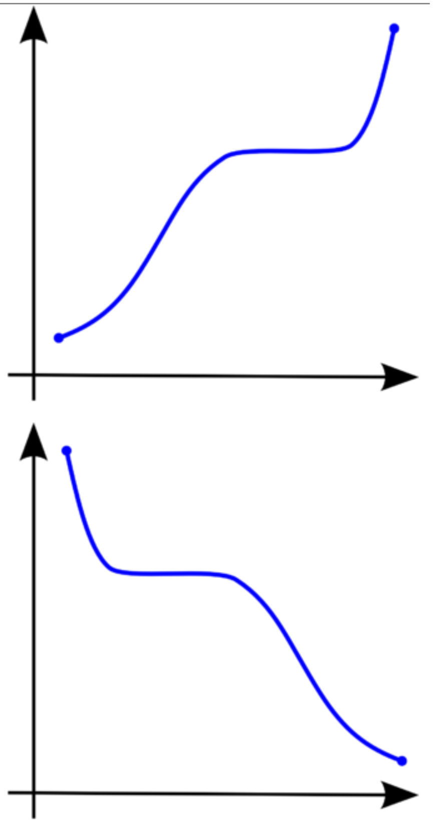
- You can use nonparametric correlation analysis to find correlation between categorical, nonlinearly related, non-normally distributed variables.

  - Spearman’s rank correlation
  - Chi-square tables

Spearman’s rank correlation

- In statistics, Spearman's rank correlation coefficient or Spearman's ρ, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables).

- Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)



Spearman's rank Correlation Coefficient

- r_s=1   Strong Positive relationship
- r_s =0   Not linearly correlated
- r_s = -1 Strong negative relationship

The Spearman's rank correlation assumes

Your variable are ordinal; numeric, but able to be ranked like a categorical variable

Your variables are related nonlinearly

Your data is non-normally distributed

Chi-square test

A chi-squared test for independence tests if there is a significant relationship between two or more groups of categorical data from the same population.

The null hypothesis for this test is that there is no relation.

In general the chi-square analysis is used when there is a need to examine the similarities between two or more populations or variables on some characteristics of interest.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying

Example: a scientist wants to know if education level and marital status are related for all people in some country.

x^2	=	chi squared
O_i	=	observed value
E_i	=	expected value

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-square test for independence

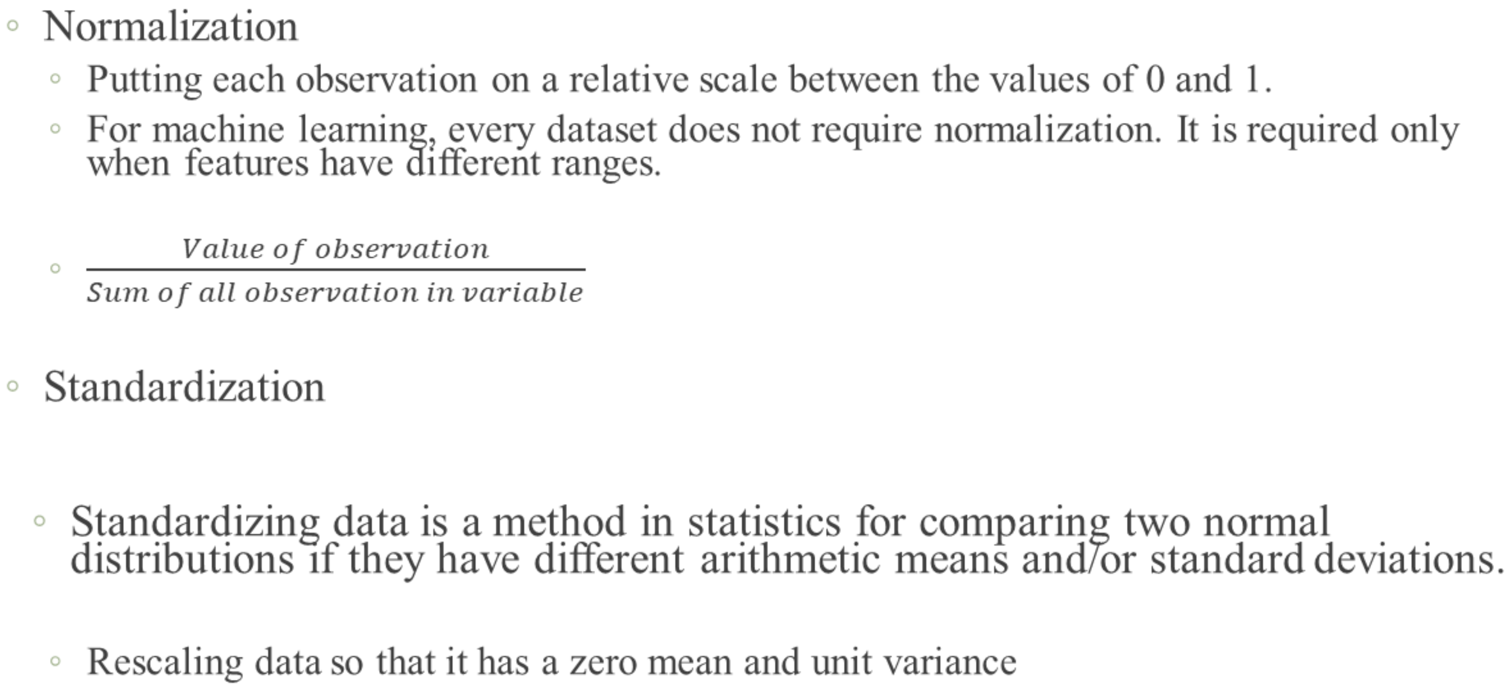
- $p < 0.05$ Reject null hypothesis and conclude that the variables are correlated
- $P > 0.05$ Accept null hypothesis and conclude that the variables are independent
- the null hypothesis refers to a general or default position: that there is no relationship between two measured phenomena

Scaling the data

- Differing magnitude among variables do not produce misleading statistic.

- To prepare your data for machine learning.

Scaling data



- Normalization
  - Putting each observation on a relative scale between the values of 0 and 1.
  - For machine learning, every dataset does not require normalization. It is required only when features have different ranges.
  - $\frac{Value\ of\ observation}{Sum\ of\ all\ observation\ in\ variable}$
- Standardization
  - Standardizing data is a method in statistics for comparing two normal distributions if they have different arithmetic means and/or standard deviations.
  - Rescaling data so that it has a zero mean and unit variance

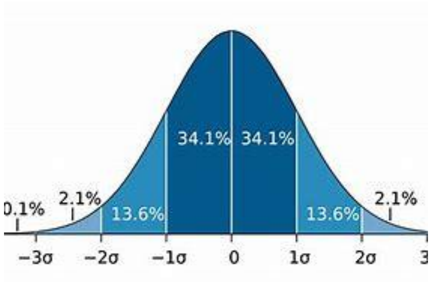
Standardization" typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean.

Descriptive statistics

- Describe the value of observations in a variable
  - Sum
  - Median
  - Mean
  - Max
  - Standard deviation
  - Variance

Standard Deviation & Variance

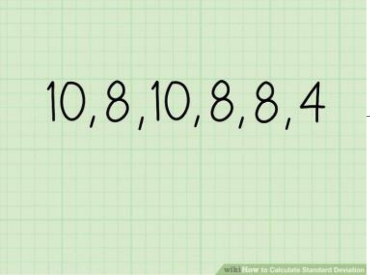
- Standard deviation
  - Standard deviation is a number used to tell how measurements for a group are spread out from the average (mean) or expected value.
  - A low standard deviation means that most of the numbers are very close to the average. A high standard deviation means that the numbers are spread out



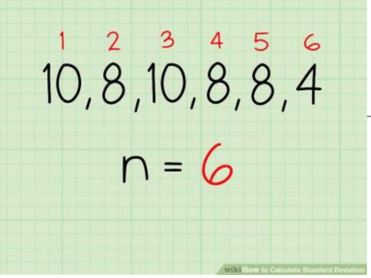
- Variance
  - The variance in probability theory and statistics is a way to measure how far a set of numbers is spread out.
  - variance is defined as the average of the squares of the differences between the individual (observed) and the expected value.

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\bar{X}$ = mean of all values in the data set,
N = number of values in the data set.

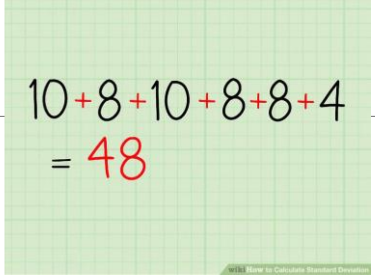


$$10, 8, 10, 8, 8, 4$$

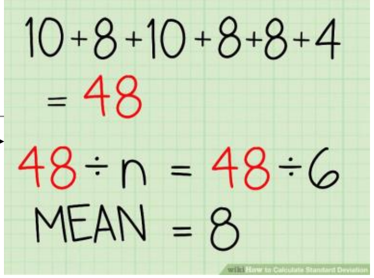


$$\overset{1}{10}, \overset{2}{8}, \overset{3}{10}, \overset{4}{8}, \overset{5}{8}, \overset{6}{4}$$

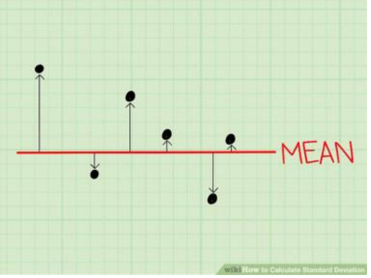
$$n = 6$$

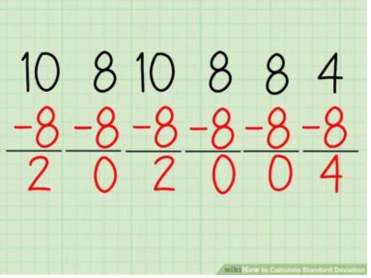


$$10 + 8 + 10 + 8 + 8 + 4$$
$$= 48$$

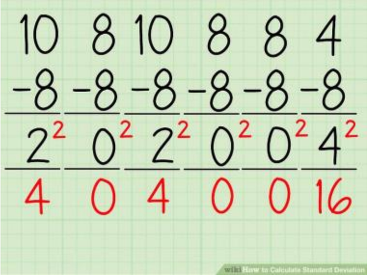


$$10 + 8 + 10 + 8 + 8 + 4$$
$$= 48$$
$$48 \div n = 48 \div 6$$
$$\text{MEAN} = 8$$

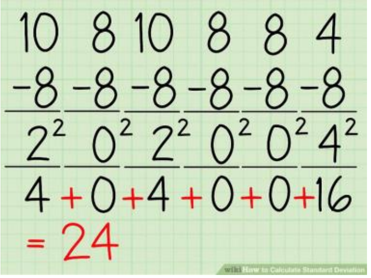




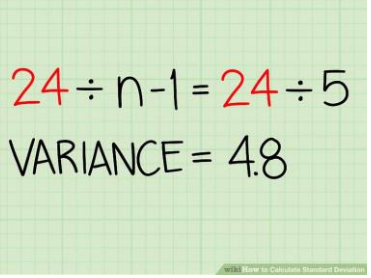
$$\begin{array}{cccccc} 10 & 8 & 10 & 8 & 8 & 4 \\ -8 & -8 & -8 & -8 & -8 & -8 \\ \hline 2 & 0 & 2 & 0 & 0 & 4 \end{array}$$



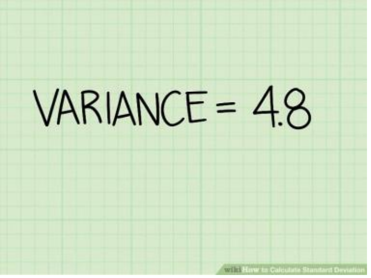
$$\begin{array}{cccccc} 10 & 8 & 10 & 8 & 8 & 4 \\ -8 & -8 & -8 & -8 & -8 & -8 \\ \hline 2^2 & 0^2 & 2^2 & 0^2 & 0^2 & 4^2 \\ \hline 4 & 0 & 4 & 0 & 0 & 16 \end{array}$$



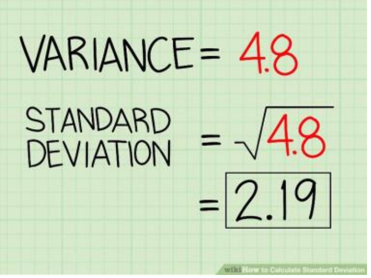
$$\begin{array}{cccccc} 10 & 8 & 10 & 8 & 8 & 4 \\ -8 & -8 & -8 & -8 & -8 & -8 \\ \hline 2^2 & 0^2 & 2^2 & 0^2 & 0^2 & 4^2 \\ \hline \end{array}$$
$$4 + 0 + 4 + 0 + 0 + 16$$
$$= 24$$



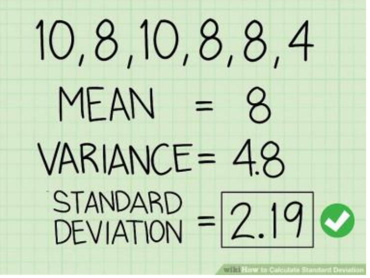
$$24 \div n - 1 = 24 \div 5$$
$$\text{VARIANCE} = 4.8$$



$$\text{VARIANCE} = 4.8$$



$$\text{VARIANCE} = 4.8$$
$$\text{STANDARD DEVIATION} = \sqrt{4.8}$$
$$= \boxed{2.19}$$



$$10, 8, 10, 8, 8, 4$$
$$\text{MEAN} = 8$$
$$\text{VARIANCE} = 4.8$$
$$\text{STANDARD DEVIATION} = \boxed{2.19}$$ ✓

Uses of Descriptive statistics

- Detecting outliers
- Planning data preparation requirements for machine learning
- Selecting features for use in machine learning