# Data Analytics

## Purpose of DA

1.main focus and method: regression, classification and clustering

2.data

* make sure you do understand the meaning of the data and sure it successfully reflects something in the reality
* types of data: find the best way to represent data, nominal, ordinal, interval or ratio (you should make it understandable and suitable)

3.visulization

* when we draw, we always want to show the trend and the relationship or meaning by comparison
* avoid to plot more than three dimensions in a chart
* be direct. Don’t let the reader spend time to guess what’s your conclusion
* some examples: box chart, line chart, bar chart, pie chart, word cloud, scatter plot

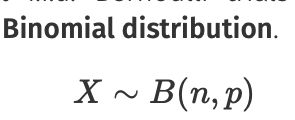
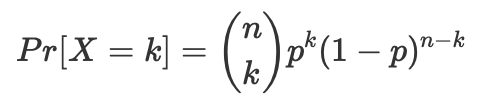
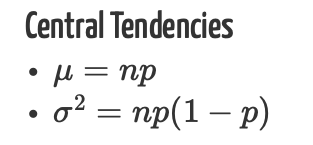
## statistic knowledge

1.statistical attributes and what kind of information they reflect

* mean, median and mode. Mean can reflect the average situations. However, it can be affected by some extreme data. Median can be more robust when faced with extreme data. Mode reflects the most case in the data.
* we can also learn something from data point ranks 25%, 75%, etc.
* correlation between variables can reflect how they related and react to each other

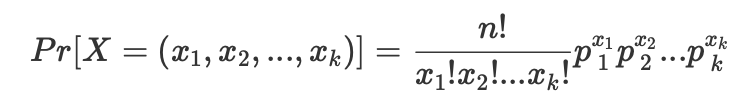
2. probability and distribution: from different ways the probability can be generated, we can have different distributions. In this distribution, we can know all the possibilities and for the possibility for some to happen. (which calculated by using the space of some event divided by sample space) This can be very direct for us to know the information contains in the data.

* two kinds of distribution
  + when it is continuous, for each point the possibility is 0, and we calculation the area below the pdf (probability density function) to calculate the probability for some events to happen.
  + When it is discrete, for each point the possibility is the possibility of this event to happen, which is called pmf(probability of mass function)
  + Some typical distributions are binomial distribution(describe the situation when a lot of Bernoulli trials has been conducted) and Gaussian distribution(norm distribution), the related formular are as follows:

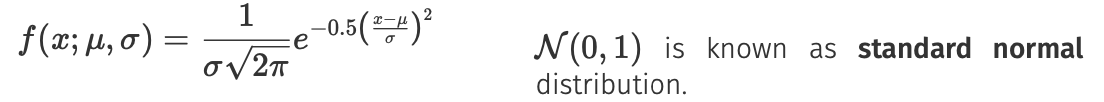
// n presents the times of trials





// n is the times of trial, and x1 is the number of x == value1 to happen



// python scipy.statis can help calculate pdf, pmf and cdf as well as inverse ppf(give probability return the possible event)

3.according to central limit theorem, we can assume the sample can reflect the population, thus, from the sample we can have some confidence about what the population is like. However, we should assure, the frenquency of some sample appear in the sample should be reflective that in the population.

