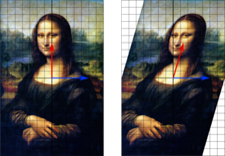
**Miscellaneous Statistical Knowledge**

Eigenvectors and values



Sum of the Squared Residuals: a type of loss function

So RSS is a **measure** of how good the model approximates the data while OLS is a **method** of constructing a good model.

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory), the **central limit theorem** (**CLT**) establishes that, in some situations, when [independent random variables](https://en.wikipedia.org/wiki/Statistical_independence) are added, their properly normalized sum tends toward a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) (informally a "*bell curve*") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

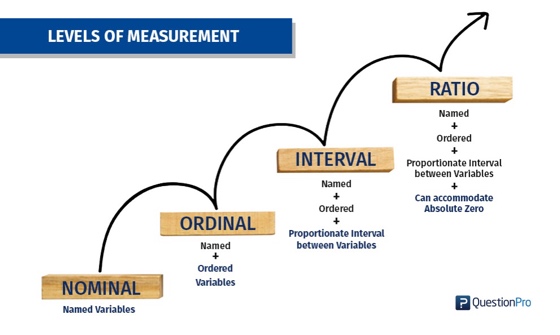
The chi-square distribution is used in the common [chi-square tests](https://en.wikipedia.org/wiki/Chi-square_test) for [goodness of fit](https://en.wikipedia.org/wiki/Goodness_of_fit) of an observed distribution to a theoretical one, the [independence](https://en.wikipedia.org/wiki/Statistical_independence) of two criteria of classification of [qualitative data](https://en.wikipedia.org/wiki/Data_analysis), and in confidence interval estimation for a population [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as [Friedman's analysis of variance by ranks](https://en.wikipedia.org/wiki/Friedman_test).

A **unit root** (also called a **unit root** process or a difference stationary process) is a stochastic trend in a time series, sometimes called a “random walk with drift”; If a time series has a **unit root**, it shows a systematic pattern that is unpredictable. A possible **unit root**.

Dickey-Fuller checks for unit root.

The **chi**-squared test is **used** to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. ... A **chi**-squared test can be **used** to attempt rejection of the null hypothesis that the data are independent.

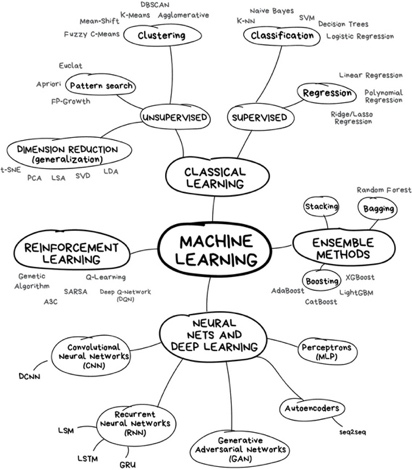
Expected vs. Observed

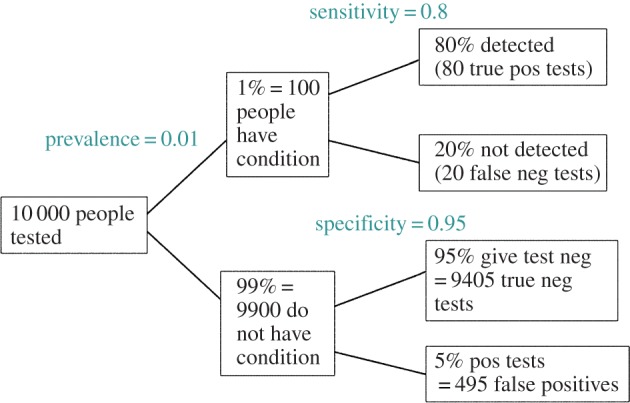


The **t distribution** (aka, **Student’s t-distribution**) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

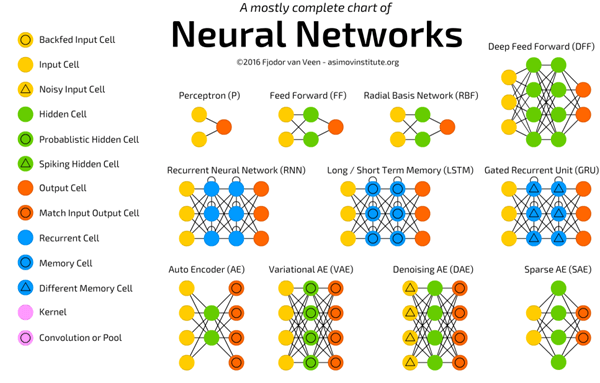
In [statistics](https://en.wikipedia.org/wiki/Statistics), the number of **degrees of freedom** is the number of values in the final calculation of a [statistic](https://en.wikipedia.org/wiki/Statistic) that are free to vary.[[1]](https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)#cite_note-1)

The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called number of degrees of freedom. In other words, the number of degrees of freedom can be defined as the minimum number of independent coordinates that can specify the position of the system completely.

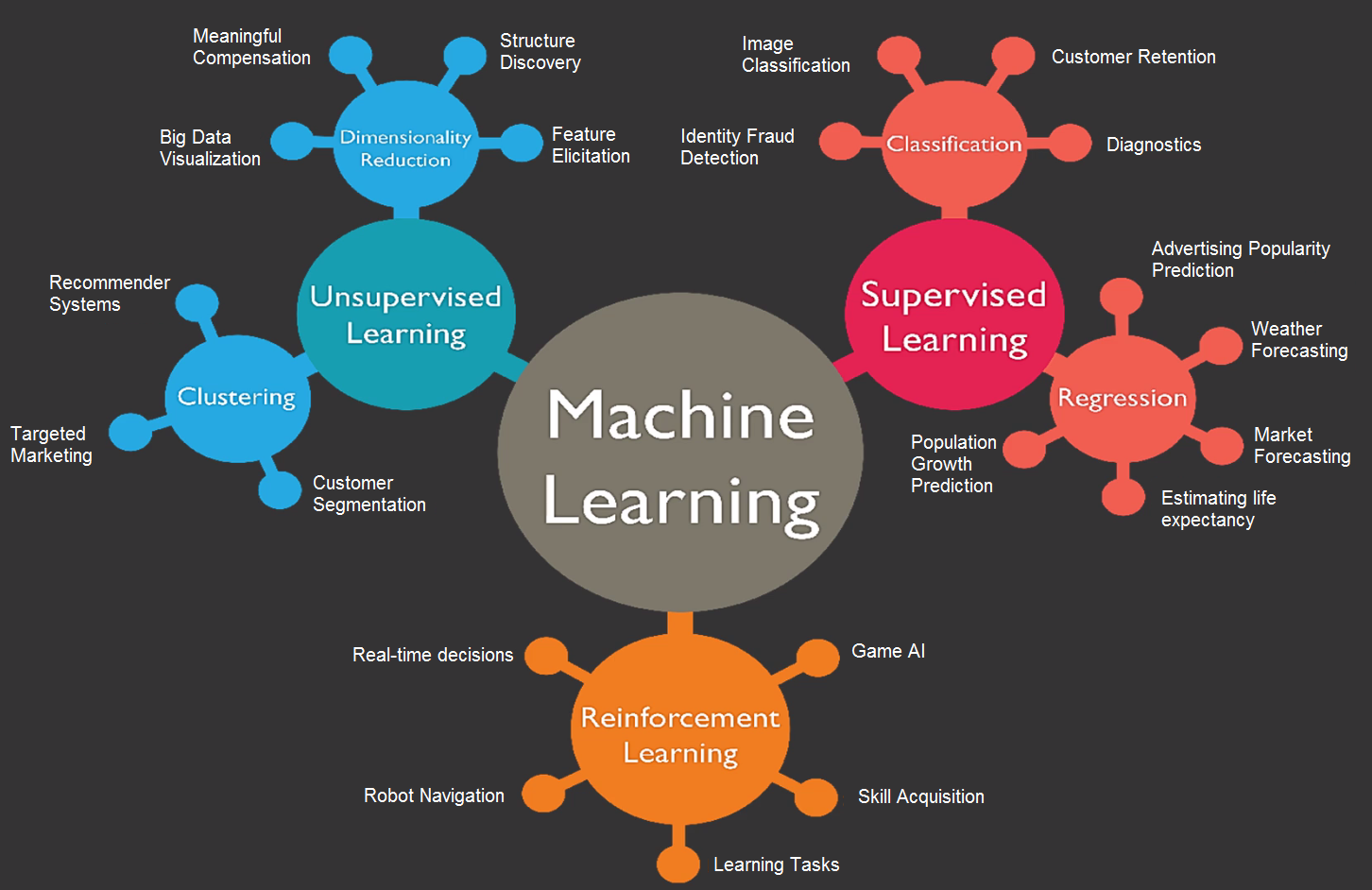




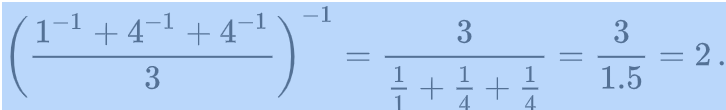






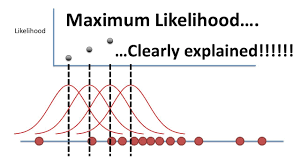


The harmonic mean can be expressed as the [reciprocal](https://en.wikipedia.org/wiki/Multiplicative_inverse) of the [arithmetic mean](https://en.wikipedia.org/wiki/Arithmetic_mean) of the reciprocals of the given set of observations. As a simple example, the harmonic mean of 1, 4, and 4 is



A picture containing text, map

Description automatically generatedkr

 Stat quest

Maximum likelihood estimation is a method that will find the values of μ and σ that result in the curve that best fits the data.

The common measure of in-sample fit is the information criterion such as Akaike (AIC) or Bayesian (BIC). These are computed using the loglikelihoods.

BIC is in between -

R -

AIC – in between R squared and BIC

BIC – penalizes for the most when you have too many parameters. Most parsimonious.

Information Criterion -

**Extra: +++++++++++++++++++++++++++++++++++++++++++**

***Structure extraction*** *– identifying fields and blocks of content based on tagging*

***Identify and mark sentence, phrase, and paragraph boundaries –*** *these markers are important when doing entity extraction and NLP since they serve as useful breaks within which analysis occurs.*

***Language identification*** *– will detect the human language for the entire document and for each paragraph or sentence. Language detectors are critical to determine what linguistic algorithms and dictionaries to apply to the text.*

***Tokenization*** *– to divide up character streams into tokens which can be used for further processing and understanding. Tokens can be words, numbers, identifiers or punctuation (depending on the use case).*

*The first thing you do in NLP is tokenization.*

***Tagging and Acronym normalization*** *– acronyms can be specified as “I.B.M.” or “IBM” so these should be tagged and normalized.*

***Lemmatization*** *– reduces word variations to simpler forms that may help increase the coverage of NLP utilities.*

* *For example, from "produced", the lemma is "produce", but the stem is "produc-".*
* *Lemmatization - Lemmatization is a process of reducing words into their lemma or dictionary. It takes into account the meaning of the word in the sentence.*
* *Example: beautiful and beautifully are lemmatized to beautiful and beautifully respectively without changing the meaning of the words. But, good, better and best are lemmatized to good since all the words have similar meaning.*
* *A lemma is a word that stands at the head of a definition in a dictionary. Lemma is preferred over stemming.*

***Stemming –*** *Stemming uses simple pattern matching to simply strip suffixes of tokens (e.g. remove “s”, remove “ing”, etc.).*

***Decompounding*** *– for some languages (typically Germanic, Scandinavian, and Cyrillic languages), compound words will need to be split into smaller parts to allow for accurate NLP. For example: “samstagmorgen” is “Saturday Morning” in German.*

***Entity extraction*** *– identifying and extracting entities (people, places, companies, etc.) is a necessary step to simplify downstream processing. There are several different methods:*

* **Regex extraction** *– good for phone numbers, ID numbers (e.g. SSN, driver’s licenses, etc.), e-mail addresses, numbers, URLs, hashtags, credit card numbers, and similar entities.*
* **Dictionary extraction** *– uses a dictionary of token sequences and identifies when those sequences occur in the text. This is good for known entities, such as colors, units, sizes, employees, business groups, drug names, products, brands, and so on.*
* **Complex pattern-based extraction** *– good for people names (made of known components), business names (made of known components) and context-based extraction scenarios (e.g. extract an item based on its context) which are fairly regular in nature and when high precision is preferred over high recall.*
* **Statistical extraction** *– use statistical analysis to do context extraction. This is good for people names, company names, geographic entities which are not previously known and inside of well-structured text (e.g. academic or journalistic text). Statistical extraction tends to be used when high recall is preferred over high precision.*
* ***Phrase extraction*** *– extracts sequences of tokens (phrases) that have a strong meaning which is independent of the words when treated separately. These sequences should be treated as a single unit when doing NLP.*

*For example, “Big Data” has a strong meaning which is independent of the words “big” and “data” when used separately. All companies have these sorts of phrases which are in common usage throughout the organization and are better treated as a unit rather than separately. Techniques to extract phrases include:*

Part of speech (POS) tagging *– identifies phrases from noun or verb clauses*

Statistical phrase extraction *- identifies token sequences which occur more frequently than expected by chance*

Hybrid *- uses both techniques together and tends to be the most accurate method.*

# **EXTRA NOTES: SUMMARY**

# **Regression Metrics**

Mean Squared Error (MSE)

It is perhaps the most simple and common metric for regression evaluation, but also probably the least useful. It is defined by the equation

Root Mean Squared Error (RMSE)

RMSE is just the square root of MSE.

The square root is introduced to make scale of the errors to be the same as the scale of targets.

Mean Absolute Error (MAE)

What is important about this metric is that it **penalizes huge errors that not as that badly as MSE does. Because it is not squared.** Thus, it’s not that sensitive to outliers as mean square error. It is not as sensitive to outliers as MSE.

MAE is widely used in finance, where $10 error is usually exactly two times worse than $5 error. On the other hand, MSE metric thinks that $10 error is four times worse than $5 error. MAE is easier to justify than RMSE.

R Squared (R²)

The coefficient of determination.

A close up of a logo

Description automatically generated