Funded by
the European Union

8. Statistical Analysis

This series of tutorials is based upon work from COST Action **Multi3Generation CA18231**, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation, cf. www.cost.eu

Importance/Relevance of a concept/theme in a corpus

- By applying regular grammars to a corpus, NooJ users can locate words, terms, lexical fields, themes, concepts and expressions, and build the corresponding concordances.

- How to determine if the resulting concordance is interesting or not?

- Does a theme appear in an *interesting* way, in a chapter of a novel (*e.g.*, love), in a political debate (immigration), in a newspaper (Middle East), in accident insurance reports (left front axle), or in product reviews (antennagate)?

Importance/Relevance of a concept/theme

- Studying frequency: if the text contains many occurrences of a term, then the term must be relevant and interesting to the text...

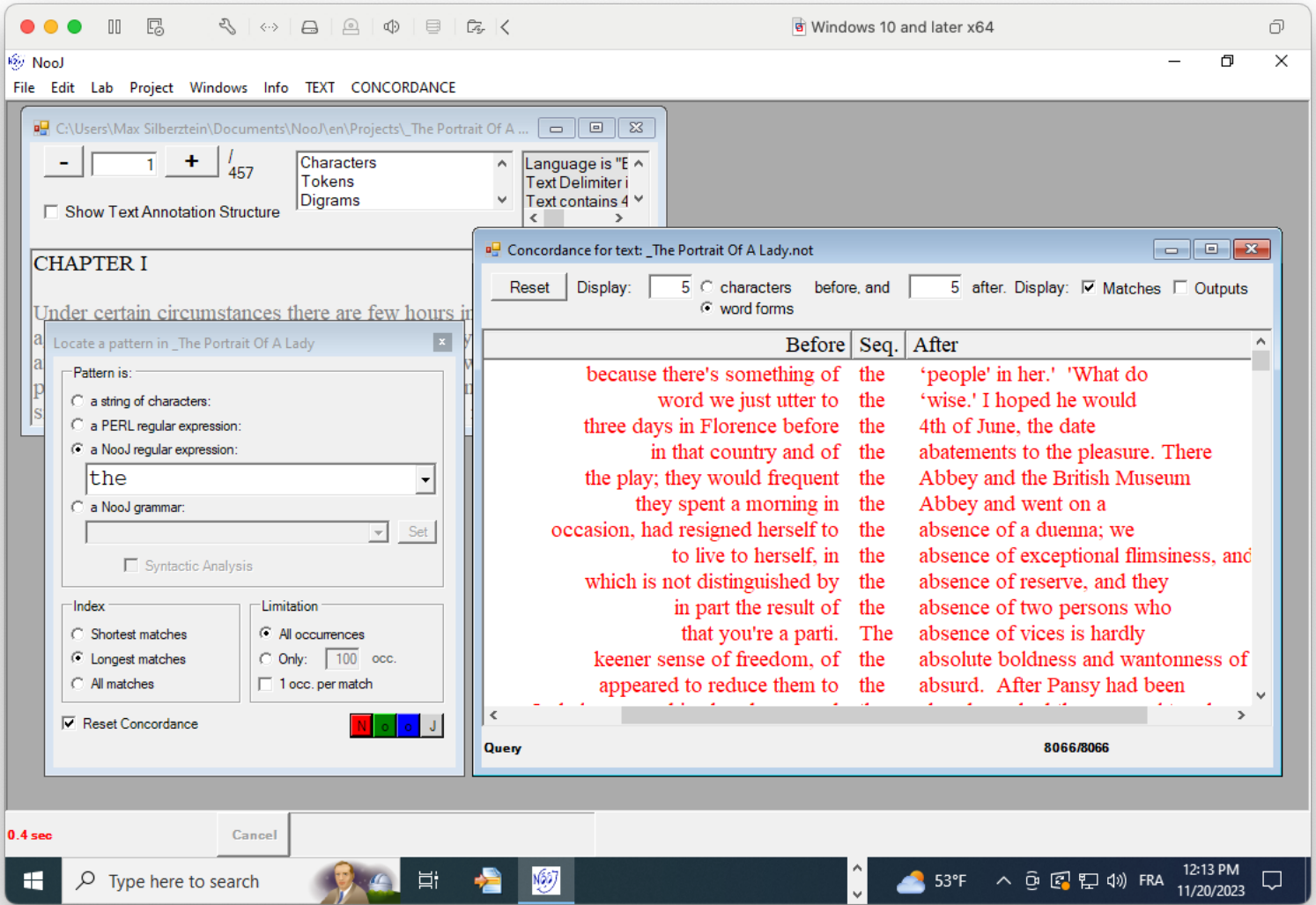
Importance/Relevance of a concept/theme

- Studying frequency: if the text contains many occurrences of a term, then the term must be relevant and interesting to the text...

- But, the word “the” occurs very frequently in any English text; it does not mean that it is specifically relevant to a given novel

Importance/Relevance of a concept/theme

- The word “the” has 8,066 occurrences in The Portrait Of A Lady. But that is not very interesting...



Importance/Relevance of a concept/theme

What is important is not the frequency by itself, but the difference between what we expect and what we find:

- We expect any English text to contain many occurrences of the word “the”; we find out that there is indeed a large number of occurrences ⇨ no surprise, not interesting.

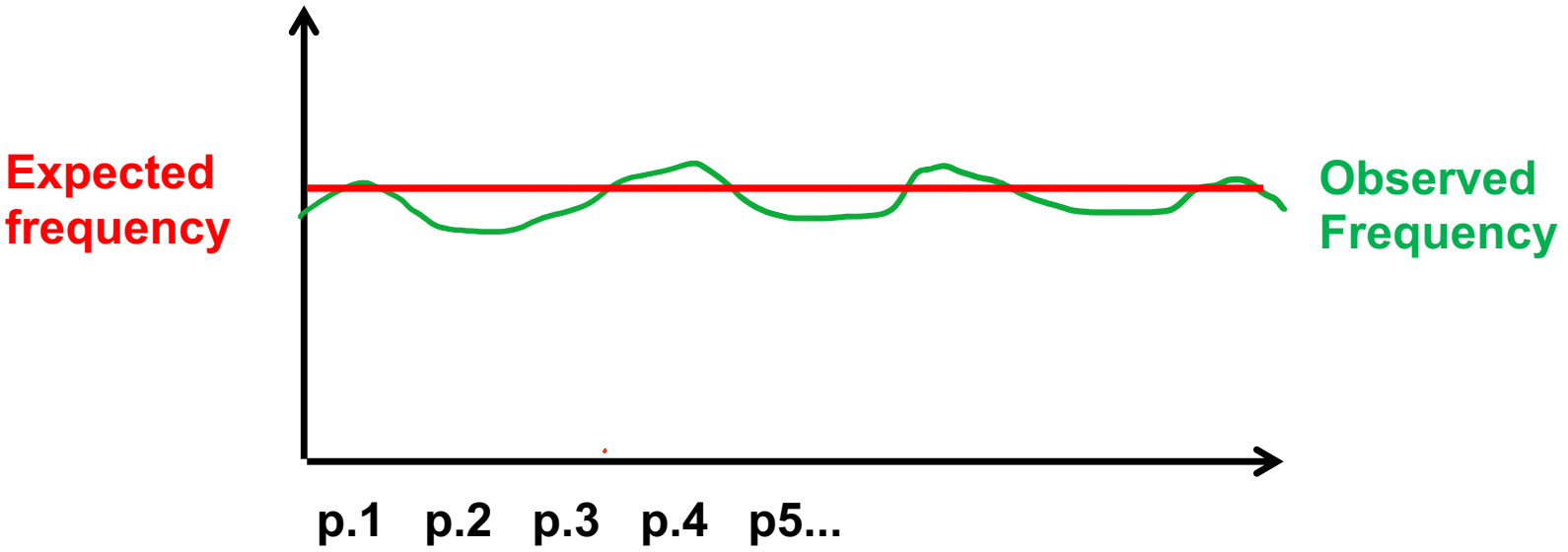
- We expect to find absolutely zero occurrence of the term COVID in newspapers from the XXth century. But if we were to find even only one occurrence
  ⇨ that would be extremely interesting!

Importance/Relevance of a concept/theme

- We are going to look for differences between the **expected** and **observed** frequencies.

Importance/Relevance of a concept/theme

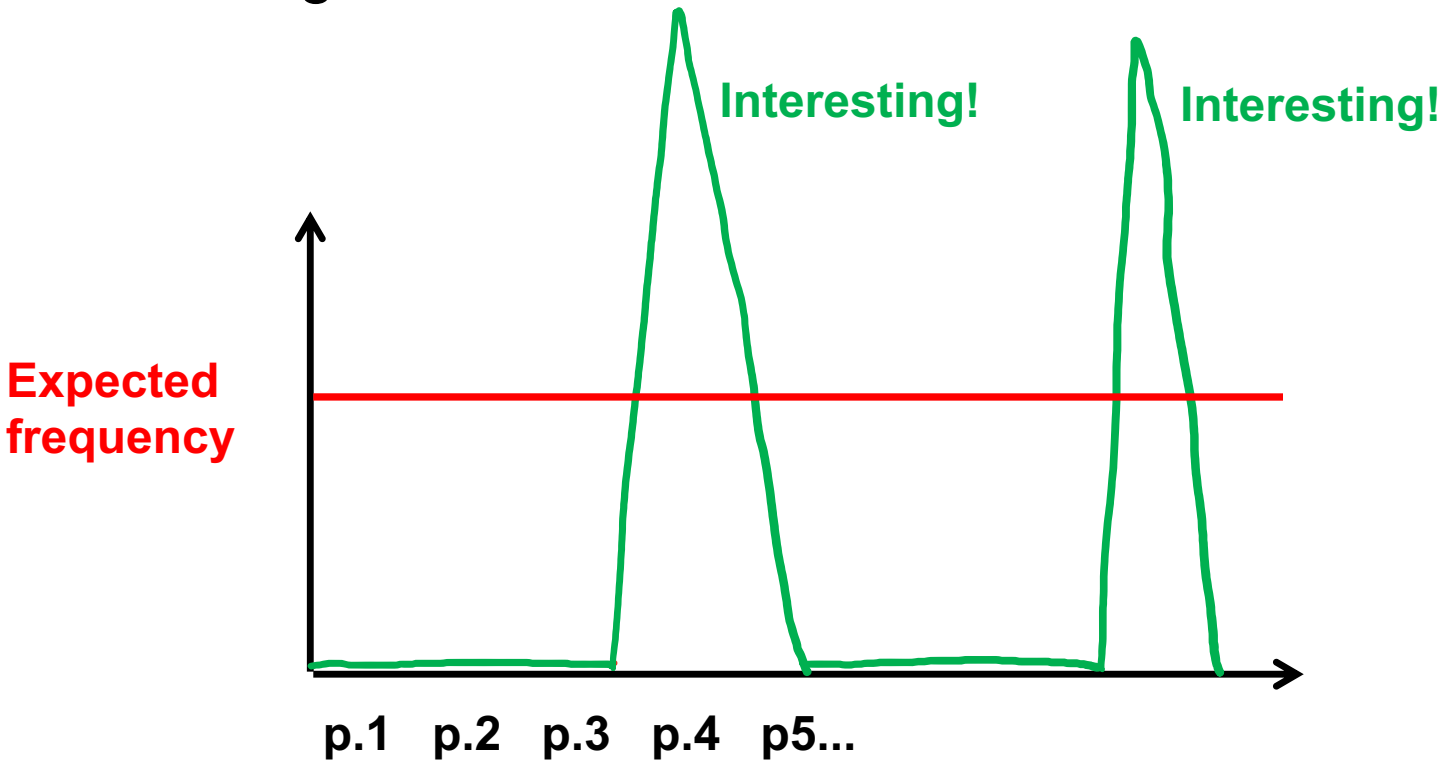
- Let's say we have a 100-page text.
- We look for the word "the".
- The resulting concordance contains 1,000 occurrences.



⇨ We expect each page of the text to contain aproximately 10 occurrences of the word "the"

Importance/Relevance of a concept/theme

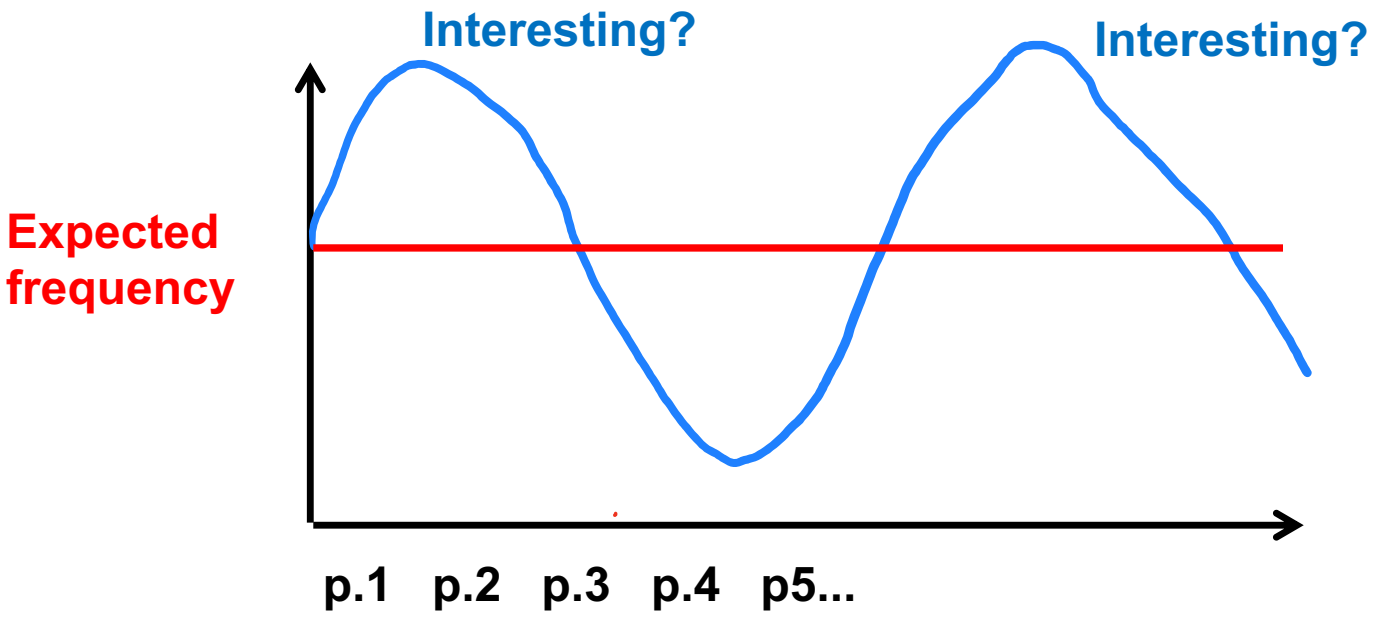
- Now we look for the word “killed”
- The resulting concordance also contains 1,000 occurrences.



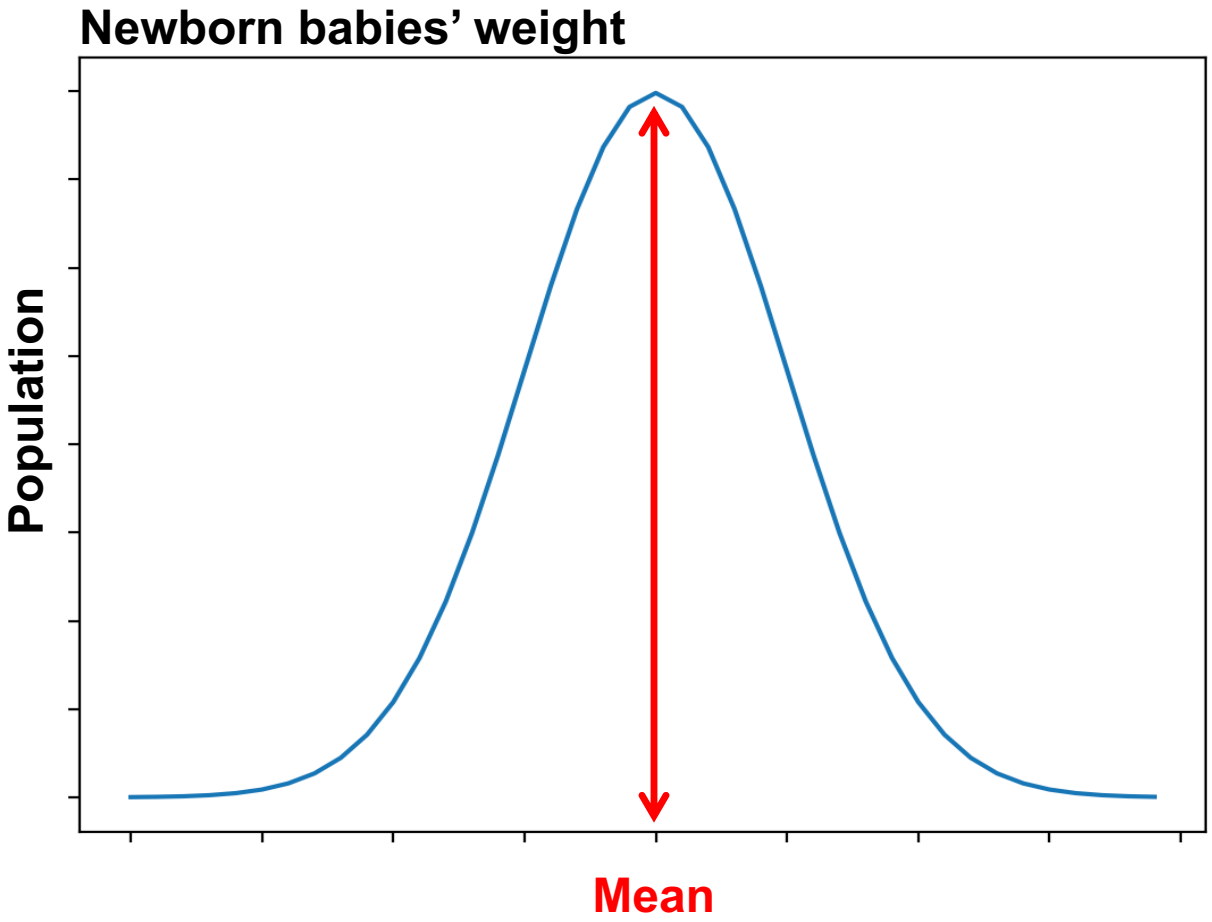
⇨ All occurrences of the word “killed” occur only in two small sections of the text

Importance/Relevance of a concept/theme

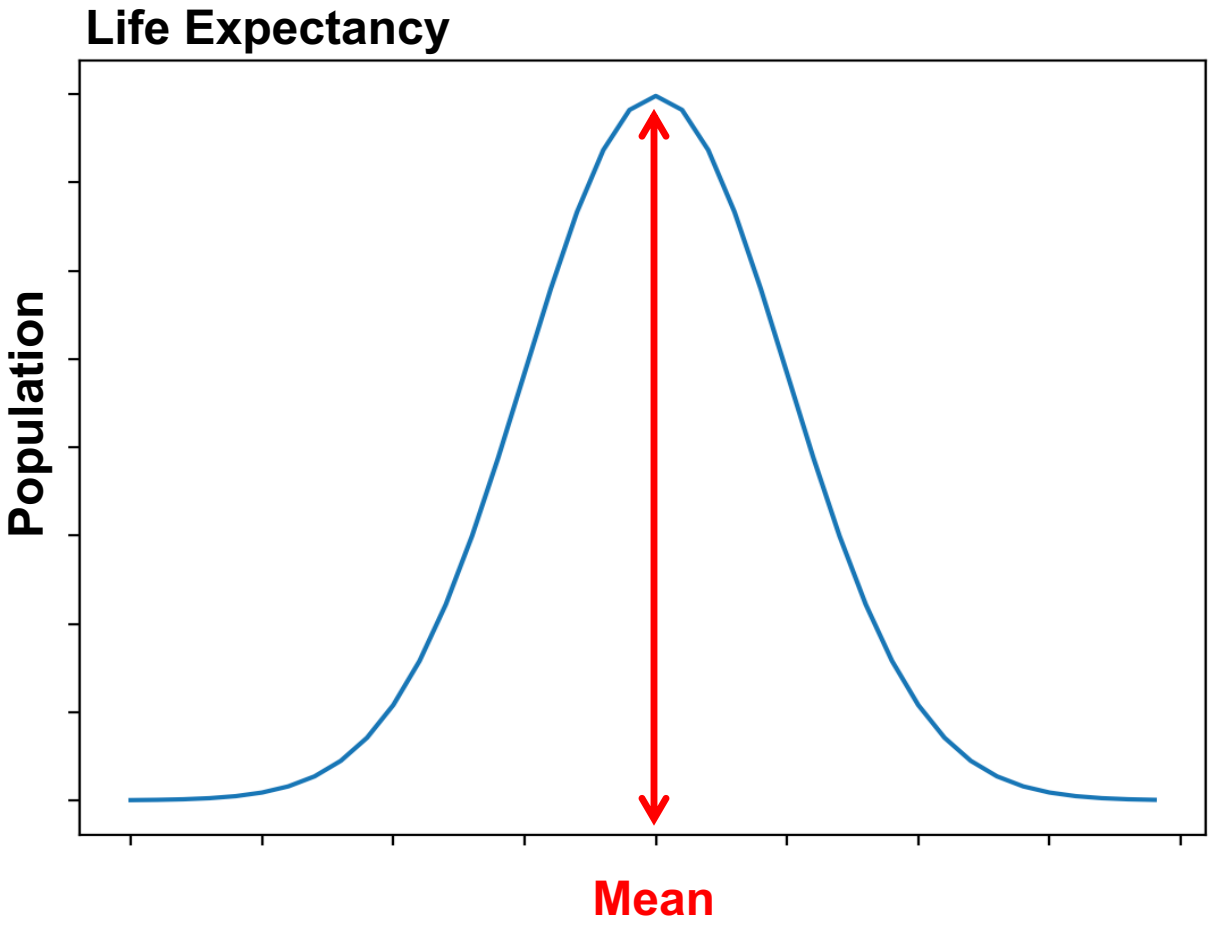
- Between these two extreme cases, how to decide if a



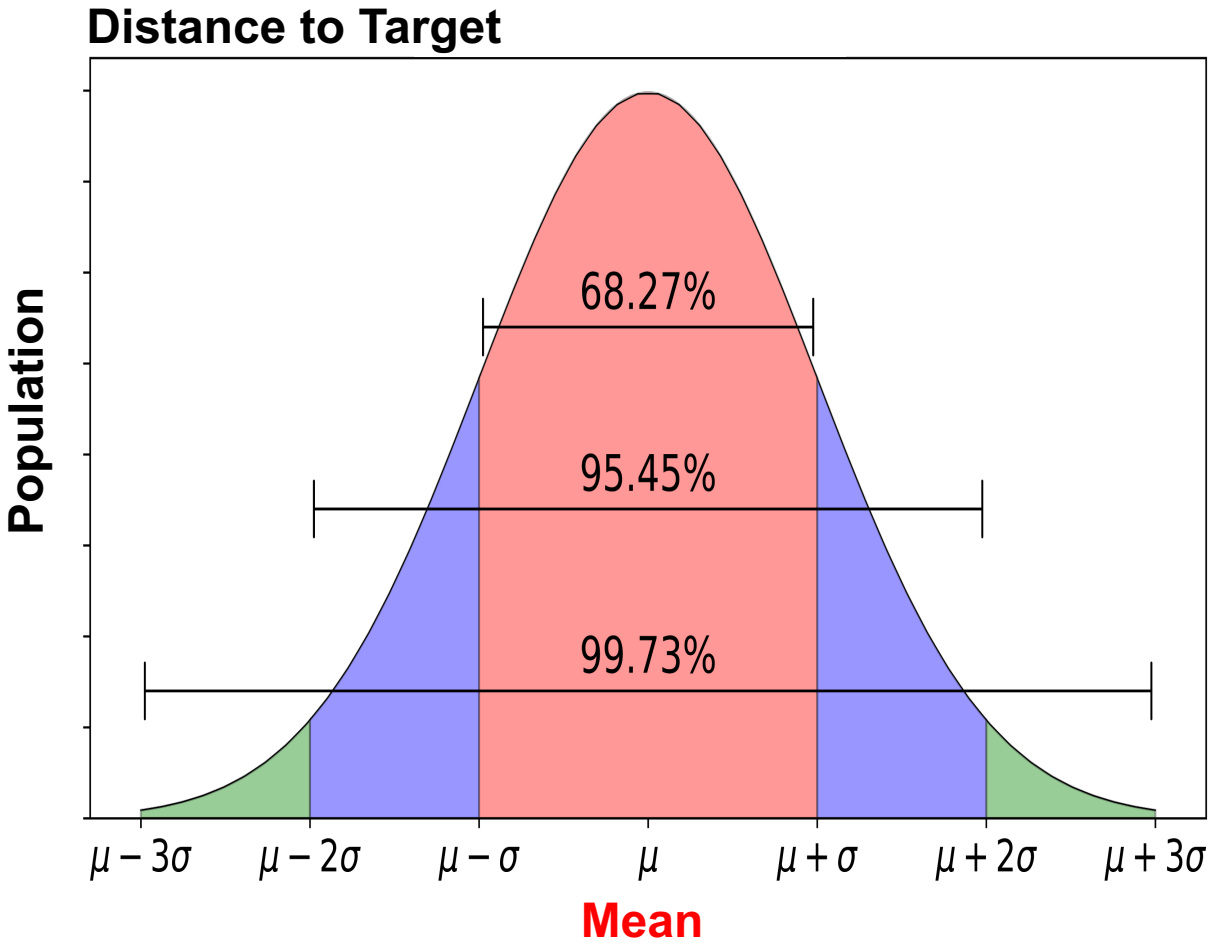
Gauss Normal Distribution



Gauss Normal Distribution

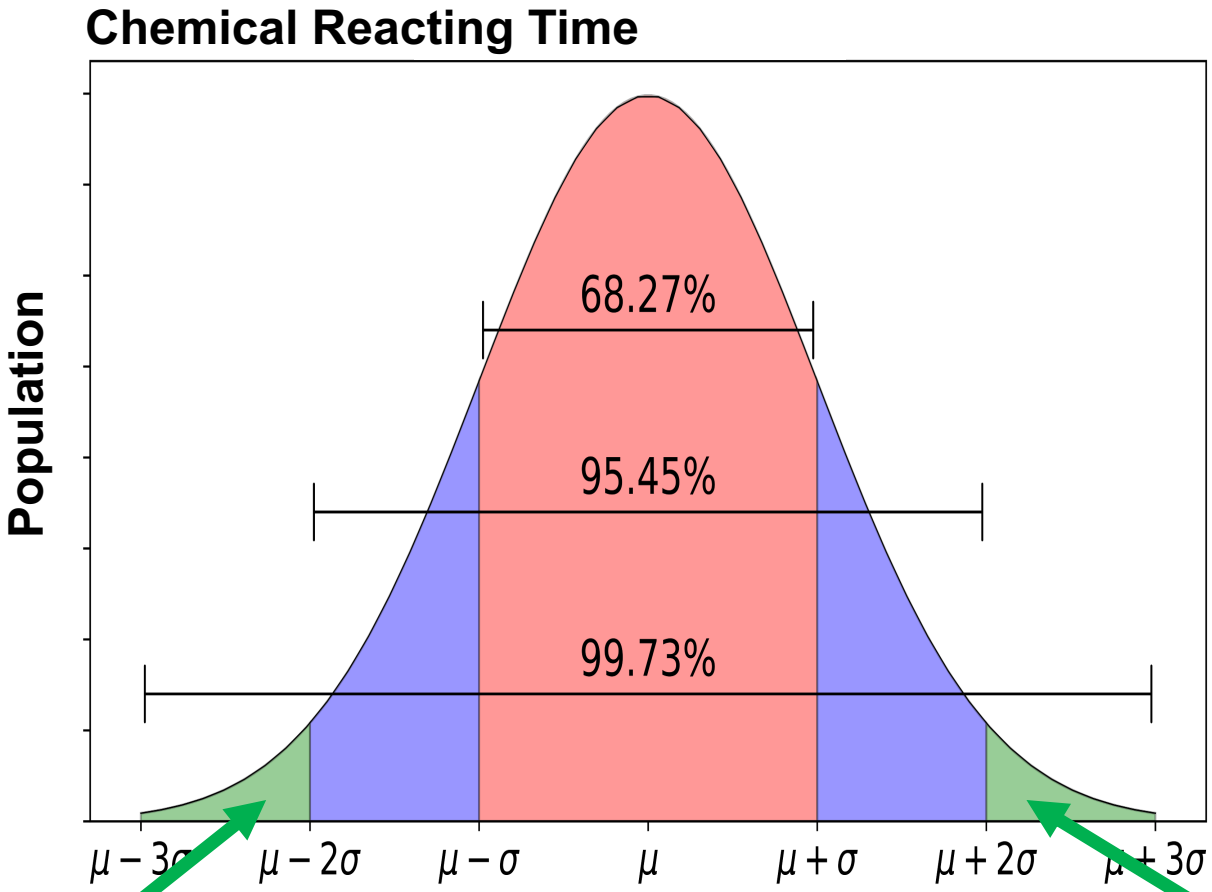


Gauss Normal Distribution



μ: Mean
σ: Standard Deviation

Gauss Normal Distribution



< μ - 2 σ
EXCEPTIONAL

> μ + 2 σ
EXCEPTIONAL

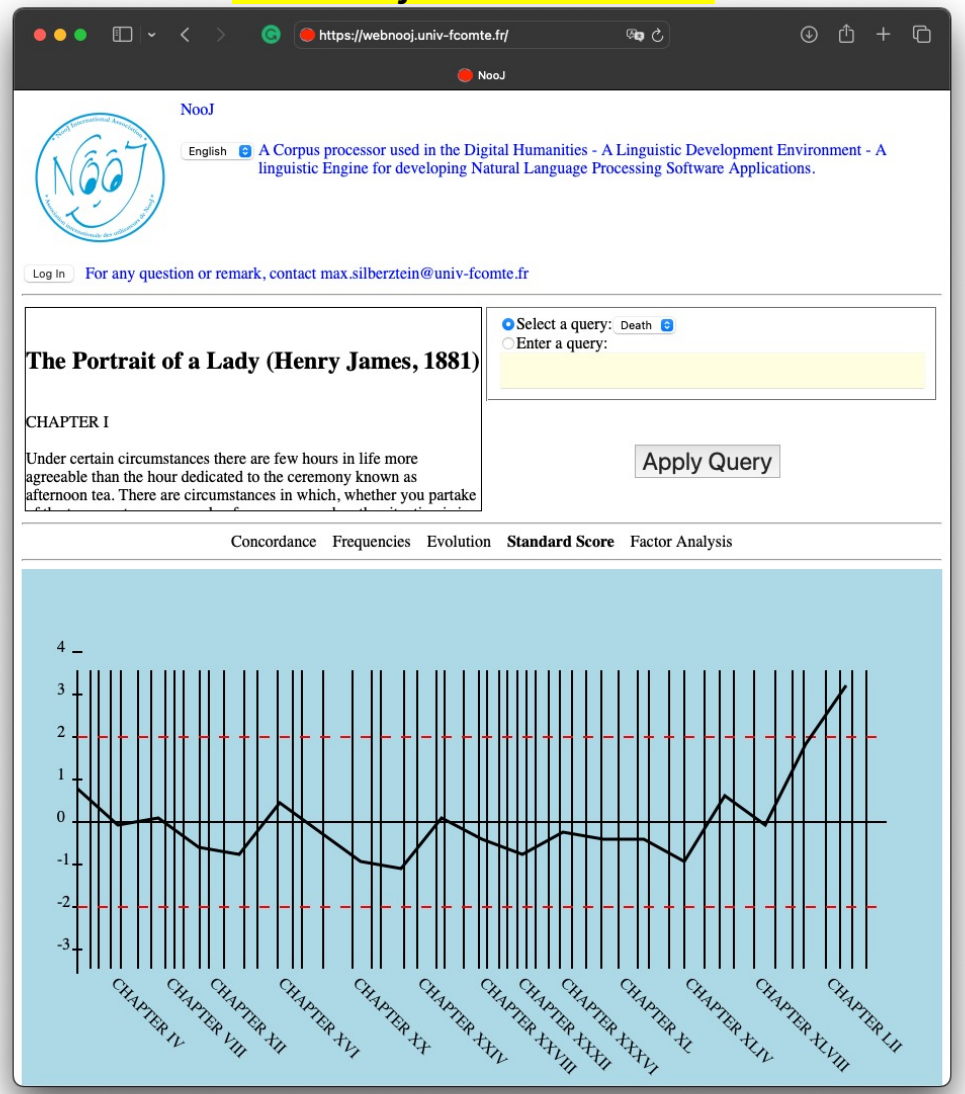
μ: Mean
σ: Standard Deviation

Standard Score

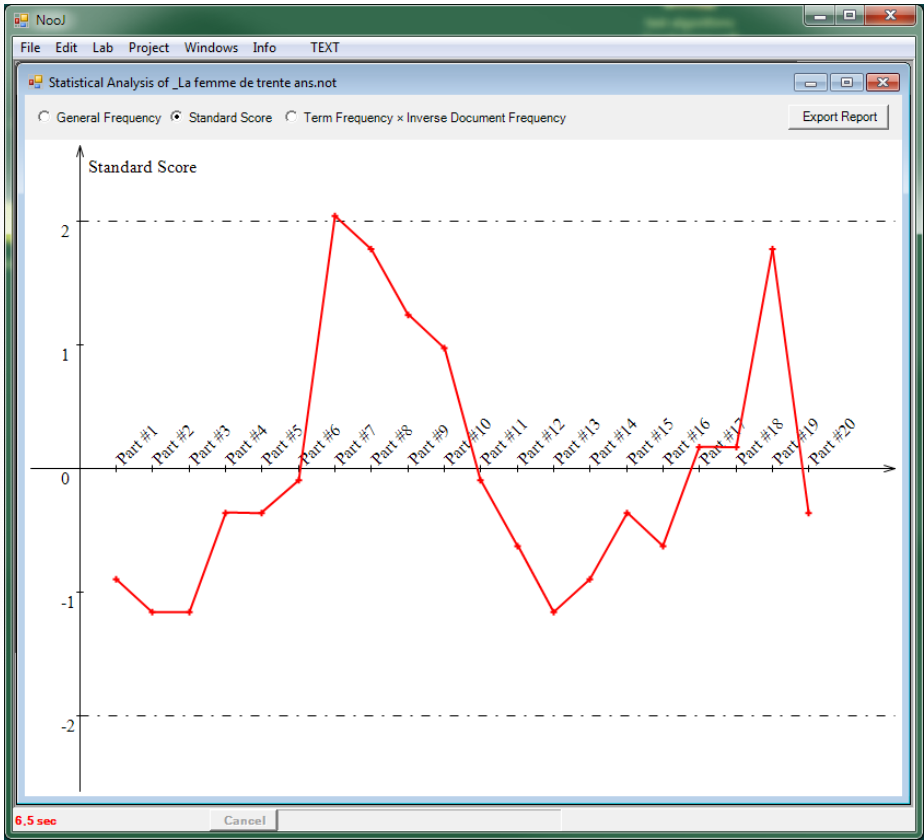
- Standard Scores evaluates the distance between a given value and the mean in a population
- If > 2, then the measure is abnormaly high ⇨ very interesting
- If < -2, then the measure is abnormaly low ⇨ very interesting
- Between -2 and +2, the value is located within the range of normal values, *i.e.*, not surprising, not interesting.

CONCORDANCE > Standard Score

webnooj.univ-fcomte.fr



NooJ





CONGRATULATIONS


You know how to perform a standard score analysis of any linguistic unit that can be recognized by a NooJ grammar