



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, Nool 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...

The screenshot shows the NooJ software interface. The main window displays the text of 'CHAPTER I' from 'The Portrait Of A Lady'. A search window is open, showing the search pattern 'the' and the results of the search. The search results are displayed in a table with columns 'Before', 'Seq.', and 'After'. The search results show the word 'the' in red text, indicating it is the search term. The search results are as follows:

Before	Seq.	After
because there's something of word we just utter to three days in Florence before in that country and of the play; they would frequent they spent a morning in occasion, had resigned herself to to live to herself, in which is not distinguished by in part the result of that you're a parti. keener sense of freedom, of appeared to reduce them to	the	'people' in her.' 'What do 'wise.' I hoped he would 4th of June, the date abatements to the pleasure. There Abbey and the British Museum Abbey and went on a absence of a duenna; we absence of exceptional flimsiness, and absence of reserve, and they absence of two persons who absence of vices is hardly absolute boldness and wantonness of absurd. After Pansy had been

The search window also shows the search pattern 'the' and the search results. The search results are displayed in a table with columns 'Before', 'Seq.', and 'After'. The search results show the word 'the' in red text, indicating it is the search term. The search results are as follows:

0.4 sec

Windows 10 and later x64

File Edit Lab Project Windows Info TEXT CONCORDANCE

C:\Users\Max Silberstein\Documents\NooJ\en\Projects_The Portrait Of A ...

Characters 457
Tokens
Digrams

Language is "E"
Text Delimiter i
Text contains 4

CHAPTER I

Under certain circumstances there are few hours in

Locate a pattern in _The Portrait Of A Lady

Pattern is:

a string of characters:

a PERL regular expression:

a NooJ regular expression:

the

a NooJ grammar:

Syntactic Analysis

Index

Shortest matches

Longest matches

All matches

Limitation

All occurrences

Only: 100 occ.

1 occ. per match

Reset Concordance

Reset Display: 5 characters before, and 5 after. Display: Matches Outputs word forms

Before Seq. After

because there's something of the 'people' in her.' 'What do

word we just utter to the 'wise.' I hoped he would

three days in Florence before the 4th of June, the date

in that country and of the abatements to the pleasure. There

the play; they would frequent the Abbey and the British Museum

they spent a morning in the Abbey and went on a

occasion, had resigned herself to the absence of a duenna; we

to live to herself, in the absence of exceptional flimsiness, and

which is not distinguished by the absence of reserve, and they

in part the result of the absence of two persons who

that you're a parti. The absence of vices is hardly

keener sense of freedom, of the absolute boldness and wantonness of

appeared to reduce them to the absurd. After Pansy had been

Query 8066/8066

Type here to search

53°F 12:13 PM 11/20/2023

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:

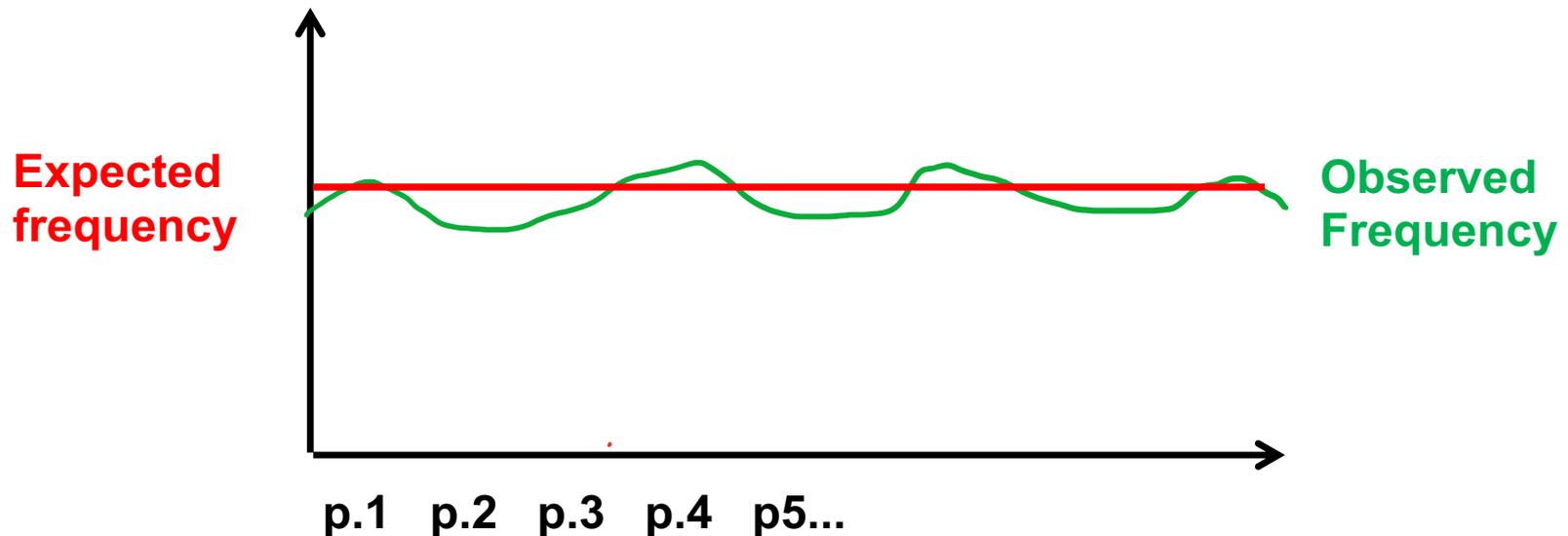
- We expect any English text to contain many occurrences of the word “the”; we find out that there is indeed a large number of occurrences ⇒ **no surprise, not interesting.**
- We expect to find absolutely zero occurrence of the term COVID in newspapers from the XXth century. But if we were to find even 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

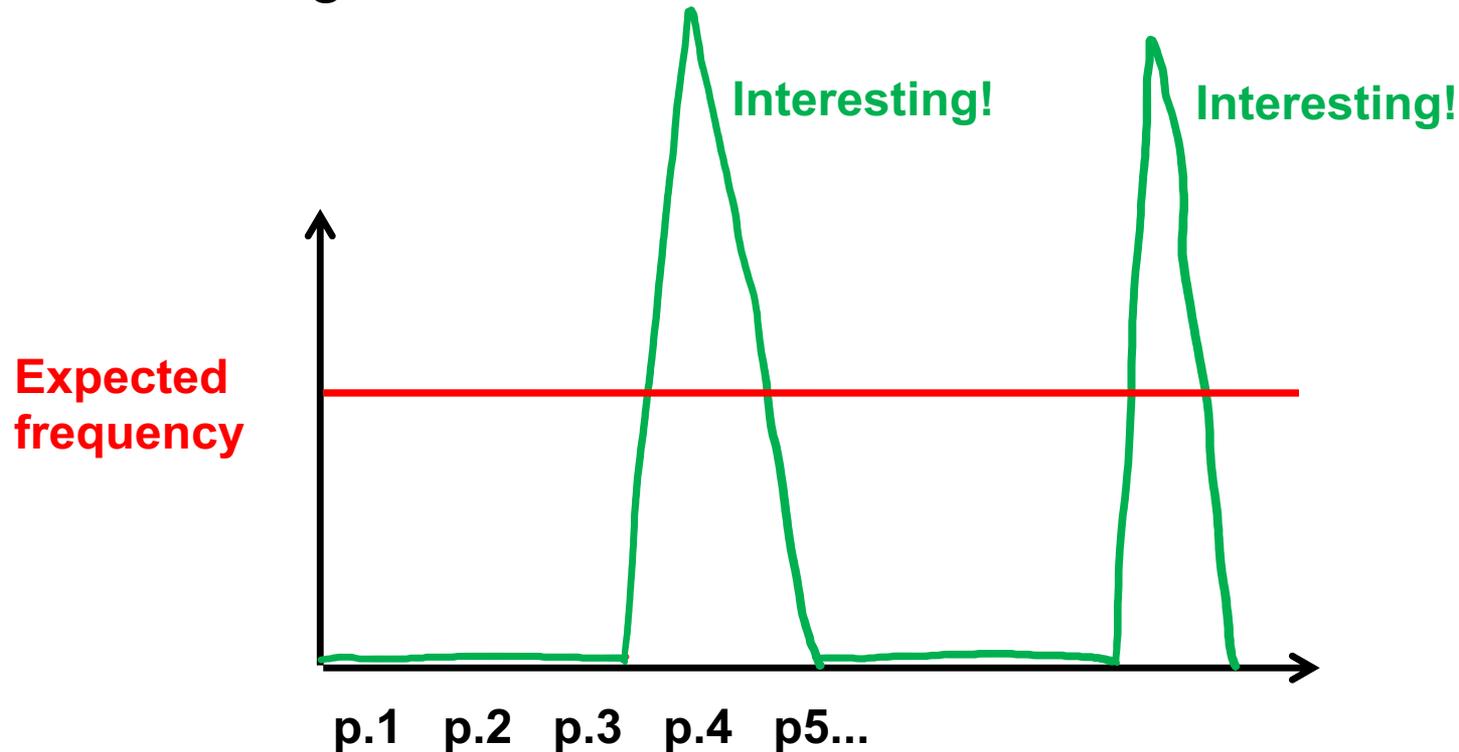
- Let's say we have a 100-page text.
- We look for the word "the".
- The resulting concordance contains 1,000 occurrences.



⇒ We expect each page of the text to contain approximately 10 occurrences of the word "the"

Importance/Relevance of a concept/theme

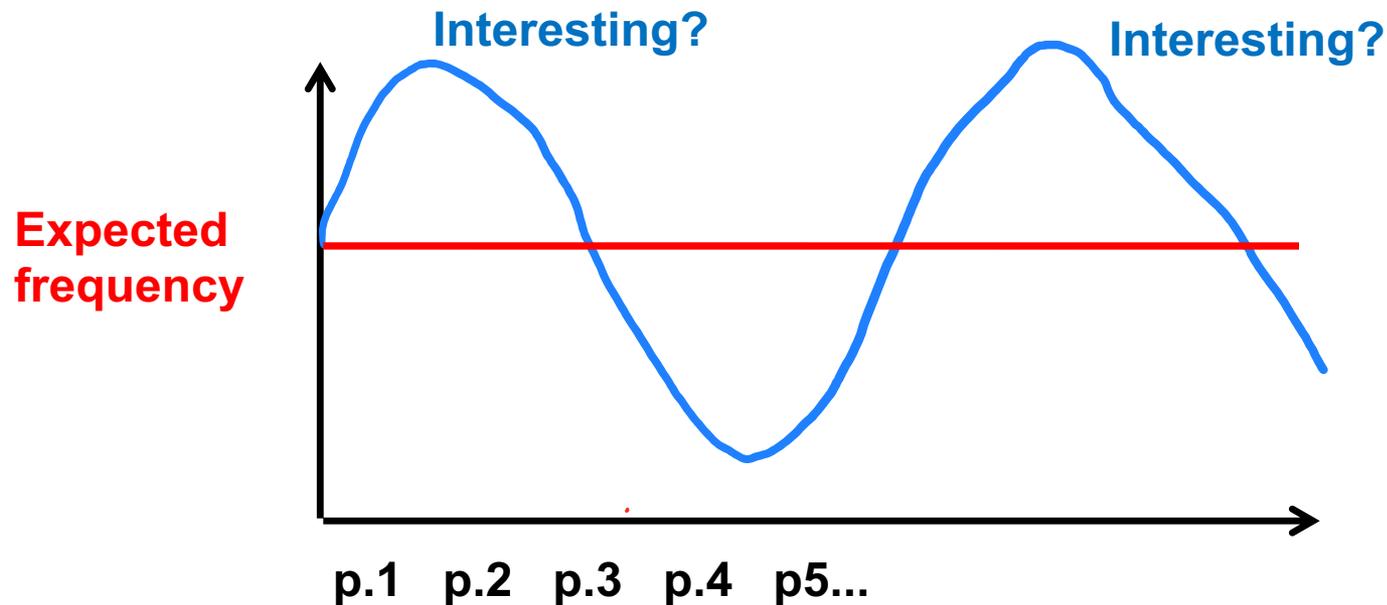
- Now we look for the word “killed”
- The resulting concordance also contains 1,000 occurrences.



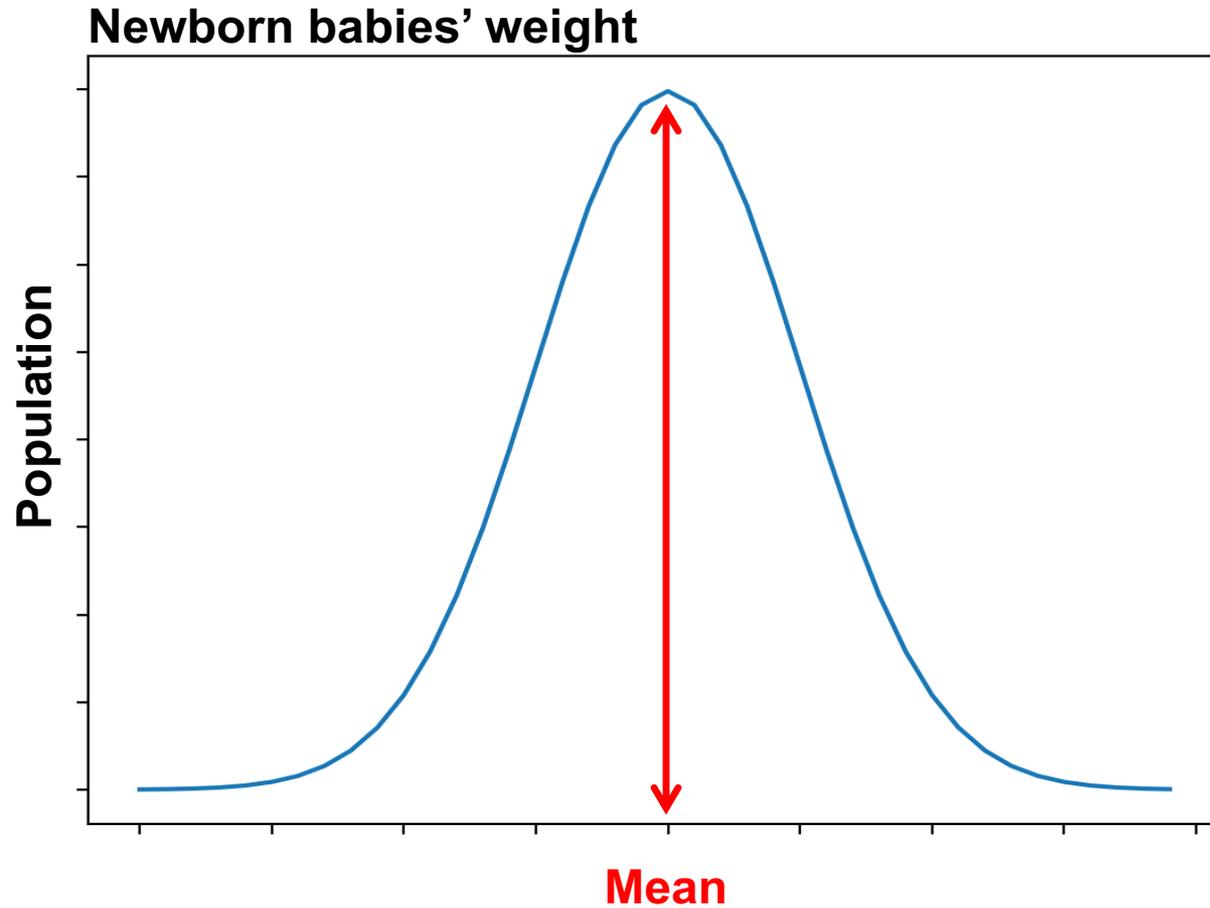
⇒ All occurrences of the word “killed” occur only in two small sections of the text

Importance/Relevance of a concept/theme

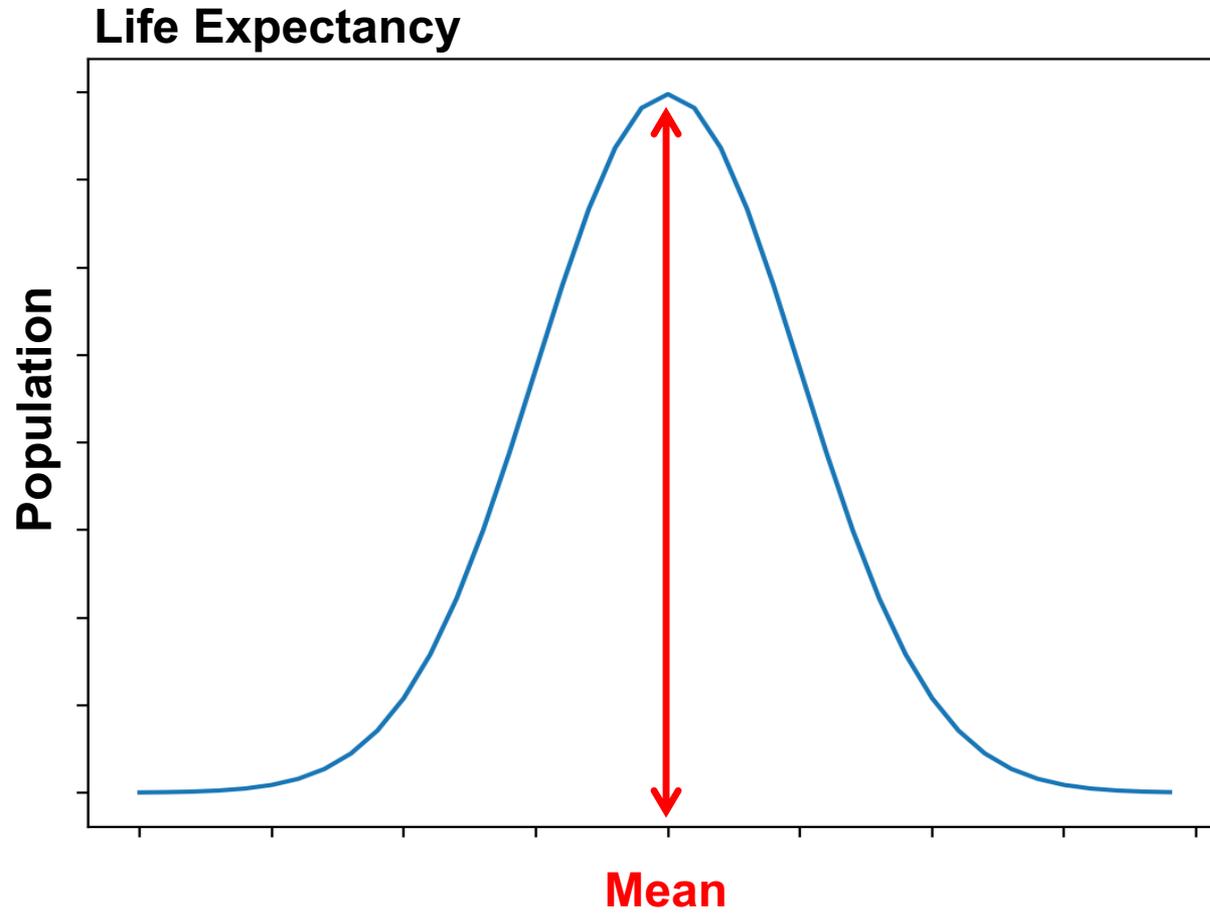
- Between these two extreme cases, how to decide if a



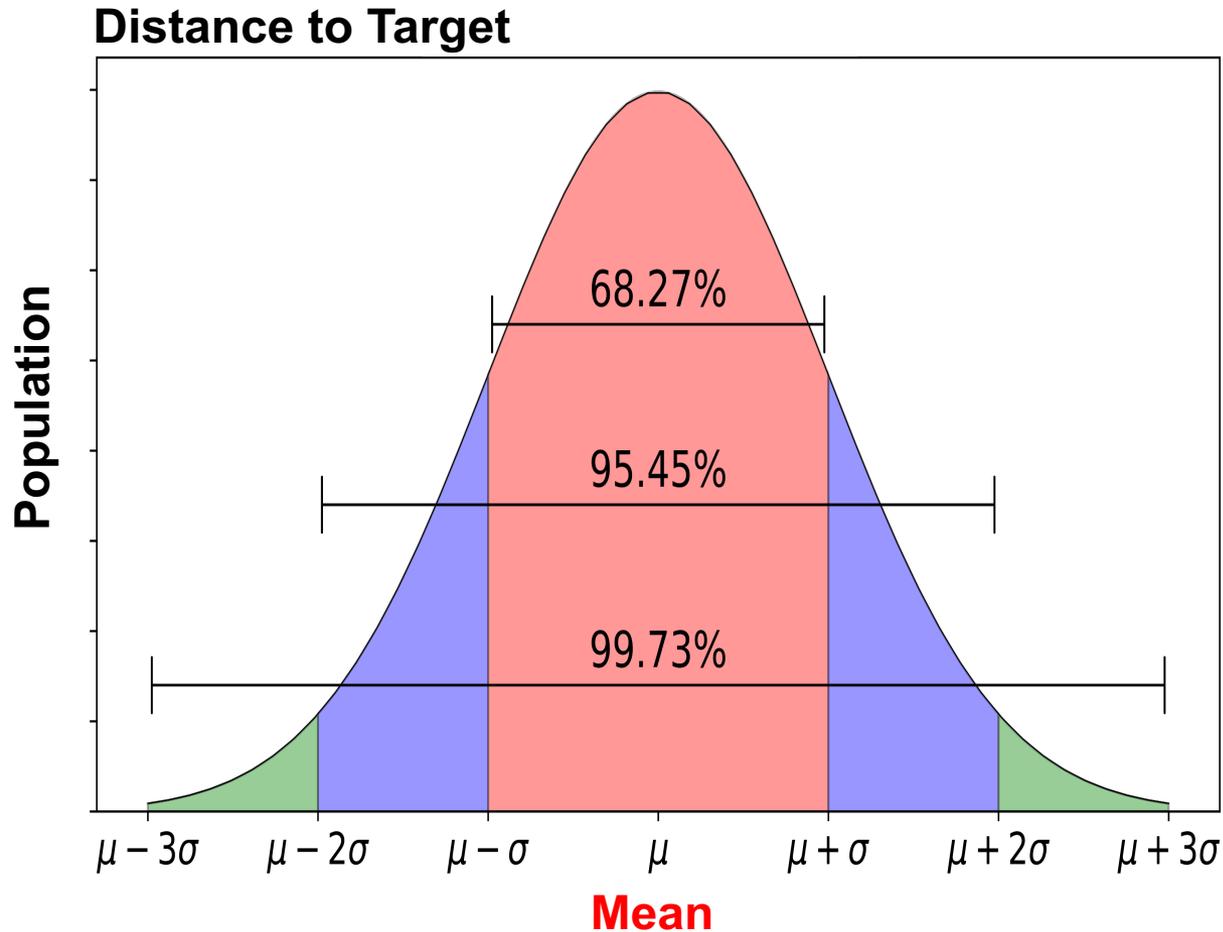
Gauss Normal Distribution



Gauss Normal Distribution



Gauss Normal Distribution

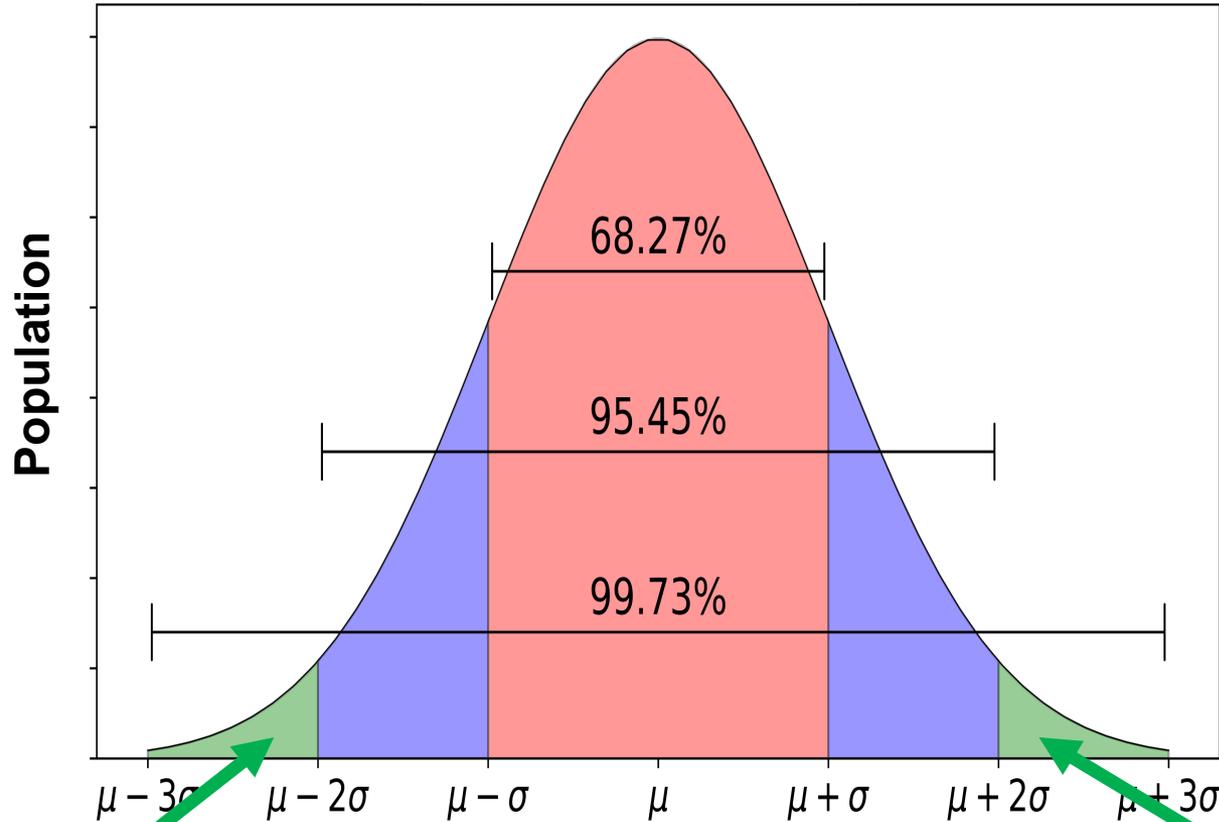


μ : Mean

σ : Standard Deviation

Gauss Normal Distribution

Chemical Reacting Time



$< \mu - 2\sigma$
EXCEPTIONAL

μ : Mean
 σ : Standard Deviation

$> \mu + 2\sigma$
EXCEPTIONAL

Standard Score

- Standard Scores evaluates the distance between a given value and the mean in a population
- If > 2 , then the measure is abnormally high \Rightarrow very interesting
- If < -2 , then the measure is abnormally low \Rightarrow 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

The Portrait of a Lady (Henry James, 1881)

CHAPTER I

Under certain circumstances there are few hours in life more agreeable than the hour dedicated to the ceremony known as afternoon tea. There are circumstances in which, whether you partake

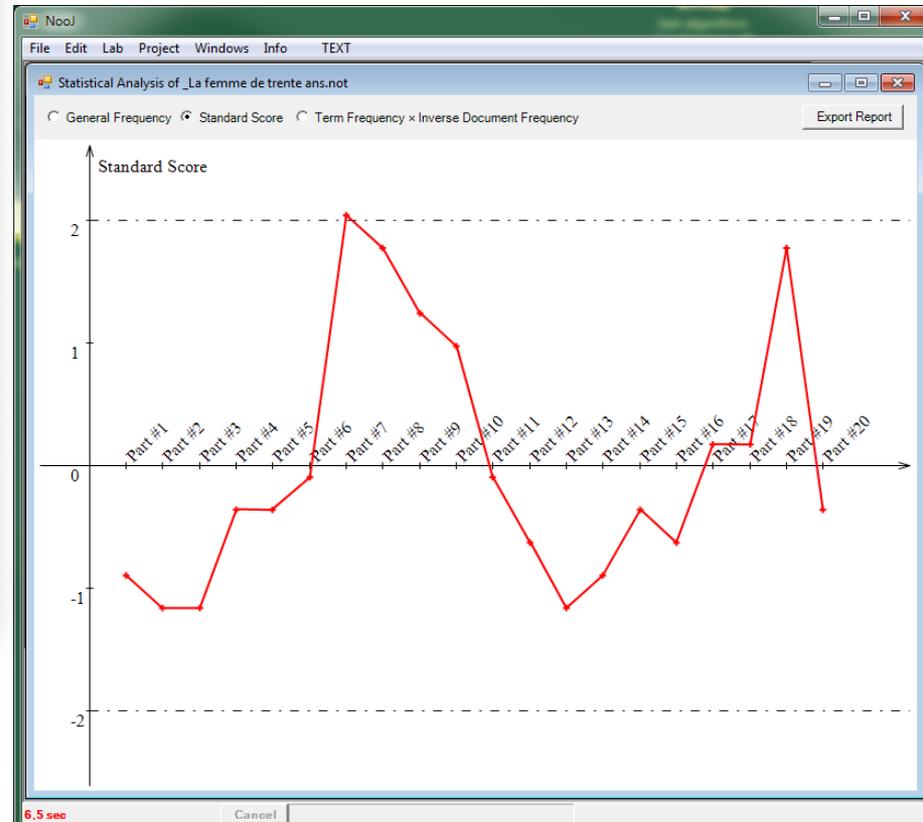
Select a query: **Death**

Enter a query:

Apply Query

Concordance Frequencies Evolution **Standard Score** Factor Analysis

NooJ





CONGRATULATIONS



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

