Minimal Ratio: An Exact Metric for Keywords, Collocations etc.

Jiří Milička[1]

Abstract

The paper defines and shows how to use the Minimal Ratio – an exact metric that expresses the ratio between the measured value and the limits of the confidence interval calculated according to the formula Fischer's exact test is based on. The metric is meant to assist with keywords and collocations extraction and comparing texts or corpora according to the word types distribution or other similar criteria.

Keywords

Collocations, confidence interval, Fisher's test, Keywords, Minimal Ratio

Introduction

Since ancient times, dividing outcomes of some system according to two more or less independent criteria was considered to be a good way to sort the outcomes and to make the system more convenient for the human mind. The idea resulted in a plenty of "quaternities", such as the Aristotelian wet / dry, hot / cold scale which affected European and Middle Eastern philosophy, science and alchemy with an admirable endurance.

In modern statistics, certain tests were introduced that help to decide, whether two binary oppositions divide the data proportionally.[2] Chi-squared and Fisher's exact test are the most popular ones. As for linguistic applications, e.g. assume the text T_1 consisting of N_1 word tokens that contains the word type *w* f_1 times and the text T_2 consisting of N_2 word tokens that contains the word w f_2 times. Fisher's test tells us what is the probability that the type *w* is in these texts distributed in this or more skewed way.

---

[1] Institute of Comparative Linguistics, Faculty of Arts, Charles University, Prague, milicka@centrum.cz
[2] (Fisher, 1922), (Yates, 1984), (Barnard, 1947) etc.

But what if this probability is only a part of our concern; what if we want more than to falsify the null hypothesis saying that the relative frequency of the type *w* in the text T_1 is the same as it is in the T_2 (at some significance level)? What if we want to know, what is the distinction between these two relative frequencies? As the problem is central for some important linguistic applications (texts comparison, automatic keyword extraction, searching for typical collocations etc.), many metrics were introduced,[3] more or less intuitive, more or less practical.

This paper defines a metric that is both exact and intuitively comprehensible. Exact, because it is based on the confidence interval calculated according to the formula the Fisher's test is based on. The main feature of the metric is that the results are easy to interpret – its name (Minimal Ratio) literally describes what it means: the minimal possible ratio between the relative frequencies (e.g. of the word type *w* in T_1 and T_2).

The Definition

Let us describe what the concept of the confidence interval means in this context. Join T_1 and T_2 into one "text" and choose randomly N_1 tokens. What is the probability that exactly *x* tokens of the type *w* were included in the sample? We have summed up 1 000 000 of such trials and the result are presented in the following chart:

[3] A basic outline to be found in (Oakes, 1998, pp 158–168).

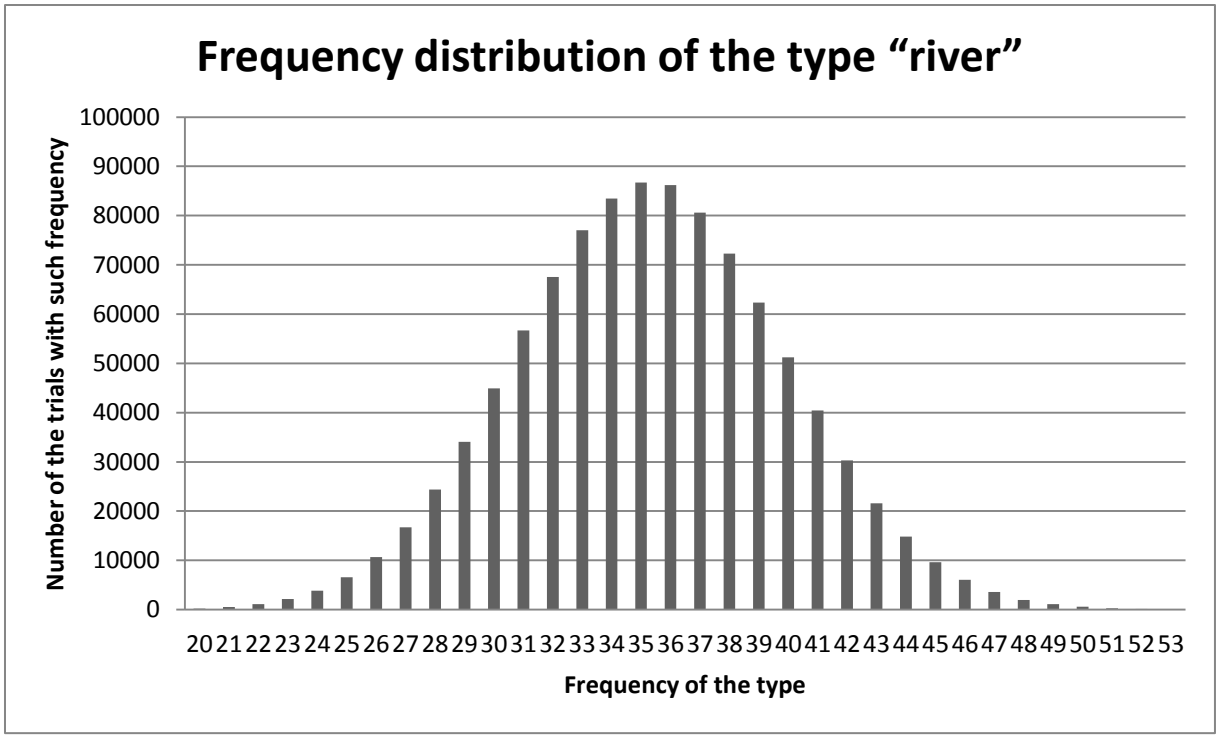

Fig. 1: T_1 is The Last of the Mohicans by James Fenimore Cooper (N_1=146 297), T_2 is The Deerslayer by the same author (N_2=213 785), *w* is the type “river” (f_1 is the variable *x*, $f_1+f_2=87$)

10 000 000 is not a small number, but random deviations still affect the results; what if we took all possible samples?

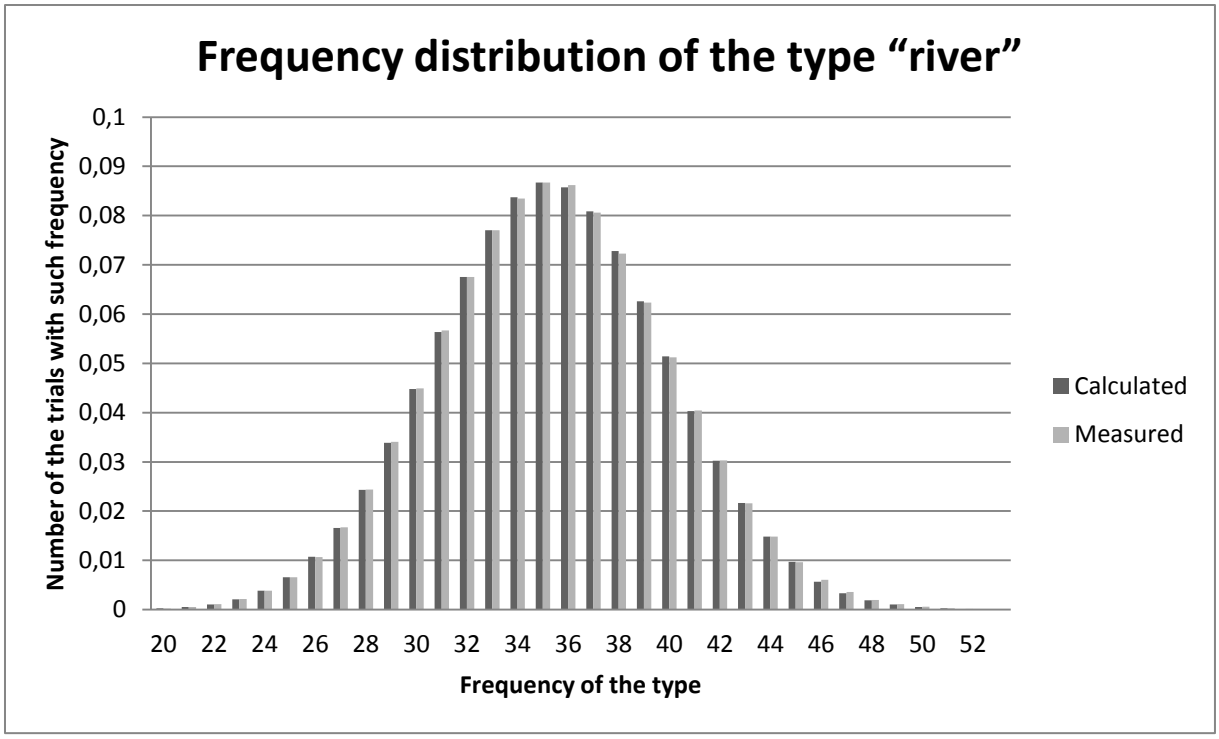

Fig. 2: T_1 is The Last of the Mohicans by James Fenimore Cooper (N_1=146 297), T_2 is The Deerslayer by the same author (N_2=213 785), *w* is the type “river” (f_1 is the variable *x*, f_1+f_2=87)

Of course, it is beyond limitations of contemporary computing technology to measure all possible samples but we have exactly calculated the result using this formula, which is also the base of Fisher's exact test:[4]

$$p = \frac{\binom{N_1}{f_1}\binom{N_2}{f_2}}{\binom{N_1+N_2}{f_1+f_2}}$$

The curve describes the distribution of frequency of the type *w* provided that the distribution is independent of any distinction between T_1 and T_2.

Using the Wolfram Mathworld definition, “a confidence interval is an interval in which a measurement or trial falls corresponding to a given probability”.[5]

A visual example of the 95% confidence interval is presented in the following chart:

[4] And which is also the base of the generalized hapax-token Combinatorial model as described in (Milička, 2009).
[5] (Weisstein, 2012)

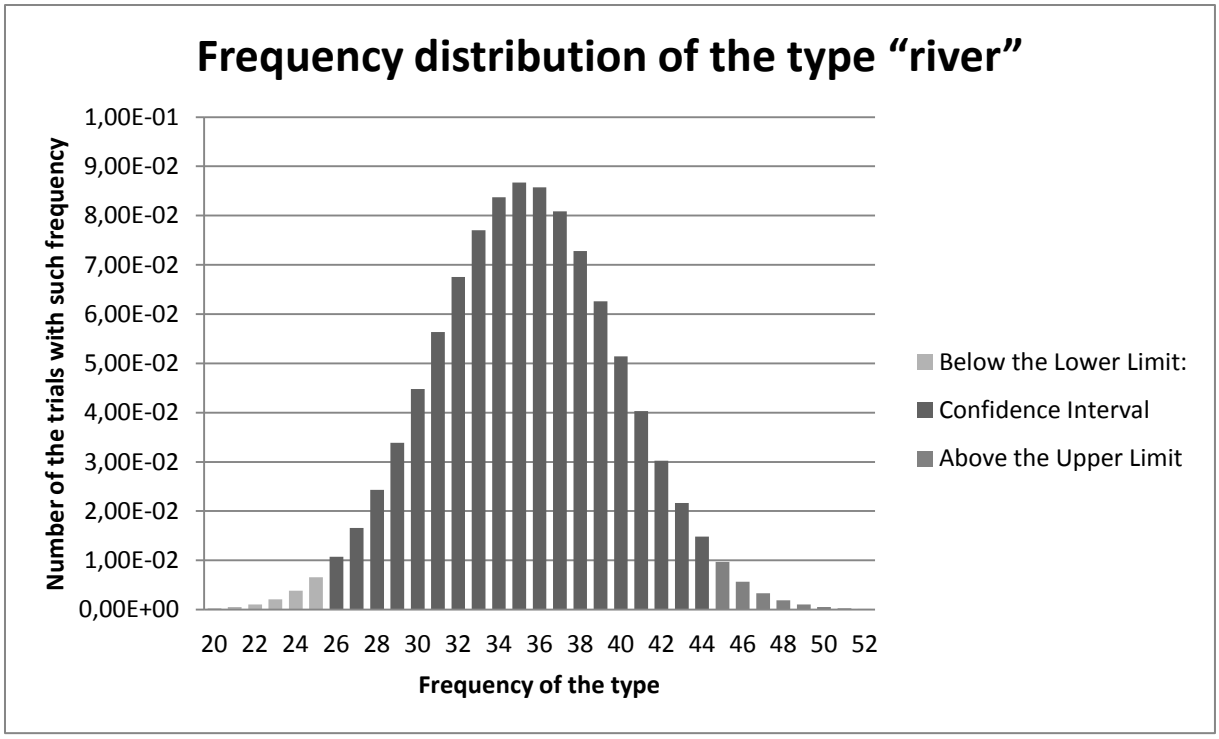

Fig. 3: T_1 is The Last of the Mohicans by James Fenimore Cooper (N_1=146 297), T_2 is The Deerslayer by the same author (N_2=213 785), *w* is the type “river” (f_1 is the variable *x*, f_1+f_2=87)

For 95 % of all samples, the frequency of *w* is somewhere between 26 and 44, these numbers are the upper and lower confidence limit (UL(0.95) and LL(0.95)). The ratio between f_1 and frequency of *w* in the sample lies somewhere between $f_1/UL(0.95)$ and $f_1/LL(0.95)$ for 95 % of all samples. If f_1 is between these two values, we cannot say that relative frequency of *w* in the text T_1 differs from the relative frequency of *w* in the text T_2.

The upper and the lower limits of the confidence interval at confidence level α are of a particular interest for our purposes.

Let us define LL(α) as the maximal z_1 that satisfies the following inequation:

$$\frac{(1-\alpha)}{2} < \sum_{i=0}^{z_1-1} \frac{\binom{N_1}{i}\binom{N_2}{f_1+f_2-i}}{\binom{N_1+N_2}{f_1+f_2}}$$

Let us define UL(α) as the minimal z_2 that satisfies the following inequation:

$$\frac{(1-\alpha)}{2} < \sum_{i=z_2+1}^{f_1+f_2} \frac{\binom{N_1}{i}\binom{N_2}{f_1+f_2-i}}{\binom{N_1+N_2}{f_1+f_2}}$$

The definition of the Minimal Ratio (MR) at confidence level α is as follows:

$$f_1 < LL(\alpha) \Rightarrow MR(\alpha) = \frac{f_1}{LL(\alpha)}$$

$$LL(\alpha) \leq f_1 \leq UL(\alpha) \Rightarrow MR(\alpha) = 1$$

$$f_1 > UL(\alpha) \Rightarrow MR(\alpha) = \frac{f_1}{UL(\alpha)}$$

Known Issues

The metric suffers from the same problems as Fisher's exact test – the distribution is a discrete one so the real confidence level could be in some cases substantially higher than the confidence level demanded by the user of the metric.

Another imperfection is the "zero division" case, i.e. $UL(\alpha) = 0$. For practical purposes the part of definition can be modified to:

$$f_1 > UL(\alpha) \Rightarrow MR(\alpha) = \frac{f_1}{UL(\alpha) + 1}$$

The solution of both of the problems could be introducing a continuous probability distribution. In fact, if N_2 is higher by several orders of magnitude than N_1, the distribution can be successfully approximated by normal distribution. As T_2 (the reference corpora) would be in many practical applications really much bigger than T_1 and as the normal distribution is much easier to implement and to calculate, this solution would be very tempting, despite of losing the exactness of the metric.

Practical Implementation

The aforementioned algorithm was implemented in the Keyworder[6]. The application uses the metric to determine word types that characterize differences between two texts or corpora chosen by the user. The author of the paper is willing to support other implementations of the metric both by consultations and sharing the source codes.

[6] Available at www.milicka.cz/kestazeni/keyworder.exe

Usage

The following table represents an example of usage of the formula and the Keyworder. The examined text is The Last of the Mohicans by James Fenimore Cooper, while the reference text is The Deerslayer by the same author. Word types are sorted by the Minimal Ratio from higher to lower, so that the most characteristic word types rank on the top.

Type	Minimal Ratio	Type	Minimal Ratio
Heyward	2,184	Renard	1,760
Duncan	2,169	cavern	1,750
scout	2,163	Sagamore	1,682
Magua	2,091	Subtil	1,667
Cora	2,066	route	1,667
Alice	2,017	rocks	1,652
David	2,017	multitude	1,625
Uncas	2,000	Mohicans	1,618
Munro	1,929	Horican	1,611
Montcalm	1,865	horses	1,600
Hawkeye	1,835	natives	1,600
Le	1,821	William	1,588

All types are either names of the main characters or notions characterizing the story (*route, cavern, scout*). More complicated results are obtained by comparing Alice's Adventures in Wonderland by Lewis Carroll with The Last of the Mohicans by James Fenimore Cooper (case insensitive):

Type	MR	Type	MR	Type	MR
she	4,242	mouse	3,583	caterpillar	3
alice	4,063	turtle	3,5	won't	3
don't	3,813	that's	3,3	oh	2,933
it's	3,8	duchess	3,25	very	2,88
queen	3,778	dormouse	3,25	there's	2,875
went	3,773	said	3,208	you're	2,875
mock	3,733	cat	3,182	tone	2,857
hatter	3,667	i'll	3,111	round	2,857
gryphon	3,667	can't	3,111	off	2,808
i'm	3,625	hare	3,1	thing	2,722
rabbit	3,615	i've	3,091	tea	2,714
herself	3,609	quite	3,056	going	2,7

All nouns here are related to the notions that play an important role in Alice's story. Also the pronoun *she* (that would be eliminated by “stoplists” of some other keyword extracting algorithms) is crucial to describe the difference between the stories. But other types are rather related to style: word types like *don't, it's, I'm, I'll* and *won't* do not say more than that L.

Carroll uses contracted forms more than J. F. Cooper does and along with types *said* and *oh* suggests that he uses more direct speech. *Quite* and *very* lead us to the assumption that L. Carroll's style tends towards subjectivity, but these ideas are not crucial for this section which is only intended to show how the metric behaves on the real data.

Conclusion

The main advantage of an exact metric like Minimal Ratio is that it gives reasonable outcomes for both small and large amounts of data. As was mentioned above, the usage of the metric is not restricted to compare two texts or corpora, the Minimal Ratio can be easily adopted to e.g. detect most important collocations (the collection of the word tokens in the left or right context should be assigned to T_1 and the whole text to $T_1 \cup T_2$).

Also *w* may be not only word type but nearly whatever distinctive feature of your interest. Of course, Minimal Ratio can be used also outside of linguistics.

References

Barnard, G. A. (1947). Significance Tests for 2×2 Tables. *Biometrika 34 (1/2)* , pp. 123–138.

Fisher, R. A. (1922). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* , pp. 87–94.

Milička, J. (2009). Type-token & Hapax-token Relation: A Combinatorial Model. *Glottotheory. International Journal of Theoretical Linguistics 2/1* , pp. 99–110.

Oakes, M. P. (1998). *Statistics for CorpusLinguistics.* Edinburgh: Edinburgh University Press.

Weisstein, E. W. (2012). *Confidence Interval.* [online] .Cit 2012-10-28. MathWorld – A Wolfram Web. Resource: http://mathworld.wolfram.com/ConfidenceInterval.html

Yates, F. (1984). Tests of Significance for 2 × 2 Contingency Tables. *Journal of the Royal Statistical Society. Series A (General)* , pp. 426–463.