wam package: computing Word association measure

Bernard Desgraupes and Sylvain Loiseau
<bernard.desgraupes@u-paris10.fr>, <sylvain.loiseau@univ-paris13.fr>

December 22, 2017

Abstract

Contents

1 Introduction: Indicators of word association **2**

2 Functions for computing words association strength **2**
- 2.1 Introduction 2
- 2.2 Arguments 2
- 2.3 Comparison with contingency table 2
- 2.4 Comparison between function 3
- 2.5 Log-likelihood 6
- 2.6 Specificities 7
  - 2.6.1 Analysis of the specificities indicator : Standard indicator (method="base") 8
  - 2.6.2 Analysis of the specificities indicator : log (method="log") 12
- 2.7 z 15
- 2.8 t 16
- 2.9 chisq 17
- 2.10 fisher 18

3 High level interface **19**

4 Bibliography **20**

---

1 Introduction: Indicators of word association

This package contains

- an implementation of serveral functions for computing word association strength ;
- a hig level set of functions for conveniently apply these functions to corpora.

Word association can serve two goals:

- analyzing the association strength between a word and a subcorpora
- analyzing the association strength between two words (the tendency of these two words to co-occur).

2 Functions for computing words association strength

2.1 Introduction

Several low-level functions allow for computing association strengh given row data and according to several indicators proposed in the literature.

All association measure functions are prefixed with ”wam.” and return a numeric vector indicating the association strengh.

2.2 Arguments

All word association measure functions have the first four arguments: (N, n, K, k), where:

1. N is the total size of the corpus
2. n is the size of the subcorpus (or the frequency of word2)
3. K is the frequency of word1
4. k is the sub-frequency of word1 in the subcorpus (or the number of co-occurrence between word1 and word2)

Arguments are recycled.

2.3 Comparison with contingency table

These four arguments can be easily turn into the ”contingency table” used in some publications:

	word1	¬ word1	Total
subcorpus (or word2)	11	12	R1
¬ subcorpus (or word2)	21	22	R2
Total	C1	C2	N

where :
Conversion from N, n, K, k :

	word1	¬ word1	Total
subcorpus (or word2)	k	$n - k$	n
¬ subcorpus (or word2)	$K - k$	$N - K - (n - k)$	$N - n$
Total	K	$N - K$	N

Conversion to N, n, K, k :

- $N = N$
- $n = O11 + O12$
- $K = O11 + O21$
- $k = O11$

  The functions `make.contingency` and `make.list` help converting from one format toward the other.

2.4 Comparison between function

TODO : max, mode (expected) for each function, negative or positive, etc.

The indicator, unless otherwise stated in the help pages of the functions, are positive when word1 is over-represented ("attracted"), and negative when word1 is under-represented.

In absolute value, the more the word is over-representend or under-represented, the more the association measure is high.

```
> data(robespierre, package="wam")
> head(robespierre)

  types parts   k     N    K    n
1    de    D1 464 61449 3173 8395
2    la    D1 365 61449 2788 8395
3   les    D1 281 61449 2123 8395
4    et    D1 227 61449 1708 8395
5    le    D1 200 61449 1351 8395
6     l    D1 188 61449 1287 8395

> peuple_D4 <- robespierre[robespierre$types=="peuple" & robespierre$parts == "D4",]
> peuple_D4
```

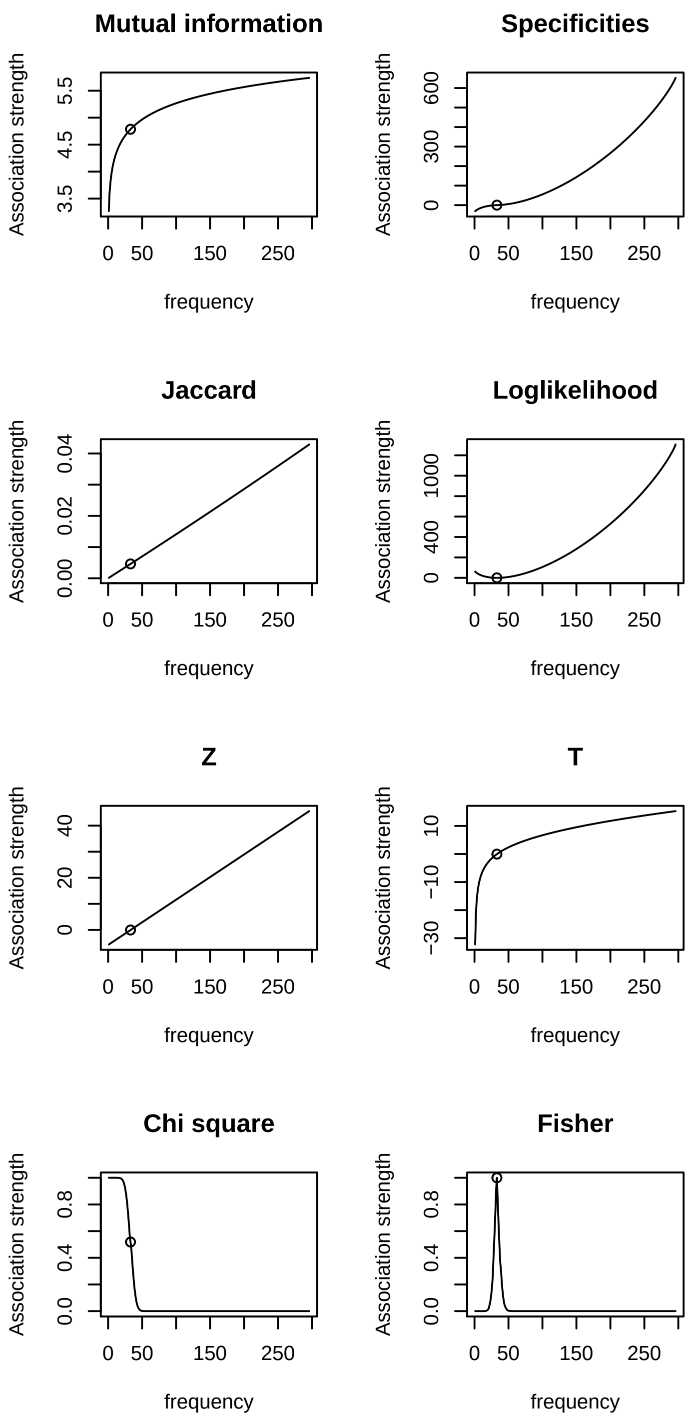
```
     types parts  k     N   K    n
495 peuple    D4 14 61449 296 6903
```

```
> N <- peuple_D4$N
> n <- peuple_D4$n
> K <- peuple_D4$K
> k <- peuple_D4$k
> maxk <- min(K,n)
> maxk
```

```
[1] 296
```

```
> allk <- 0:maxk
> expected = round(K * n / N)
> expected
```

```
[1] 33
```


Mutual information
Specificities
Jaccard
Loglikelihood
Z
T
Chi square
Fisher
Association strength
frequency
0
50
150
250
3.5
4.5
5.5
0
300
600
0.00
0.02
0.04
0
400
1000
0
20
40
−30
−10
10
0.0
0.4
0.8

2.5 Log-likelihood

See Dunning 1993.

```
> wam.loglikelihood(N, n, K, k);
```

```
[1] 15.7202
```

```
> expected <- round(K * n / N)
> wam.loglikelihood(N, n, K, expected);
```

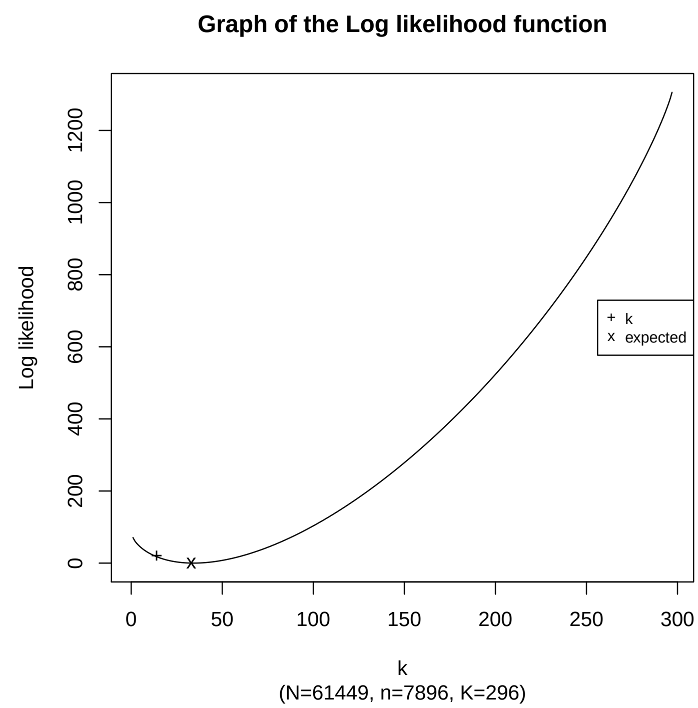
```
[1] 0.002162702
```

Graph of the function :

```
> maxk <- min(K,n)
> maxk
```

```
[1] 296
```

```
> allk <- 0:maxk
> plot(wam.loglikelihood(N, n, K, allk),
+      type="l", xlab="k", ylab="Log likelihood",
+           main="Graph of the Log likelihood function",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.loglikelihood(N, n, K, k), pch="+")
> points(expected, wam.loglikelihood(N, n, K, expected), pch="x")
> legend("right",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



2.6 Specificities

See Lafon 1980.

```
> wam.specificities(N, n, K, k, method="base");
```

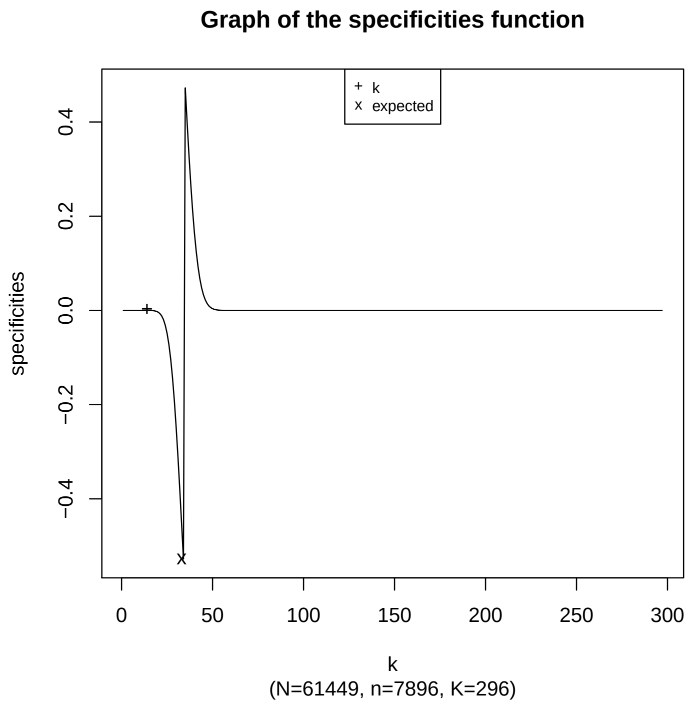
```
[1] -6.693709e-05
```

```
> wam.specificities(N, n, K, expected, method="base");
```

```
[1] -0.5276713
```

Graph of the function:

```
> plot(wam.specificities(N, n, K, allk, method="base"),
+      type="l", xlab="k", ylab="specificities",
+           main="Graph of the specificities function",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.specificities(N, n, K, k, method="base"), pch="+")
> points(expected, wam.specificities(N, n, K, expected, method="base"), pch="x")
> legend("top",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



2.6.1 Analysis of the specificities indicator : Standard indicator (method="base")

The presentation below follows (Lafon, 1980).

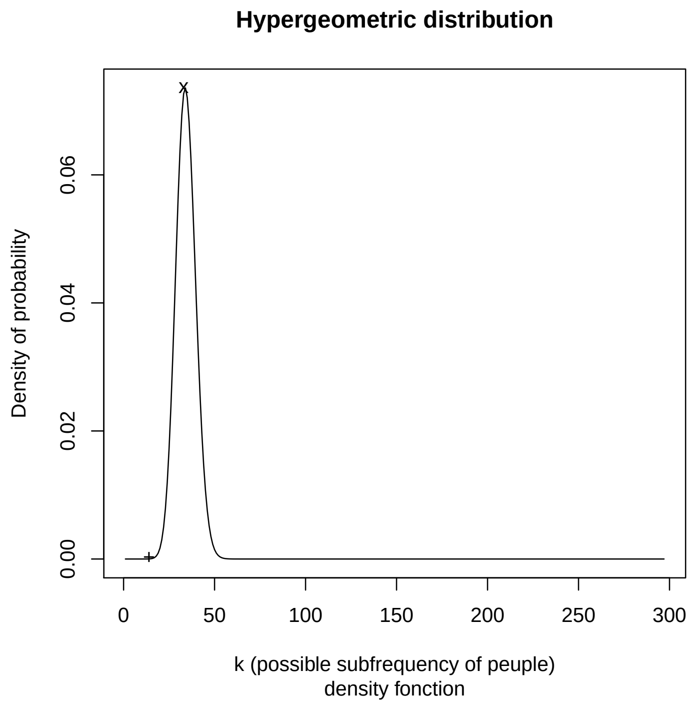
The specificities indicator is based on the hypergeometric distribution. This distribution give the probability associated with a drawing without replacement.

For all the possible subfrequencys of *peuple* in the fourth discourses we can compute the density of probability in the hypergeometric distribution. The graph contains also the observed frequency as well as the mode. The mode is the closest positive integers to the expected frequency.

```
> mode <- floor((n+1)*(K+1)/(N+2));
> mode
```

```
[1] 33
```

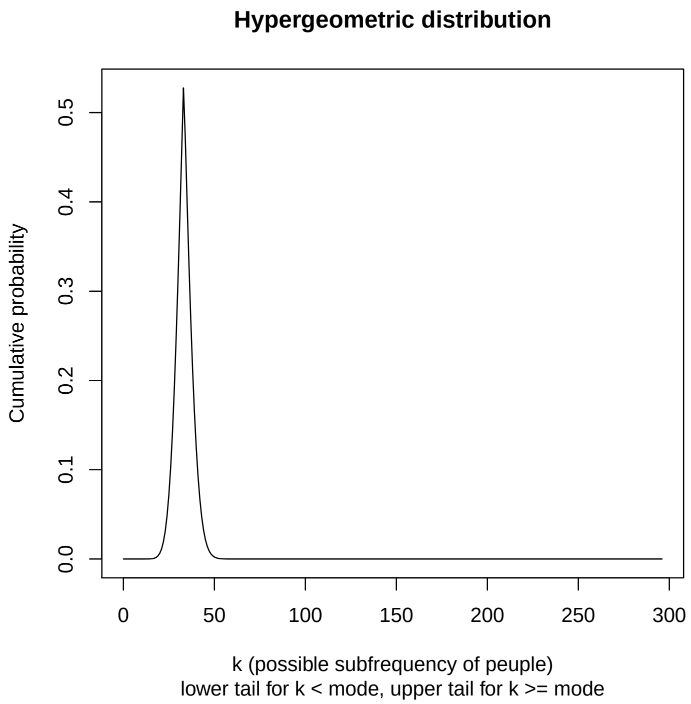
```
> plot(dhyper(allk, K, N-K, n),
+ type="l", xlab="k (possible subfrequency of peuple)", ylab="Density of probability",
+ main="Hypergeometric distribution", sub="density fonction")
> points(k, dhyper(k, K, N-K, n), pch="+")
> points(mode, dhyper(mode, K, N-K, n), pch="x")
```



If the observed frequency is less than the expected frequency, we compute the sum of the probability for a frequency lesser or equal to the observed frequency ($Prob(X \leq k)$) – that is, the cumulative probability.

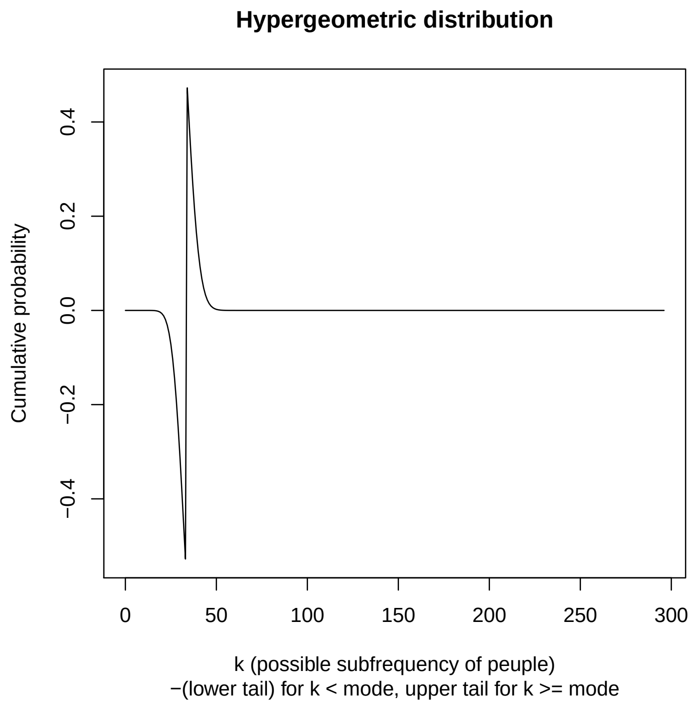
If the observed frequency is greater than the expected frequency, we compute the sum of the probability for a frequency greater to the observed frequency ($Prob(X > k)$) (Lafon 1980 : 141) – that is, the cumulative probability for the upper tail of the distribution.

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n),
+                            phyper(allk-1, K, N-K, n, lower.tail=FALSE))
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="Cumulative probability",
+         main="Hypergeometric distribution",
+         sub="lower tail for k < mode, upper tail for k >= mode")
```



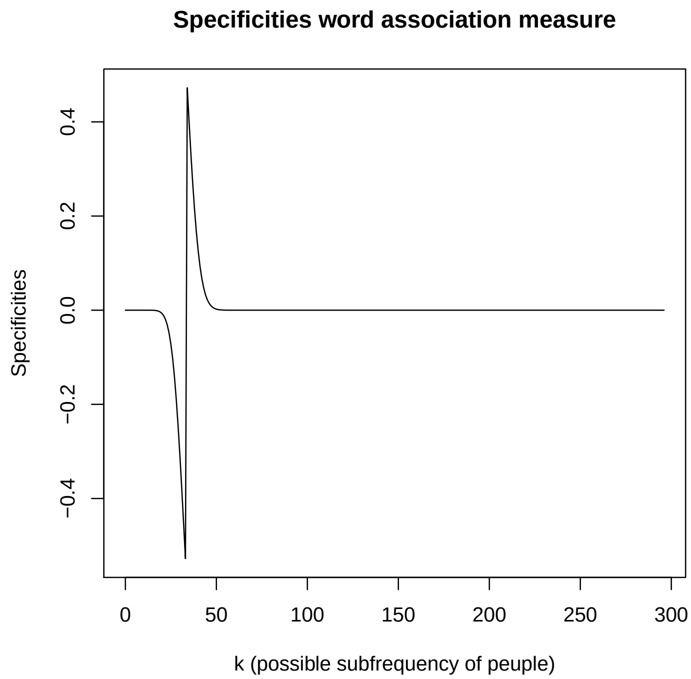
We add a sign: negative if the frequency is lower than expected, positive if it is greater.

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n),
+                         phyper(allk-1, K, N-K, n, lower.tail=FALSE))
> y <- ifelse(allk <= mode, -y, y);
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="Cumulative probability",
+         main="Hypergeometric distribution",
+         sub="-(lower tail) for k < mode, upper tail for k >= mode")
```



It is the standard Specificities function (Lafon 1980), as implemented in the function wam.specificities with method="base" :

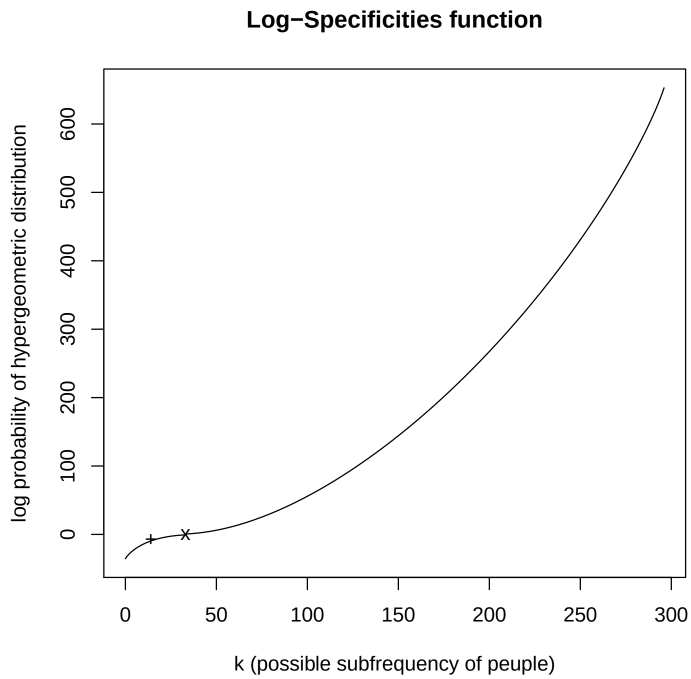
```
> plot(allk, wam.specificities(N, n, K, allk, method="base"),
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="Specificities",
+         main="Specificities word association measure")
```



2.6.2 Analysis of the specificities indicator : log (method="log")

In order to ease the reading, log are used:

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n, log.p=TRUE),
+                          phyper(allk-1, K, N-K, n, lower.tail=FALSE, log.p=TRUE))
> y <- ifelse(allk <= mode, -abs(y), abs(y));
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="log probability of hypergeometric distribution",
+         main="Log-Specificities function");
> points(k, phyper(k, K, N-K, n, log.p=TRUE), , pch="+")
> points(mode, phyper(mode, K, N-K, n, log.p=TRUE), pch="x")
```



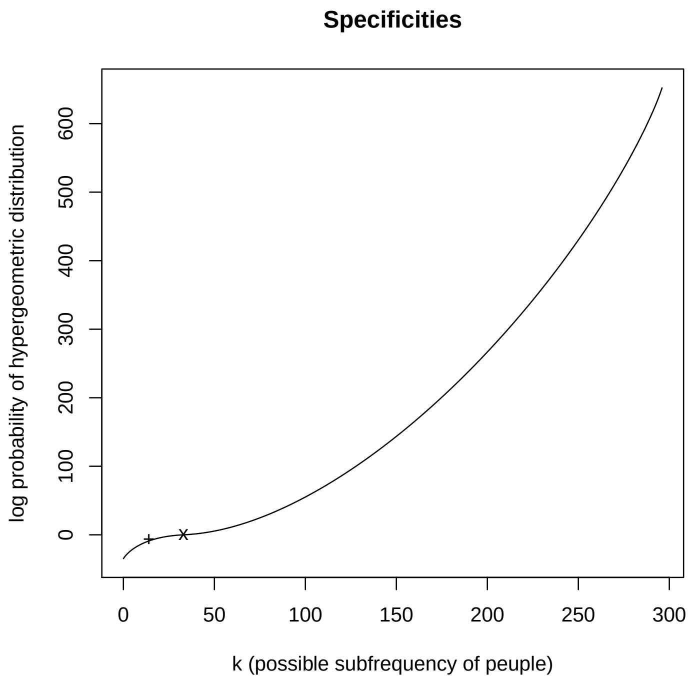
Another issue is that the mode is different from 0:

```
> phyper(mode, K, N-K, n, log.p=TRUE)
```

```
[1] -0.6392818
```

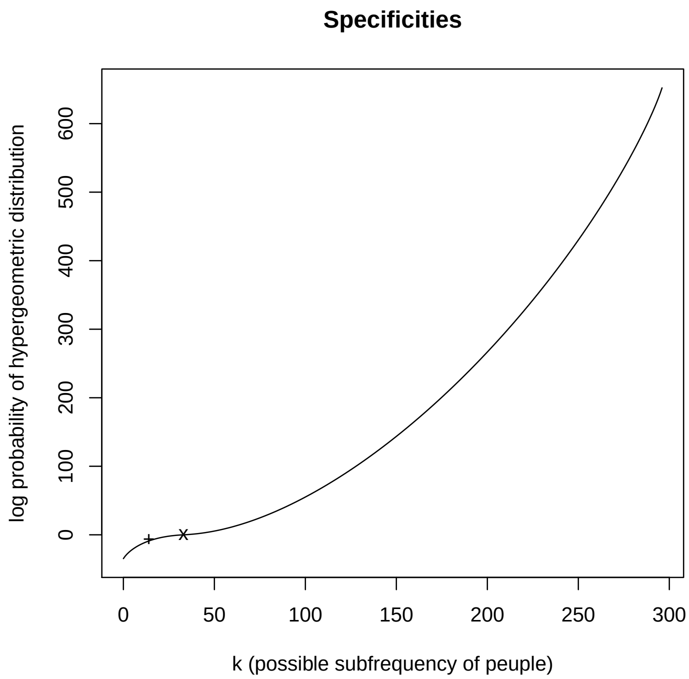
in order to have $mode = 0$, the value of the mode is substracted from all values:

```
> y <- ifelse(allk <= mode, phyper(allk, K, N-K, n, log.p=TRUE),
+                          phyper(allk-1, K, N-K, n, lower.tail=FALSE, log.p=TRUE))
> cdmo <- phyper(mode, K, N-K, n, log.p=TRUE);
> y <- ifelse(allk <= mode, -abs(cdmo-y), abs(cdmo-y));
> plot(allk, y,
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="log probability of hypergeometric distribution",
+         main="Specificities")
> points(k, wam.specificities(N, n, K, k, method="log"), pch="+")
> points(mode, wam.specificities(N, n, K, mode, method="log"), pch="x")
```



That is the wam.specificities function with method="log":

```
> plot(allk, wam.specificities(N, n, K, allk, method="log"),
+         type="l", xlab="k (possible subfrequency of peuple)",
+         ylab="log probability of hypergeometric distribution",
+         main="Specificities")
> points(k, wam.specificities(N, n, K, k, method="log"), pch="+")
> points(mode, wam.specificities(N, n, K, mode, method="log"), pch="x")
```



where the mode is 0 :

```
> wam.specificities(N, n, K, mode, method="log");
```

```
[1] 0
```

2.7 z

```
> wam.z(N, n, K, k);
```

```
[1] -3.338591
```

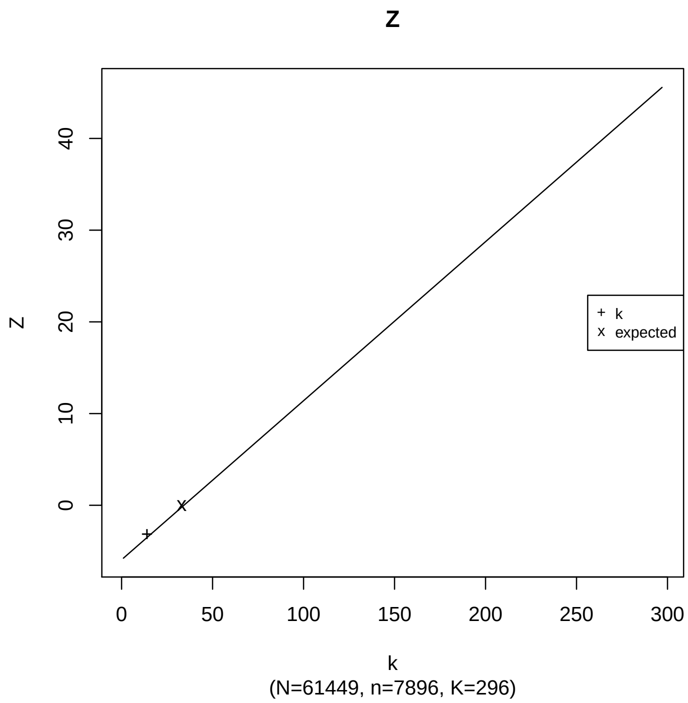
```
> wam.z(N, n, K, mode);
```

```
[1] -0.04366125
```

Graph of the function :

```
> plot(wam.z(N, n, K, allk),
+       type="l", xlab="k", ylab="Z",
+            main="Z",
+            sub="(N=61449, n=7896, K=296)")
> points(k, wam.z(N, n, K, k), pch="+")
```

```
> points(mode, wam.z(N, n, K, mode), pch="x")
> legend("right",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



2.8 t

See Church et al. 1991.

```
> wam.t(N, n, K, k);
```

```
[1] -5.145252
```

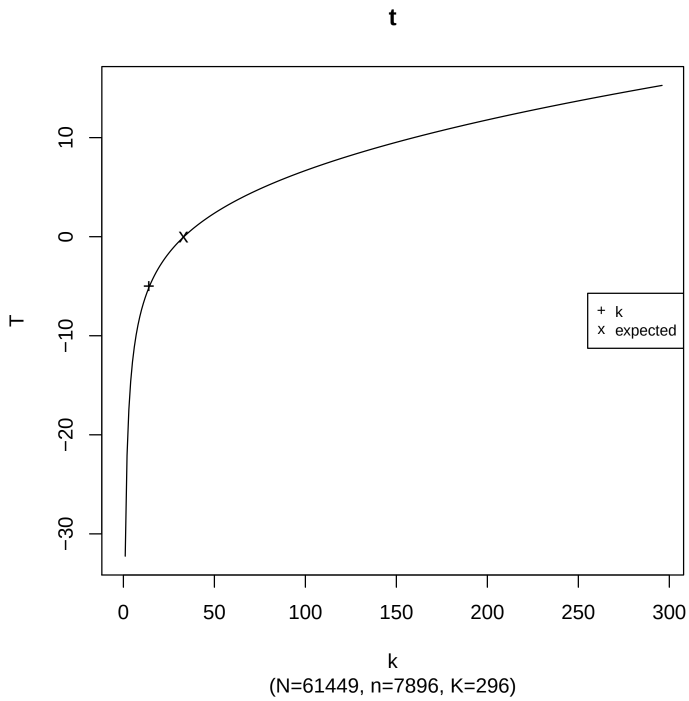
```
> wam.t(N, n, K, mode);
```

```
[1] -0.04382749
```

Graph of the function :

```
> plot(allk, wam.t(N, n, K, allk),
+      type="l", xlab="k", ylab="T",
+           main="t",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.t(N, n, K, k), pch="+")
```

```
> points(mode, wam.t(N, n, K, mode), pch="x")
> legend("right",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



2.9 chisq

```
> wam.chisq(N, n, K, k);
```

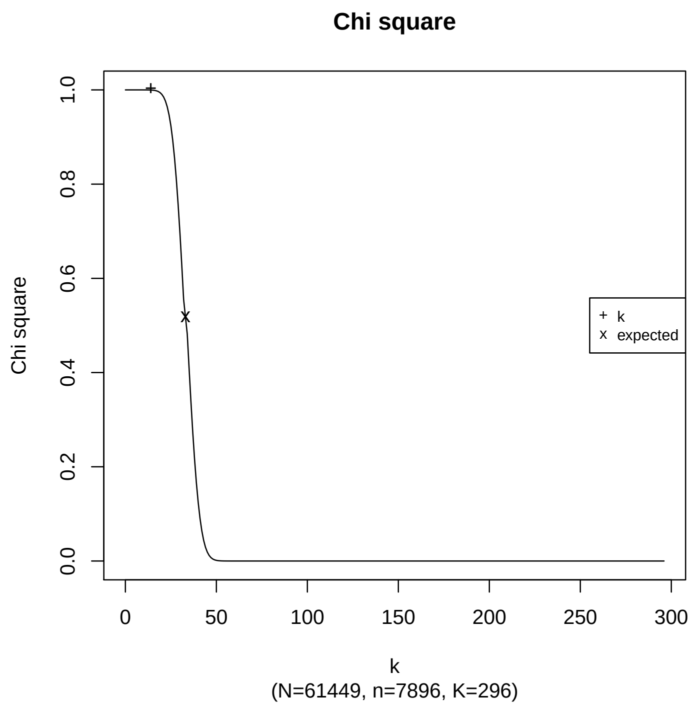
```
[1] 0.9997298
```

```
> wam.chisq(N, n, K, mode);
```

```
[1] 0.5182654
```

Graph of the function :

```
> plot(allk, wam.chisq(N, n, K, allk),
+      type="l", xlab="k", ylab="Chi square",
+           main="Chi square",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.chisq(N, n, K, k), pch="+")
> points(mode, wam.chisq(N, n, K, mode), pch="x")
> legend("right",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



2.10 fisher

```
> wam.fisher(N, n, K, k);
```

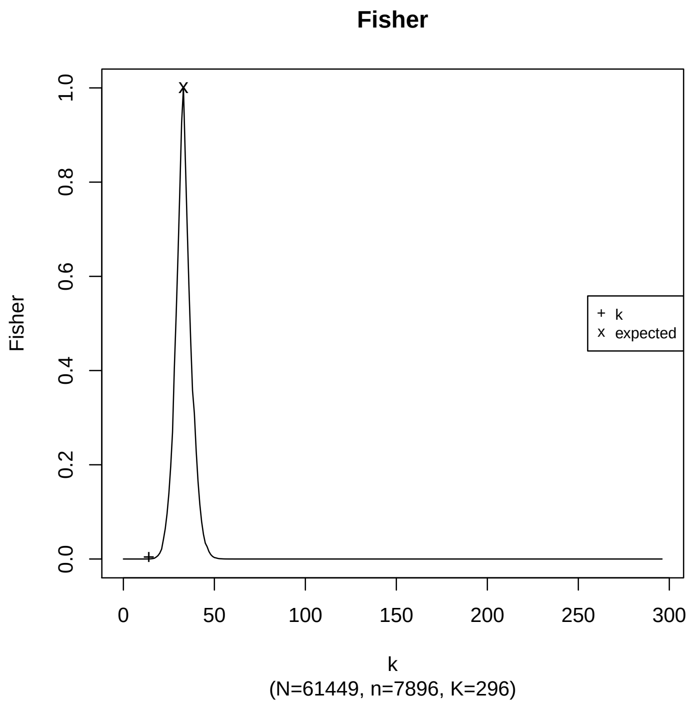
```
[1] 0.0001353407
```

```
> wam.fisher(N, n, K, expected);
```

```
[1] 1
```

Graph of the function :

```
> plot(allk, wam.fisher(N, n, K, allk),
+      type="l", xlab="k", ylab="Fisher",
+           main="Fisher",
+           sub="(N=61449, n=7896, K=296)")
> points(k, wam.fisher(N, n, K, k), pch="+")
> points(expected, wam.fisher(N, n, K, expected), pch="x")
> legend("right",legend=c("k","expected"), pch=c("+","x"), cex=0.75)
```



3 High level interface

The function *wam* provide a more high level interface to the actual functions.

This function allows for computing several indicators at the same time.

It takes also as argument the subcorpus name and lexical types.

```
> rm(list=ls())
> data(robespierre, package="wam")
> attach(robespierre)
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"),
+ types=types, parts=parts)
```

function allows for retrieving the basic information (see the manual page for WordAssociation):

```
> head(k(wam.res));
```

```
[1] 464 365 281 227 200 188
```

```
> head(association(wam.res));
```

```
     loglikelihood specificities
[1,]     2.5758482     2.1848933
[2,]     0.8133694    -0.9840740
[3,]     0.3408187    -0.5627793
[4,]     0.2069236    -0.3991217
[5,]     1.4940429     1.4588965
[6,]     0.9787803     1.0712966
```

```
> indicator.name(wam.res)
```

```
[1] "loglikelihood" "specificities"
```

The print function allows for an easy to use reading of the results. See :

```
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"),
+ types=types, parts=parts)
> print(wam.res, from=1, to=10, parts="D4");
```

```
Printing association measure for 1 part(s); from: 1 to: 10
Corpus size: 61449
Sorted by: loglikelihood
-----------------------------------------------------------------------------
word                | sub freq | tot freq |   loglikelihood |   specificities
-----------------------------------------------------------------------------
.............................................................................
Part name: D4
Part size: 6903 tokens.
Positive specificities printed: 10
Negative specificities printed: 0
bourdon             |       20 |       20 |           87.50 |           43.25
nous                |        3 |      430 |           79.45 |          -40.51
vous                |        6 |      424 |           63.13 |          -32.58
salut               |       33 |       91 |           39.00 |           21.12
comité             |       35 |      103 |           37.31 |           20.20
fabre               |       13 |       19 |           34.59 |           18.43
convention          |       37 |      126 |           30.52 |           16.88
public              |       32 |      108 |           26.84 |           14.98
ils                 |       21 |      419 |           20.14 |          -11.35
il                  |      105 |      605 |           20.12 |           11.70
```

4 Bibliography

Kenneth Ward Church, Patrick Hanks (1990) ”Word Association Norms, Mutual Information, and Lexicography” *Computational Linguistics*, 16/1, pages 22-29. http://www.aclweb.org/anthology/P89-1010.pdf

Chaudhari, D. L., Damani, O. P. & Laxman, S. 2011. "Lexical co-occurrence, statistical significance, and word association". In: Conference on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK, July 27-31). pp. 1058-68

http://www.aclweb.org/anthology-new/D/D11/D11-1098.pdf

Church K., Gale W., Hanks P., Hindle D. 1991. "Using Statistics in Lexical Analysis", In: Zernik U. (ed.) Lexical Acquisition: Exploiting on-line resources to build a lexicon. Hillsdale NJ: Lawrence Erlbaum, pp. 115-164.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence." In: Computational Linguistics. 19(1). Pp 61–74.

http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf

Hofland, K. and Johanssen, S. 1989. Frequency analysis of English vocabulary and grammar, based on the LOB corpus. Oxford: Clarendon.

Kilgarriff, A. 1996. "Which words are particularly characteristic of a text? A survey of statistical approaches." In: Proceedings, ALLC-ACH '96. Bergen, Norway.

http://www.cse.iitb.ac.in/~shwetaghonge/prec_recall.pdf

Lafon P. 1980. "Sur la variabilit? de la fr?quence des formes dans un corpus". *Mots*, 1, 1980, 127–165.

http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008.

Stefanowitsch A. & Gries St. Th. 2003 "Collostructions: Investigating the interaction of words and constructions", *International Journal of Corpus Linguistics*, 8/2, 209-234.