Package 'kldtools'

November 15, 2021

Type Package

Title Kullback-Leibler Divergence and Other Tools to Analyze Frequencies

Version 1.2

Date 2021-11-15

Description
Most importantly, calculates Kullback-Leibler Divergence (KLD), Turing's perspective estimator and their confidence intervals.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation no

Author Alexey Shipunov [cre],
Kateryna Krykoniuk [aut],
Jasjeet Sekhon [ctb]

Maintainer Alexey Shipunov <dactylorhiza@gmail.com>

Repository CRAN

Date/Publication 2021-11-15 12:10:29 UTC

R topics documented:

`kldtools` *Kullback-Leibler Divergence (KLD) and Turing's perspective estimator*

Description

Calculates three estimators of Kullback-Leibler Divergence (KLD): KLD, symmetrized KLD and Turing's perspective KLD along with their confidence intervals.

Usage

```
kldtools(x, y, threshold=0.975)
```

Arguments

`x`	The first vector of frequencies with a length of 2 or more
`y`	The second vector of frequencies with a length of 2 or more
`threshold`	The threshold for declaring statistical significance, the default is 0.975 (5%)

Details

The function computes three estimators of Kullback-Leibler Divergence (KLD): KLD, symmetrized KLD and Turing's perspective KLD, based on Zhang (2017). It compares two empirical, discrete distributions with the confidence interval estimate. The limiting distribution in this method is normal. Simply speaking, these estimators measure the difference between two probability distributions.

More specifically, the function calculates the confidence intervals for these KLD estimators with the measure of standard deviation (sd), for which the equation is different from that of the most known theories in the field (for more detail, see Zhang 2017, Section 5.3):

sd^2 = g(t(v))*sum(v)*g(v)

However, our function uses a corrected formula of the vector g(v) in the above-given formula, the inconsistency in which has been identified by comparing the results obtained from the application of this formula and its simplified version for k = 2 (Zhang 2017: 187).

A deeper enquiry into this problem has led to detecting the formatting error in formula (5.94) of the book (Zhang 2017: 185): it is missing the elements for the p distribution. We would like to thank Dr. Jialin Zhang for verifying the correct formula of the vector in the calculation of the variance of the KLD plug-in estimator.

Also, note that if there are only two elements in probability distributions, an infinite bias emerges (Zhang & Grabchak 2014), which makes the estimation less reliable.

The measure of symmetrized KLD (S) is calculated with the following formula:

S = S(p, q) = 1/2*(KLD(p||q) + KLD(q||p)

Although similar to Jensen-Shannon Divergence (JSD), this measure is different in that it is not a smoothed version of divergence.

Please note that the discussed estimators (which are based on empirical data) allow for a negative value of KLD (i.e. in the values of the CIs and Turing's perspective estimator), despite the proven fact that theoretical KLD should always take a non-negative value. Even though precluded by Zhang's (2017: 152, Theorem 5.1) theory, it is possible to apply these tests, since, in fact, they compare the absolute values of fluctuation of the properly normalized empirical counterparts of KLD around a theoretical value of KLD with the quantile of normal distribution.

The application areas for these three estimators might be different. By way of suggestion, the KLD estimator is appropriate for a study of differences between distributions in systems (samples), whose design criteria differ. On the other hand, symmetrized KLD may be more suitable for the systems (samples) with a more similar design.

In addition, the differences between KLD and symmetrized KLD can inform researchers on a degree of symmetry between the systems (samples): the larger the difference between these measures, the greater the asymmetry between their distributions.

Finally, Turing's perspective estimator is believed to yield more precise results and is appropriate for the data containing zeros (i.e. only in the q distribution), which are handled with the augmentation added to the formula of standard deviation (i.e. to the vector g(v) in the above-mentioned formula of sd).

Value

The list with the following components: "KLD" stands for the measure of KLD; "KLD.s" for the measure of symmetrized KLD; "Turing" for the measure of Turing's perspective KLD - and "*sd" for their respective standard deviations; and "*ci.left" for the lower and "*ci.right" for the upper limits of their respective confidence intervals - "*CI".

Author(s)

Kateryna Krykoniuk, Alexey Shipunov

References

Zhang, Zh. & Grabchak, M. (2014). Nonparametric estimation of Kullback-Leibler divergence. Neural computation 26 (11), pp. 2570-2593. DOI: 10.1162/NECO_a_00646

Zhang, Zh. (2017). Statistical Implications of Turing's Formula. Newark: John Wiley & Sons, Incorporated.

Examples

```
data <- data.frame(V1=c(1213, 57683, 74466, 44419, 17481, 3403, 42252, 7045,
 29445, 15004, 21337, 1892, 21861, 238, 26574, 17579),
 V2=c(3185, 29692, 12570, 26081, 4992, 1659, 16592, 1748, 37583, 6751, 10188,
 355, 8116, 9, 5064, 1846))

kldtools(data$V1, data$V2)
```

ksboot *Bootstrapping based on the Kolmogorov-Smirnov test*

Description

Performs bootstrapping with the Kolmogorov-Smirnov test to estimate differences between frequencies

Usage

```
ksboot(x, y, nboots=1000, alternative=c("two.sided", "less", "greater"))
```

Arguments

x	The first vector
y	The second vector
nboots	The number of bootstraps to perform
alternative	The type of alternative hypothesis (the default is "two.sided")

Details

This bootstrap version of the Kolmogorov-Smirnov test is suitable for estimating not only continuous but also frequency distributions. This is because bootstrap theories suggest that the asymptotic theory of estimates (which is built on the bootstrapping data) is, in a sense, similar to the asymptotic theory of large data sets. Hence, although the Kolmogorov-Smirnov test is initially designed for continuous distributions, in bootstrapping, it is possible to apply this method to discrete random variables, for which an empirical distribution function is built on the observed frequencies (see Abadie 2002 for an example).

Value

The list with the following components: "ksboot.pvalue" for the bootstrap p-value of the Kolmogorov-Smirnov test, calculated for the null hypothesis that the probability densities of two compared distributions are the same; "nboots" for the number of the completed bootstraps.

Author(s)

Jasjeet S. Sekhon, Alexey Shipunov

References

Abadie A. 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. Journal of the American statistical Association. 97: 284-292.

See Also

ks.test

Examples

```
data <- stack(data.frame(V1=c(1213, 57683, 74466, 44419, 17481, 3403, 42252, 7045,
 29445, 15004, 21337, 1892, 21861, 238, 26574, 17579),
 V2=c(3185, 29692, 12570, 26081, 4992, 1659, 16592, 1748, 37583, 6751, 10188, 355,
 8116, 9, 5064, 1846)))

ksboot(data$values[data$ind == "V1"], data$values[data$ind == "V2"])

pairwise.table(function(i, j)
 suppressWarnings(ksboot(data$values[as.integer(data$ind) == i],
 data$values[as.integer(data$ind) == j])$ksboot.pvalue),
 levels(data$ind), p.adjust.method="bonferroni")

pairwise.table(function(i, j)
 suppressWarnings(ksboot(data$values[as.integer(data$ind) == i],
 data$values[as.integer(data$ind) == j])$ksboot.pvalue),
 levels(data$ind), p.adjust.method="none")
```

Index