## Package 'heteroplasmy'

## December 12, 2023

**Title** Calculation of the standard error of the variance for heteroplasmy data **Version** 0.0.3.1

**Description** The package offers different methods to quantify uncertainty by calculating the standard errors of the variance of given data. The included functions are primarily aimed to heteroplasmy data, which are not assumed to follow a predefined distribution and the sample size is usually low. Included, there is a set of synthetic datasets to use, as well as real heteroplasmy data from mouse speci-

men, found in Burgstaller et al. (2018) [https://doi.org/10.1038/s41467-018-04797-

2]. The code and the methods described in the package are used in [Giannakis, K., Broz, A.K., Sloan, D.B. and Johnston, I.G., 2023. Avoiding misleading estimates us-

ing mtDNA heteroplasmy statistics to study bottleneck size and selection. G3: Genes, Genomes, Genetics, 13(6)], so please cite that work in case you use the hetero-

**License** `use\_mit\_license()`, `use\_gpl3\_license()` or friends to pick a license

**Depends** R (>= 3.1.0)

plasmy package.

URL https://github.com/kostasgian21/heteroplasmy

**Encoding** UTF-8 **LazyData** true

**Roxygen** list(markdown = TRUE)

RoxygenNote 7.2.1

Imports kimura,

stats

Remotes lbozhilova/kimura

## R topics documented:

analyticVar	2
analyticVarKimura	3
bootstrapVar	4
estimate_parameters_ks	5
estimate_parameters_ml	
heteroplasmyShift	7
hstats	8
invtransfun	8
jackVar	9

2 analyticVar

	joint_neg_log_lik	10
	kimura_lrt	10
	kimura_neg_loglik	11
	ks_dist	12
	maxlik	12
	maxlikboot	13
	mousedataBroz	14
	mousedataFreyer	14
	mousedataHB	15
	mousedataLE	15
	plotStdErrVar	16
	readHeteroplasmyData	16
	test_kimura_par	17
	transfun	18
Index		19

analyticVar

Analytic calculation of the standard error of the variance

#### Description

This model-free method to calculate analytically the standard error of the variance. It's based on the use of the appropriate h-statistic as an estimator, as default. It is offers a corrected version of the method described in *Wonnapinij et al.*. If instead the *Wonnapinij et al.* method shall be used, use *method="Wonnapinij"*.

#### Usage

```
analyticVar(data, normal = FALSE, method = "hstatistic")
```

#### Arguments

The input data in the form of a vector. NA values are omitted.

normal Parameter that indicates if the normal approximation should be used instead of the general formula from (Wilks, S. S. (1962).Mathematical Statistics). Default is FALSE.

method What method to use for the estimation of the standard error of the variance.

Accepted values are "hstatistic" (default) and "Wonnapinij".

#### Value

The analytically derived standard error of the variance of data.

```
# size of the sample n=50 #generate a random sample of size n from a normal distribution data_ex=rnorm(n,0.5,0.1) analyticVar(data)
```

analyticVarKimura 3

```
mouseData=readHeteroplasmyData("HB")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
analyticVar(mouseData1,method="hstatistic")

# use the package data and load it to variable mouseData
mouseData=mousedataLE
# calculate the standard error of the variance for the LE oocyte sample #3
bootstrapVar(mouseData[,3])
```

analyticVarKimura

Kimura-based calculation of the standard error of the variance

#### **Description**

A Kimura-distribution-based method that calculates analytically the standard error of the variance (see Eq. 11 in *Wonnapinij et al.*, 2010). It uses only the Kimura distribution parameters p and b (passed as arguments).

#### Usage

```
analyticVarKimura(data, p, b)
```

#### **Arguments**

The input data in the form of a vector. NA values are omitted.

p The p parameter of the Kimura distribution. Should be in [0,1].

b The b parameter of of the Kimura distribution. Should be in [0,1].

#### Value

The analytically derived standard error of the variance of data.

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n,0.5,0.1)
analyticVar(data)

mouseData=readHeteroplasmyData("HB")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
analyticVar(mouseData1)

# use the package data and load it to variable mouseData
mouseData=mousedataLE
# calculate the standard error of the variance for the LE oocyte sample #3
analyticVarKimura(mouseData[,3])
```

4 bootstrap Var

bootstrapVar	A boostap method to calculate the standard error of the variance
bootsti apvai	A boostap memou to carefulle the standard error of the variance

#### Description

This function uses the bootrstap method to calculate the uncertainty of the variance of a given sample based on random resampling. The number of the resamples is a parameter (default is 1000). Given that the resampling methods underestimate the uncertainty and thus provide a biased estimation, we offer the unbiased method as a default, although the user may change this option through the biased parameter for experimental purposes (they are strongly advised not to do for real problems with small samples).

#### Usage

```
bootstrapVar(data, nrep = 1000, biased = FALSE)
```

#### **Arguments**

data	The	input	data	in	the	form	of a	vector.	NA	values are	omitted.

nrep The number of bootstrap resamples. Default is 1000. The higher the number of

the samples, the better the bootstrap outcome.

biased A logical parameter to indicate if the user wants the biased version. Resampling

techniques always underestimate statistics like the variance or the standard error

of it for small samples.

#### Value

The standard error of the variance of data and the mean of the bootsrap samples means.

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n,0.5,0.1)
bootstrapVar(data)

mouseData=readHeteroplasmyData("HB")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
bootstrapVar(mouseData1)

# use the package data and load it to variable mouseData
mouseData=mousedataLE
# calculate the standard error of the variance for the LE oocyte sample #3
bootstrapVar(mouseData[,3])
```

estimate\_parameters\_ks 5

```
estimate_parameters_ks
```

Max likelihood estimation of Kimura parameters that minimize the KS statistic

#### Description

Using maximum likelihood to estimate the parameters of a fitted Kimura distribution to the input sample values that minimizes the KS statistic. It is used to showcase that using the KS statistic to support selection needs extra caution.

#### Usage

```
estimate_parameters_ks(h)
```

#### **Arguments**

h

A vector containing heteroplasmy measurements. Every observation should be in [0,1].

#### Value

The maximum likelihood estimates for a fitted Kimura distribution parameters that minimze the KS statistic of a KS test.

#### Author(s)

Kostas and Iain, <us@example.com>

#### References

Site or paper

#### See Also

readHeteroplasmyData

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n,0.5,0.1)
estimate_parameters_ml(data_ex)

mouseData=readHeteroplasmyData("LE")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
estimate_parameters_ks(mouseData1)
```

```
estimate_parameters_ml
```

Max likelihood estimation for Kimura distribution parameters

#### Description

Using maximum likelihood to estimate the parameters of a fitted Kimura distribution to the input sample values.

#### Usage

```
estimate_parameters_ml(h)
```

#### **Arguments**

h

A vector containing heteroplasmy measurements. Every observation should be in [0,1].

#### Value

The maximum likelihood estimates for a fitted Kimura distribution parameters.

#### Author(s)

Kostas and Iain, <us@example.com>

#### References

Site or paper

#### See Also

readHeteroplasmyData

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n,0.5,0.1)
estimate_parameters_ml(data_ex)

mouseData=readHeteroplasmyData("LE")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
estimate_parameters_ml(mouseData1)
```

heteroplasmyShift 7

heteroplasmyShift

Transformed heteroplasmy shift

#### **Description**

A numerical transformation of the heteroplasmy samples in order to work with the heteroplasmy shifts across diverse samples (e.g., due to time or different tissue samples). This transformation is used for comparing a heteroplasmy observation h to a reference value  $h_0$ . It corresponds to the formula:

$$\Delta h = \ln \left( \frac{h(h_0 - 1)}{h_0(h - 1)} \right)$$

#### Usage

heteroplasmyShift(h, h0)

#### **Arguments**

h The heteroplasmy observation. Can be either a single value or a vector of obser-

vations. Every observation should be in [0,1].

h0 The reference heteroplasmy value. Should be in [0,1].

#### Value

The Transformed heteroplasmy shift.

#### Author(s)

Kostas and Iain, <us@example.com>

#### References

Site or paper

#### See Also

readHeteroplasmyData

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n,0.5,0.1)
heteroplasmyShift(data_ex)

mouseData=readHeteroplasmyData("HB")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
heteroplasmyShift(mouseData1,nrep=10000)
```

8 invtransfun

h	S	t a	4†	S

Calculate various statistics for a heteroplasmy set h

#### **Description**

Can enforce an initial h0 or leave as a free parameter. Can can use population or sample statistics. analytic Var offers a simplified version of this function to compute the standard error of the variance.

#### Usage

```
hstats(h, h0 = F, usepopn = F)
```

#### **Arguments**

h A vector containing heteroplasmy measurements. Every observation should be in [0,1].

[ט,ו]

h0 Logical parameter. A particular h0 value Default is to treat h0 as a fit parameter usepopn Logical parameter. Use of population or sample statistics (T and F, respectively)

#### Value

The maximum likelihood for the input data according to the Kimura distribution (using bootstrapping)

#### **Examples**

```
X.1 = rnorm(50, 0.5, 0.1)
hstats(X.1)
```

invtransfun

The inverse function to transfun

#### **Description**

A transformation function to inveret the effect of transfun (ie, the inverse logit transform).

#### Usage

```
invtransfun(x)
```

## Arguments

Х

a real number in [0,1] to be transformed into a real.

#### Value

The inveresed treansformation of transfun.

```
invtransfun(0.71)
```

jackVar 9

jackVar

A jackknife method to compute the uncertainty of heteroplasmy data

#### **Description**

Similarly to the main *bootstrapVar* function that implements the bootstrap method to measure the standard error of the variance, the jackknife technique is another resampling method that can be used for the same purpose. Unlike *bootstrapVar*, *jackVar* (and every jackknife-like method) is deterministic and does not rely on randomness, but instead it uses removals of the sample points, one each time to calculate different sub-samples of size (n-1). Note that the size of the input data should be strictly greater than 1.

#### Usage

```
jackVar(data)
```

#### **Arguments**

data

The input data in the form of a vector. NA values are omitted.

#### Value

The analytically derived standard error of the variance of data and the mean of the bootsrap samples means..

```
# size of the sample
n=50
#generate a random sample of size n from a normal distribution
data_ex=rnorm(n, 0.5, 0.1)
jackVar(data)
mouseData=readHeteroplasmyData("HB")
mouseData1 = mouseData[which(!is.na(mouseData[,1])),1]
jackVar(mouseData1)
# use the package data and load it to variable mouseData
mouseData=mousedataLE
# calculate the standard error of the variance for the LE oocyte sample #3
bootstrapVar(mouseData[,3])
## Not run:
#input data of size 1 will fail
data_ex=rnorm(1,0.5,0.1)
jackVar(data)
## End(Not run)
```

10 kimura\_lrt

joint_neg_log_lik	Joint negative log-likelihood function for several heteroplasmy mea-
	surements

#### Description

Joint negative log-likelihood function for several families' heteroplasmy measurements. theta = [b, h0.1, h0.2, ...] (use h values if use.h0s=F, otherwise initial heteroplasmies are enforced via h0s).

#### Usage

```
joint_neg_log_lik(theta, hlist, use.h0s = F, h0s = -1)
```

#### **Arguments**

theta Kimura parameters p (or h0 here) and b.

hlist TBDD list of different sets of heteroplasmy measurements

use.h0s Logical parameter. TBD

h0s TBDD

#### Value

The negative log likelihood for the list of inputs.

#### **Examples**

```
X.1 = rnorm(50, 0.5, 0.1)

joint_neg_log_lik(c(0.5, 0.91), X.1)
```

kimura\_lrt

Likelihood ratio between two heteroplasmy samples

#### **Description**

A function to perform likelihood ratio test exploring difference in bottleneck size between two heteroplasmy samples.

#### Usage

```
kimura_lrt(h1, h2, use.h0s = F, h1.h0set = 0, h2.h0set = 0)
```

## Arguments

h1	The first heteroplasmy sample.
h2	The two heteroplasmy sample.
use.h0s	Logical parameter. TBD
h1.h0set	Logical parameter.TBD
h2.h0set	Logical parameter.TBD

kimura\_neg\_loglik 11

#### Value

The maximum likelihood for the input data according to the Kimura distribution (using bootstrapping)

#### **Examples**

```
X.1 = rnorm(50,0.5,0.1)
X.2 = rnorm(50,0.5,0.1)
kimura_lrt(X.1,X.2)
```

kimura\_neg\_loglik

Calculate negative log-likelihood for heteroplasmy measurements

#### Description

Calculate negative log-likelihood for a given set of heteroplasmy measurements h and parameters theta = logit(p), logit(b) (we write h0 for p). We can do this by enforcing a particular h0 value (passed as an argument) or treating h0 as a fit parameter (default). The logit transform is used to ensure h0 and b remain in the [0,1] interval regardless of what real-valued argument the numerical optimiser attempts.

#### Usage

```
kimura_neg_loglik(theta, h, h0 = F)
```

#### **Arguments**

theta	Kimura parameters p (or h0 here) and b.
h	A vector containing heteroplasmy measurements. Every observation should be in $[0,1]$ .
h0	Logical parameter. A particular h0 value Default is to treat h0 as a fit parameter

#### Value

The negative log likelihood for the input.

```
X.1 = rnorm(50,0.5,0.1)
kimura_neg_loglik(c(0.5,0.91),X.1)
```

12 maxlik

ks	dist	
ĸs	uist	

Kolmogorov-Smirnov distance function

#### **Description**

A function to calculate the Kolmogorov-Smirnov distance. It is used in estimate\_parameters\_ks to estimate the parameter that minimize the distance in the *optim* function from the *stats* library.

#### Usage

```
ks_dist(theta, ecdf)
```

#### **Arguments**

theta A vector. with two elements. The two Kimura parameters h0 and b.

ecdf A vector containing 10000 values from the empirical cumulative distribution

function of the input heteroplasmy data vector. To be used in the optim function

in estimate\_parameters\_ks.

#### Value

The maximum distance from the theoretical Kimura distribution.

#### **Examples**

```
\begin{array}{l} h = rnorm(20, 0.5, 0.1) \\ ecdf_h <- (stats::ecdf(h))(seq(0, 1, 1e-04)) \\ ks_dist(c(0.5, 0.95), ecdf_h) \end{array}
```

maxlik

Compute maximum likelihood parameters and confidence intervals for heteroplasmy data

## Description

Compute maximum likelihood parameters and confidence intervals for a given heteroplasmy set. We can do this while imposing a specific h0 as an argument or allowing a search over h0 values.

## Usage

```
maxlik(h, conf.level = 0.95, h0 = F)
```

#### Arguments

h	A vector containing heteroplasmy measurements. Every observation should be in
	[0,1].

conf. level The preferred confidence interval calculation, Default value is 0,95 (95%).

h0 Logical parameter. A particular h0 value Default is to treat h0 as a fit parameter

maxlikboot 13

#### Value

The maximum likelihood for the input data according to the Kimura distribution

## **Examples**

```
X.1 = rnorm(50,0.5,0.1)
maxlik(X.1,conf.level=0.95)
```

maxlikboot

Bootstrap estimates for parameters and confidence intervals for heteroplasmy data

#### **Description**

Compute bootstrap estimates for parameters and confidence intervals for a given heteroplasmy set We can do this while imposing a specific h0 as an argument or by allowing a search over h0 values.

#### Usage

```
maxlikboot(h, nboot = 1000, conf.level = 0.95, h0 = F)
```

## Arguments

h	A vector containing heteroplasmy measurements. Every observation should be in $[0,1]$ .
nboot	The number of bootstrap samples. Default value is 1000
conf.level	The preferred confidence interval calculation, Default value is 0,95 (95%).
h0	Logical parameter. A particular h0 value Default is to treat h0 as a fit parameter

## Value

The maximum likelihood for the input data according to the Kimura distribution (using bootstrapping)

```
X.1 = rnorm(50,0.5,0.1)
maxlik(X.1,nboot=10000,0.95)
```

14 mousedataFreyer

mousedataBroz

3 organellar heteroplasmy samples

#### Description

3 distinct organellar heteroplasmy samples from Broz et al. (2022). Heteroplasmy data of the package are in the range [0,1].(Add reference!)

## Usage

mousedataBroz

#### **Format**

A data frame with 32 rows and 3 columns (TO FIX)

## Source

```
http://www.example.info/
```

 ${\tt mousedataFreyer}$ 

mouse PGCs heteroplasmy data

## Description

A dataset containing heteroplasmy values from mouse PGCs taken from Freyer et al. Each column corresponds to a different specimen. NA values have been added to make the number of rows equal. Please remove them after loading. Heteroplasmy data of the package are in the range [0,1].(Add reference!)

#### Usage

mousedataFreyer

## Format

A data frame with 111 rows and 18 columns (TO FIX)

#### Source

http://www.example.info/

mousedataHB 15

mousedataHB

HB oocyte heteroplasmy data

#### Description

A dataset containing heteroplasmy values for the HB oocyte mouse lines. Each column corresponds to a different specimen. NA values have been added to make the number of rows equal. Please remove them after loading. Heteroplasmy data of the package are in the range [0,1].

#### Usage

mousedataHB

#### **Format**

A data frame with 25 rows and 56 columns (TO FIX):

#### **Source**

```
http://www.example.info/
```

mousedataLE

LE oocyte heteroplasmy data

#### Description

A dataset containing heteroplasmy values for the HB oocyte mouse lines. Each column corresponds to a different specimen. NA values have been added to make the number of rows equal. Please remove them after loading. Heteroplasmy data of the package are in the range [0,1].

#### Usage

mousedataLE

#### **Format**

A data frame with 20 rows and 43 columns (TO FIX)

#### Source

```
http://www.example.info/
```

plotStdErrVar

An example plotting function

#### **Description**

This function is used as a toy example on how to represent the data statistics regarding the variance of the sample. The mean variance and its standard error are depicted. Note that this is just an illustration to show that the analytic and the resampling approaches almost match each other.

#### Usage

```
plotStdErrVar(
  data,
  functions = c("normalApr", "analytic", "bootstrap", "jackknife"),
   ...
)
```

#### **Arguments**

data

The input data in the form of a vector. NA values are omitted.

functions

Choose the subset of the functions you wish use for the calculation and subsequent plot of the standard error of the variance. You can use one or a combination of "normalApr", "analytic", "bootstrap", "correctedBoot", and "jackknife".

For now, it outputs all of the aforementioned methods!

#### Warning

This is a plotting function just for demonstration purposes.

#### **Examples**

```
# size of the sample n=50 #generate a random sample of size n from a normal distribution data_ex=rnorm(n,0.5,0.1) plotStdErrVar(data_ex)
```

#### Description

This function allows you to read mouse heteroplasmy data from external files. Use with caution (for now). Calling this function is equivalent to loading the already stored dataframe with heteroplasmy measurements (ie, *mousedataHB,mousedataLE,mousedataFreyer*).

#### Usage

```
readHeteroplasmyData(nameD = "HB")
```

test\_kimura\_par 17

#### **Arguments**

nameD Either "HB" or "LE" or "Freyer" or "Broz".

#### Value

A dataframe containing mouse heteroplasmy data.

#### **Examples**

```
df1=readHeteroplasmyData(nameD="LE")
df2=readHeteroplasmyData(nameD="HB")
df3=readHeteroplasmyData(nameD="Freyer")
df4=readHeteroplasmyData(nameD="Broz")

# or equivalent
df1=mousedataLE
df2=mousedataHB
df3=mousedataFreyer
df4=mousedataBroz
```

test\_kimura\_par

Generalised MC KS test for genetic drift

#### Description

This function is a generalisation of the *test\_kimura* function from the *lbozhilova/kimura* package. It corresponds to a Monte Carlo Kolmogorov-Smirnov test to detect genetic drift by examining deviation from a Kimura distribution.

## Usage

```
test_kimura_par(h, p, b, num_MC = 1000, round = FALSE)
```

#### Arguments

h	A vector containing heteroplasmy measurements. Every observation should be in
---	---

[0,1].

p The p parameter of the Kimura distribution. Should be in [0,1].

b The b parameter of of the Kimura distribution. Should be in [0,1].

num\_MC number of Monte Carlo runs

round a logical argument. True if heteroplasmy fractions are rounded to two significant

digits. Be cautious as rounding can impact the calculated p-values. Default value

is FALSE

#### Value

object of class htest

18 transfun

#### **Examples**

```
data_ex=rnorm(n,0.5,0.1)
  fit = estimate_parameters_ml(data_ex)
  p=fit[1]
  b=fit[2]
test_kimura_par(data_ex,p,b)
```

transfun

A Function to cast real numbers to [0,1]

#### Description

A transformation function to cast any real number onto the interval [0,1]. Equivalent to the inverse logit transform.

#### Usage

```
transfun(x)
```

#### **Arguments**

Х

a real number to be transformed.

#### Value

The transformed cast of the input value to the interval [0,1].

```
transfun(5.1)
```

# Index

* Kimura	kimura_neg_loglik, 11
analyticVarKimura,3	maxlik, 12
* Kolmogorov	maxlikboot, 13
ks_dist, 12	* likelihood
* Smirnov	estimate_parameters_ks, 5
ks_dist, 12	estimate_parameters_ml, 6
* bootstrap	<pre>joint_neg_log_lik, 10</pre>
bootstrapVar, 4	kimura_neg_loglik,11
* datasets	maxlik, 12
mousedataBroz, 14	maxlikboot, 13
mousedataFreyer, 14	* logit
mousedataHB, 15	invtransfun, 8
mousedataLE, 15	transfun, 18
* data	* log
readHeteroplasmyData, 16	joint_neg_log_lik, 10
* distance	kimura_neg_loglik, 11
ks_dist, 12	maxlik, 12
* error	maxlikboot, 13
analyticVar,2	* negative
analyticVarKimura, 3	joint_neg_log_lik, 10
* h-statistic	kimura_neg_loglik, 11
analyticVar,2	maxlik, 12
* heteroplasmy,maximum	maxlikboot, 13
estimate_parameters_ks,5	* plot,standard,error
estimate_parameters_ml, 6	plotStdErrVar, 16
* heteroplasmy,transformation,shift	* resampling
heteroplasmyShift,7	bootstrapVar,4
* heteroplasmy	jackVar, 9
analyticVar, 2	* reverse
analyticVarKimura, 3	transfun, 18
bootstrapVar, 4	* standard
jackVar, 9	analyticVar, 2
readHeteroplasmyData, 16	analyticVarKimura, 3
* inverse	* statistic
invtransfun, 8	hstats, 8
* jackknife	kimura_lrt, 10
jackVar,9	* summary
* joint	hstats, 8
joint_neg_log_lik, 10	kimura_lrt, 10
maxlik, 12	* uncertainty
maxlikboot, 13	bootstrapVar, 4
* kimura	jackVar, 9
joint_neg_log_lik, 10	* variance
J - 11 - 11 - 12 - 12 - 12 - 12 - 12 - 1	

20 INDEX

```
analyticVar, 2
    analyticVar, 2
bootstrapVar, 4
estimate_parameters_ks, 5
\verb"estimate_parameters_m1", 6
heteroplasmyShift, 7
hstats, 8
invtransfun, 8
jackVar, 9
\texttt{joint\_neg\_log\_lik}, \textcolor{red}{10}
\texttt{kimura\_lrt}, \textcolor{red}{10}
kimura_neg_loglik, 11
ks_dist, 12
maxlik, 12
maxlikboot, 13
mousedataBroz, 14
mousedataFreyer, 14
mousedataHB, 15
mousedataLE, 15
{\tt plotStdErrVar}, {\color{red} 16}
readHeteroplasmyData, 5-7, 16
test_kimura_par, 17
transfun, 18
```