Package 'abc.data'

May 5, 2015

Type Package

Title Data Only: Tools for Approximate Bayesian Computation (ABC)

Version 1.0

Date 2015-05-04

Depends R (>= 2.10)

Description Contains data which are used by functions of the 'abc' package.

Repository CRAN

License GPL (>= 3)

NeedsCompilation no

Author Csillery Katalin [aut],
Lemaire Louisiane [aut],
Francois Olivier [aut],
Blum Michael [aut, cre]

Maintainer Blum Michael <michael.blum@imag.fr>

Date/Publication 2015-05-05 11:34:13

R topics documented:

---

human — *A set of R objects containing observed data from three human populations, and simulated data under three different demographic models. The data set is used to illustrate model selection and parameter inference in an ABC framework (see the vignette of the* abc *package for more details).*

Description

data(human) loads in four R objects: stat.voight is a data frame with 3 rows and 3 columns and contains the observed summary statistics for three human populations, stat.3pops.sim is also a data frame with 150,000 rows and 3 columns and contains the simulated summary statistics, models is a vector of character strings of length 150,000 and contains the model indices, par.italy.sim is a data frame with 50,000 rows and 4 columns and contains the parameter values that were used to simulate data under a population bottleneck model. The corresponding summary statistics can be subsetted from the stat.3pops.sim object as subset(stat.3pops.sim, subset=models=="bott").

Usage

```
data(human)
```

Format

The stat.voight data frame contains the following columns:

pi The mean nucleotide diversity over 50 loci in 3 human populations, Hausa, Italian, and Chinese.

TajD.m The mean of Tajima's D statistic over 50 loci in 3 human populations, Hausa, Italian, and Chinese.

TajD.v The variance of Tajima's D statistic over 50 loci in 3 human populations, Hausa, Italian, and Chinese.

Each row represents a simulation. Under each model 50,000 simulations were performed. Row names indicate the type of demographic model.

The stat.3pops.sim data frame contains the following columns:

pi The mean of nucleotide diversity over 50 simulated loci under 3 demographic scenarios: constant size population, population bottleneck, and population expansion.

TajD.m The mean of Tajima's D statistic over 50 simulated loci under 3 demographic scenarios: constant size population, population bottleneck, and population expansion.

TajD.v The variance of Tajima's D statistic over 50 simulated loci under 3 demographic scenarios: constant size population, population bottleneck, and population expansion.

Each row represents a simulation. Under each model 50,000 simulations were performed. Row names indicate the type of demographic model.

The par.italy.sim data frame contains the following columns:

Ne The effective population size.

a The intensity of the bottleneck (i.e. the ratio of the population sizes before and during the bottleneck).

duration The duration of the bottleneck.

start The start of the bottleneck.

Each row represents a simulation.

models contains the names of the demographic models.

Details

Data is provided to estimate the posterior probabilities of classical demographic scenarios in three human populations: Hausa, Italian, and Chinese. These three populations represent the three continents: Africa, Europe, Asia, respectively. `par.italy.sim` may then used to estimate the ancestral population size of the European population assuming a bottleneck model.

It is generally believed that African human populations are expanding, while human populations from outside of Africa have gone through a population bottleneck. Tajima's D statistic has been classically used to detect changes in historical population size. A negative Tajima's D signifies an excess of low frequency polymorphisms, indicating population size expansion. While a positive Tajima's D indicates low levels of both low and high frequency polymorphisms, thus a sign of a population bottleneck. In constant size populations, Tajima's D is expected to be zero.

With the help of the `human` data one can reach these expected conclusions for the three human population samples, in accordance with the conclusions of Voight et al. (2005) (where the observed statistics was taken from), but using ABC.

Source

The observed statistics were taken from Voight et al. 2005 (Table 1.). Also, the same input parameters were used as in Voight et al. 2005 to simulate data under the three demographic models. Simulations were performed using the software ms and the summary statistics were calculated using sample_stats (Hudson 1983).

References

B. F. Voight, A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson and A. Di Rienzo (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. PNAS **102**, 18508-18513.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18** 337-338.

`musigma2` *A set of objects used to estimate the population mean and variance in a Gaussian model with ABC (see the vignette of the* `abc` *package for more details).*

Description

`musigma2` loads in five R objects: `par.sim` is a data frame and contains the parameter values of the simulated data sets, `stat` is a data frame and contains the simulated summary statistics, `stat.obs` is a data frame and contains the observed summary statistics, `post.mu` and `post.sigma2` are data frames and contain the true posterior distributions for the two parameters of interest, μ and σ^2, respectively.

Usage

```
data(musigma2)
```

Format

The `par.sim` data frame contains the following columns:

`mu` The population mean.

`sigma2` The population variance.

The `stat.sim` and `stat.obs` data frames contain the following columns:

`mean` The sample mean.

`var` The logarithm of the sample variance.

The `post.mu` and `post.sigma2` data frames contain the following columns:

`x` the coordinates of the points where the density is estimated.

`y` the posterior density values.

Details

The prior of σ^2 is an inverse χ^2 distribution with one degree of freedom. The prior of μ is a normal distribution with variance of σ^2. For this simple example, the closed form of the posterior distribution is available.

Source

The observed statistics are the mean and variance of the sepal of *Iris setosa*, estimated from part of the `iris` data.

The data were collected by Anderson, Edgar.

References

Anderson, E. (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2-5.

---

ppc — *Data to illustrate the posterior predictive checks for the data* human. ppc *and* human *are used to illustrate model selection and parameter inference in an ABC framework (see the vignette of the* abc *package for more details).*

Description

`data(ppc)` loads in the data frame `post.bott`, which contains the summary statistics calculated from data simulated a posteriori under the bottleneck model (see `data(human)` and the package's vignette for more details).

Usage

```
data(ppc)
```

Format

The `post.bott` data frame contains the following columns:

`pi` The mean nucleotide diversity over 50 loci.

`TajD.m` The mean of Tajima's D statistic over 50 loci.

`TajD.v` The variance of Tajima's D statistic over 50 loci.

Each row represents a simulation. 1000 simulations were performed under the bottleneck model.

Index