Manual for CMRPopHet version: 2.2.3

S.J. Puechmaille

Updated: 14/04/2011

Please, report any suggestions to improve the script, input or output files to: s [dot] puechmaille [at] gmail [dot] com

TABLE OF CONTENT

Here we provide some information on how to calculate population size from Capture-Mark-Recapture data as in Petit and Valière (2006) and perform heterogeneity test as described in Puechmaille and Petit (2007). For further details on the methods themselves and which estimates to use, please read the above mentioned papers as well as Gazey and Staley (1986). These three papers are provided along with the R scripts.

BEFORE RUNNING ANALYSES

Before you run any analyses, you need to make sure you know which working directory is used by R as you will need to copy the 'CMRPopHet-2.2.3.R' file and input files there.
If you are not sure which working directory R is using, simply type in the R console 'getwd()'
On my computer, it gives:

```
> getwd()
[1] "C:/Program Files/R/R-2.10.0/adata"
```

Therefore, in my case, the 'CMRPopHet.R' file and input files will have to be copied in the 'adata' folder.

If you want to change the working directory, you can select your own by typing 'setwd("new directory")' as below:

```
>setwd("C:/Program Files/R/R-2.10.0/adata/CMRPopHet")
```

To verify that the change has been made, we can type 'getwd()' as explained above.

```
> getwd()
[1] "C:/Program Files/R/R-2.10.0/adata/CMRPopHet"
```

ANALYSES

After you have placed the file 'CMRpop-2.2.3.R' into the default R working directory (or your selected working directory), you need to install the functions by typing in the R console: 'source("CMRPopHet-2.2.3.R")'. Then, to run the functions, simply call them and give their arguments. A file 'CMRPopHet-Script.R' (a simple text file) gives example of command lines that can be typed or paste onto the R console to obtain results.

I - Population size estimates following Petit & Valière (2006)

A-For single dataset

The function is named 'CMRpopsize' and needs 4 arguments to run. The arguments are as detailed below:

1- 'name': name of the population (using double inverted comas as shown below)
2- 'samples': number of samples analysed
3- 'indiv': number of individuals analysed (unique genotypes)
4- 'maxPop': Maximum population size to be considered (cf. Petit & Valière, 2006 for details about this parameter)

Two facultative arguments for graphics' settings can also be provided:

5-‘plotNvalue’:	which of the population size estimate to be plotted; must be one of "Mean" (default), "Median" or "Mode "
6-‘plotLimit’:	which of the interval of population size estimate to be plotted; must be one of "HPD95" (default) or "IC95"

Example from Puechmaille & Petit (2007):

```
> CMRpopsize(name="ST2003",samples=46,indiv=19,maxPop=200)
Population: ST2003
Mode-Median-Mean:       22 - 23 - 24
Quantile 95%: [ 19 ; 29 ]
HPD 95%:     [ 20 ; 28 ]
```

A ‘CMR-Pop_ name_ plotNvalue_ plotLimit.pdf’’ file will be created by default and in the case of the above example, it will be named ‘CMR-Pop_ Graph_ST2003 _Mean_HPD95.pdf’. This PDF file shows Posterior probability distribution of the different population sizes. The estimated population size (Mean by default) is represented as a vertical line and the HPD 95% (by default) is depicted as vertical dashed lines.
If we want to plot the Median and IC 95% on the graph, the following should be used:

```
>CMRpopsize(name="ST2003",samples=46,indiv=19,maxPop=200,plotNvalue="Median",plotLimit="IC95")
```

In this case, a file named “CMR-Pop_Graph_ST2003_Median_IC95.pdf” would be created.

B-For multiple datasets

The function is named ‘CMRpopsizeM’ and needs 1 argument to run, the input file name (see example below). The input file content should be as detailed below:
1- *i* rows and 4 columns with no header for the columns (*i*=number of dataset):
2- First column containing the name of the population
3- Second column containing the number of samples analysed
4- Third column containing the number of individuals analysed (unique genotypes)
5- Fourth column containing the maximum population size to be considered
6- See example below (identical to example file named "inputfile-pop.txt")

```
E2003 114   53     500
E2004 145   58     500
P2003 97    38     500
P2004 95    38     500
ST2003      46     19     500
ST2004      37     16     500
```

Example from Puechmaille & Petit (2007):

```
> CMRpopsizeM("inputfile-pop.txt")
Population: E2003
Mode-Median-Mean:        64 - 65 - 66
Quantile 95%: [ 57 ; 77 ]
HPD 95%:      [ 57 ; 75 ]
Population: E2004
Mode-Median-Mean:        66 - 66 - 67
Quantile 95%: [ 60 ; 74 ]
HPD 95%:      [ 60 ; 73 ]
Population: P2003
Mode-Median-Mean:        43 - 43 - 44
Quantile 95%: [ 39 ; 50 ]
HPD 95%:      [ 39 ; 49 ]
Population: P2004
Mode-Median-Mean:        43 - 44 - 44
Quantile 95%: [ 39 ; 51 ]
HPD 95%:      [ 40 ; 50 ]
Population: ST2003
Mode-Median-Mean:        22 - 23 - 24
Quantile 95%: [ 19 ; 29 ]
HPD 95%:      [ 20 ; 28 ]
Population: ST2004
Mode-Median-Mean:        19 - 20 - 21
Quantile 95%: [ 16 ; 27 ]
HPD 95%:      [ 17 ; 26 ]
null device
          1
```

A message with ‘null device 1’ appears at the end of the R console (see above example) and means the PDF file has been created. The results are shown on the R console but are also saved as a text file called "CMR-Pop_ Results.txt"; this file contains 12 columns:

(1) Population name,
(2) Number of samples analysed,
(3) Number of individuals analysed,
(4) Maximum population size considered,
(5) Mode,
(6) Median,
(7) Mean,
(8) 2.5% quantile,
(9) 97.5% quantile,
(10) 2.5% border of the HPD,
(11) 97.5% border of the HPD,

(12) Date and time when the estimate was run.

NB: if the file 'CMR-Pop_ Results.txt 'already exists (i.e. if you previously used this function and did not delete the 'CMR-Pop_ Results.txt' file), the results will be appended at the end of the file.

A second results file (graphics) called 'CMR-Pop_Graph_ plotNvalue_ plotLimit.pdf' will be created by default. This PDF file shows Posterior probability distribution of the different population sizes for each data set on a different page. The estimated population size (Mean by default) is represented as a vertical line and the HPD 95% (by default) is depicted as vertical dashed lines. If you already have a 'CMR-Pop_Graph_ plotNvalue_ plotLimit.pdf' PDF in the folder, it will be replaced. For plotting different estimated values of N, see details in (A). For example, we could use:

```
> CMRpopsizeM("inputfile-pop.txt",plotNvalue="Median",plotLimit="IC95")
```

II - Heterogeneity test on population size estimates following Puechmaille & Petit (2007)

A- For single dataset

The function is named 'CMRpophet' and needs 5 arguments to run (6th argument is facultative). The arguments are as detailed below:

1- 'name':	name of the population (using double inverted comas as shown below)
2- 'obser':	number of individuals'captured' respectively once, twice, three times, etc. In the below example, if no individual had been captured three times, the argument should have been 'obser=c(20,18,0,4,1,1,1)'
3- 'samples':	number of samples analysed
4- 'popsize':	estimated population size (i.e. obtained with the 'CMRpopsize' function)
5- 'replicates':	number of replicates to perform for the test
6- 'alpha':	facultative argument for specifying the significance level (alpha); default=0.05

Example from Puechmaille & Petit (2007):

```
>CMRpophet(name="E2003",obser=c(20,18,8,4,1,1,1),samples=114,popsize=64,replicates=1
000)
```

The significance level in the above example is set to 0.05 (5%) by default but it can be customized using the 'alpha' argument as below:

```
>CMRpophet(name="E2003",obser=c(20,18,8,4,1,1,1),samples=114,popsize=64,replicates=1
000,alpha=0.1)
```

The graph showing the observed and expected number of capture per individual is automatically saved as a PDF with the file name "CMR-Het_Graph_name_alpha=x.xx.pdf". For the example above with alpha set by default, the graph will be saved in a file named

“CMR-Het_Graph_E2003_alpha=0.05.pdf”. This file will be saved in the folder set by default (see paragraph “Before running analyses” for details). If you already have in the folder a PDF named ‘CMR-Het_Graph_E2003_alpha=0.05.pdf’, it will be replaced by the new one.

B- For multiple datasets

The function is named ‘CMRpophetM’ and needs 1 argument to run, the input file name which content should be detailed below:
i rows and k columns with no header for the columns (*k*=number of dataset):
1- First column containing the name of the population
2- Second column containing the number of samples analysed
3- Third column containing the population size estimate (i.e. obtained with the 'CMRpopsize' function)
4- Fourth column containing the number of replicates needed
5- Fifth column containing the number of individuals found once
6- Sixth column containing the number of individuals found twice
7- Seventh column containing the number of individuals found three times,
8- k^{th} column containing the number of individuals found *k*-4 times,
Note that k has to be the same for all datasets, therefore, fill in with '0' whenever needed
See example file named "inputfile-het.txt"
Example from Puechmaille & Petit (2007):

```
>CMRpophetM("inputfile-het.txt")
```

The significance level in the above example is set to 0.05 (5%) by default but it can be customized using the 'alpha' argument as below:

```
>CMRpophetM("inputfile-het.txt",alpha=0.1)
```

The results are automatically saved as a csv (comma separated value) files called "CMR-Het_Results_alpha=x.csv" and "CMR-Het_Summary_alpha=x.csv"
The "CMR-Het_Summary_alpha=x.csv" file contains 3 columns:
(1) Population name
(2) Result of the test (Homogene or Heterogene),
(3) Number of times the observed value is outside the 95% Confidence Interval,
The file name contains the alpha level used for the test (default value or the value set by the user) so this value is not again reported in the results file.

The "CMR-Het_Results_alpha=x.csv" file contains the following (!!!5 rows per population [dataset]!!!):
(1) Population name, (data inputted by the user)
(2) Number of samples analysed, (data inputted by the user)
(3) Population size estimate, (data inputted by the user)
(4) The number of replicates performed, (data inputted by the user)
For each population, there are 5 rows of data containing (From column 5 onwards):

(5) Row 1: vector of the upper limit of the 95%CI of the number of capture (simulation)
(6) Row 2: vector of the observed number of capture (data inputted by the user)
(7) Row 3: vector of the lower limit of the 95%CI of the number of capture (simulation)
(8) Row 4: vector of the average number of capture (simulation)
(9) Row 5: if for a given number of capture, the observed data is comprised within the 95%CI, a value of '0' is shown, if the observed data falls outside, a value of '1' is given.

Similarly to what is explained for the single data set function, the graph showing the observed and expected number of capture per individual is automatically saved with the file name “CMR-Het_Graph_alpha=x.xx.pdf”. This PDF contains the graph for each data set.
This file will be saved in the folder set by default (see paragraph “Before running analyses” for details).

CHANGES FROM PREVIOUS VERSIONS

Since 2.2.2
1-Possibility to choose which population size estimate and interval to plot on the graphs. See Gazey and Staley (1986) for the relative merits of the different point estimators and confidence intervals (p944-946).
2-Provides also the ‘Mean’ and ‘Median’ for the population size estimate.

Since 2.2.1
1-Graphs are now saved as PDFs.
2-The population size function has been partially recoded to perform faster, especially for large population sizes (N>100; at least twice faster).
3-The calculation of the 2.5% and 97.5% quantiles is now done as follows. Gazey & Staley (1986) said that “*the appropriate quantiles can be reported, i.e., the a and b such that P(N<a)=α/2 and P(N>b)=α/2*”. Nevertheless, P(N<*a*) or P(N>*b*) rarely equal α/2 and instead of reporting the 2.5% and 97.5% quantiles, we should report an interval of N in which the quantiles would be included. Nevertheless, because we consider every integer (between ‘indiv’ and ‘maxPop’) for population size estimate, the interval of quantiles would be very narrow. Let’s take an example with the 97.5% quantiles for the “E2003” data set:
P(N>76)= **0.02916**225 and P(N>77)= **0.02110**531
The value of *b* satisfying P(N>*b*)=α/2 is comprised between 76 and 77, therefore, the 97.5% quantile is comprised in the interval [76-77]. For simplicity reasons, we only provide the value of *b* giving the closest value to α/2, in this case, 77.

REFERENCES

Gazey, W.J. & Staley, M.J. (1986) Population estimation from mark–recapture experiments using a sequential Bayes algorithm. *Ecology*, **67**, 941–951.

Petit E. and Valière N. (2006). Estimating population size with noninvasive Capture-Mark-Recapture Data. Conservation Biology, 20: 1062-1073.

Puechmaille S. and Petit E. (2007). Empirical evaluation of non-invasive capture-mark-recapture estimation of population size based on a single sampling session. Journal of Applied Ecology, 44: 843-852.