MCMP user guide and technical description

2013-05-28 PsN 3.6.2

Introduction

MCMP – Monte-Carlo Mapped Power – is a tool for power computations.

Method presented at PAGE2010:

Camille Vong, Martin Bergstrand, Mats O. Karlsson. Rapid sample size calculations for a defined likelihood ratio test-based power in mixed effects models.

Examples

mcmp -full_model=mod1.mod -stratify_on=DOSE -reduced_model=mod2.mod -seed=123

Input and options

mcmp-specific input

-simulation_model=sim.mod	The filename of the simulation model, optional. Default is the -full_model filename. If both -table_reduced and -table_full are specified, then option -simulation_model will be ignored. Cannot be used with option -simdata.
-full_model=full.mod	The filename of the full model is required, unless -table_full (see below) is used.
-reduced_model =red.mod	The name of the reduced model is required, unless -table_reduced (see below) is used.
-stratify_on=<column>	The name of the variable to stratify on, optional. Must be all uppercase and (unless NM7) at most 4 characters. Unless -table_strata is specified, the variable must be possible to request in $TABLE, i.e. either present in $INPUT or defined in $PK/ $PRED/$ERROR. PsN does not check that the variable is defined. If -reduced_model is specified then PsN will set $TABLE there, otherwise $TABLE will be set in the full model. The stratification table will not be generated from the simulation model. PsN will set $TABLE ID <stratify_on> FIRSTONLY NOAPPEND NOPRINT ONEHEADER FILE=strata.tab
-curve	Set by default. Can be disabled with -no-curve. This option controls whether the complete power curve up to the target power should be generated, or if the program should only compute the sample size required to achieve the target power.
-start_size=N	First total sample size (sum of samples from all strata) to test. Optional, default is 3*increment (see below).
-increment=N	Optional, default is the number of strata (1 if stratification is not used). Only consider setting this option if the stratification groups do not have equal size (the design is not balanced). Option

	-increment is the stepsize for the total sample size (sum of samples from all strata), i.e. the distance on the x-axis between the points in a total sample size vs. power plot. See section Discussion on strata sample sizes for a more detailed discussion of this option.
-max_size=N	The largest total sample size to plot for, optional, default equal to number of individuals in dataset.
-target_power=X	Default 98. If the whole power curve is generated (option -curve is set), the computations will stop when the computed power exceeds the target power 3 times in a row, even if max_size is not reached. If option curve is not set, then the program determines the sample size only for the target power.
-table_full=filename	Optional. The name of the table containing iofv for the full model. If this option is used, PsN will skip the estimation of the full model and read the table directly instead. For this option, iotab tables generated using PsN with option -iofv and NONMEM6 will work, as well as .phi-files generated with NONMEM7. Files generated in other ways must follow the following rules: There must be exactly one row per individual, all other lines (headers) must start with a non-number, the columns must be space or tab separated, and iofv must come in the last column.
-table_reduced=filename	Optional. The name of the table containing iofv for the reduced model. If this option is used, PsN will skip the estimation of the reduced model and read the table directly instead. The file must have the format defined above in the -table_full help text.
-simdata=filename	Optional. The name of a previously generated file with simulated data. Cannot be used together with option -simulation_model. If this option is set, no simulation will be performed by mcmp. Instead the file simdata will be used as the datafile when estimating the full and reduced model.
-table_strata=filename	Optional unless both -table_full and -table_reduced is used and stratify_on is set, then required. Table with stratification column. Must have only one row per individual, i.e. for example generated in NONMEM with FIRSTONLY (see option -stratify_on), and must have a header with ID and stratification variable name. Option -table_strata may be used even if neither of -table_full or -table_reduced are used, then the stratification column in -table_strata will be used instead of a table generated from the estimation of the full or reduced model.
-n_bootstrap=N	Optional, default 10000. The number of bootstrapped delta-ofv:s to generate for each total sample size.
-df=N	Optional, default 1. The number of degrees of freedom for the chi-square test. Allowed values are 1-30 in steps of 1 and 40-100 in steps of 10. PsN will compute the power for significance levels 5%, 1% and 0.1% for the number of degrees of freedom.
-significance_level	Optional, default 5 (percent). Permitted values are 5, 1 and 0.1. Convergence check will be based on critical ofv for this

	significance level.
-critical_ofv	Optional, no default. If specified, -critical_ofv will override the setting of -df. PsN will work with positive values internally (delta_ofv=reduced-full, check if delta_ofv > critical_ofv), but will automatically change the sign if the user gives a negative value.

Some general PsN-options which are useful in combination with mcmp

For a complete list of common options see common_options_defaults_versions.pdf, or psn_options -h on the command-line.

-directory=mcmp_dirX	The directory in which the script will run NONMEM can be named. The default name is "mcmp_dirX" where X is increased by 1 each time you run the script.
-seed=X	A seed for the random number generator can be specified. This makes the run reproducible. It acts as a starting point for the random number generator to produce the random number that will be used for the bootstrapping of the iofv:s.
-threads=X	The number of parallel processes to start for the model runs on a parallel computer. Setting this value to something larger than the number of models given as input (1or 2) will have no effect. The bootstrapping of the iofv:s is not parallelized.
-help	With -help mcmp will print a longer help message.

Output

The output from each mcmp is mcmp_results.csv which contains a table with header total_X,power at 5%,power at 1%,power at 0.1%, (headers for N from each strata). One row for each total sample size (total_X). The table can be plotted in excel. The results are also printed to screen as they are produced, and the user can terminate the run with Ctrl-C if the obtained power is deemed sufficient. The file mcmp_results.csv will contain the values that were computed before the interruption.

Recovering a crashed/stopped mcmp

If the simulation finished but none of the estimations finished, then start over in a new run directory but use option -simdata with the dataset from the simulation step. If one or two of the estimations finished, then start in new directory using option -table_reduced and/or -table_full and possibly -table_strata.

Discussion on strata sample sizes

The increment and start_size options may seem complicated, so here is a detailed background to the design of those options in PsN.

We start with some examples to explain a method which is *not* implemented in PsN. When the design is perfectly balanced, choosing the number of individuals to sample from each strata in each iteration is trivial. It is more complicated to define an algorithm that works in all cases with any design. Consider

the following three cases:

1) N_{total}=400 where N_A=200 and N_B=200

2) N_{total}=400 where N_A=100 and N_B=300

3) N_{total}=424 where N_A=233 and N_B=191

It is easy to see that the ideal sampling scheme in case 1 is to take 1 individual from each strata in the first iteration, 2 from each in the second iteration, and so on. Then the 1:1 balance is perfect in every iteration. The increment, the increase of the total sample size in each iteration, is 2 in this case.

Case 2 is also simple, take 1 individual from strata A and 3 from strata B in the first iteration, 2 from strata A and 6 from strata B in the second iteration and so on, thus always preserving the 1:3 relation of the group sizes.

Case 3 is more difficult and it is not obvious what the best strategy is.

The implicit strategy in for case 1 and 2 is the following algorithm:

i) Find the greatest common divisor D of the strata sizes.

ii) In each iteration, increase the sample size from strata A with N_A/D and from strata B with N_B/D

For case 1 D=200 and $N_A/D=N_B/D=1$ and for case 2 D=100, $N_A/D=1$ and $N_B/D=3$. For case 3 it turns out the greatest common divisor is 1, giving a useless strategy since it is not okay to sample 233+191 individuals in the first iteration. Hence finding the greatest common divisor is not a strategy which is suitable to implement in PsN.

PsN must be able to handle all cases, keep as good a balance as possible between the strata regardless of their original sizes, make small enough increases in the sample sizes to give a good power plot and allow the user to affect the sampling strategy as much as possible without making the input options too complicated.

PsN uses the following algorithm:

1) Compute the desired total sample size $X_{desired\_total}$=start_size+(i-1)*increment, where i, i=1,2,3... is the iteration number, increment has default equal the number of strata but can be set by the user, and start_size has default 3*increment but can be set by the user.

2) Compute the number of individuals X_i to sample from strata i using the following formula: X_i =round_to_nearest_integer($X_{desired\_total}$*N_i/N_{total}).

With the formula in step 2 strata i:s fraction of the total sample size will always be as close as possible to the fraction of strata i in the original population. The rounding is necessary since the division often has a non-zero remainder, and it is important to note that the actual total sample size X_{actual_total}, which is the sum of the individual sample sizes X_i, can differ slightly from $X_{desired\_total}$ because of the necessary rounding. If the increment is small it can happen that X_{actual_total} is the same in two consecutive iterations, without there being any error. PsN always reports X_{actual_total} in all output. $X_{desired\_total}$ is an internal variable and is never displayed.

If the user has a dataset as case 2 (N_{total}=400 where N_A=100 and N_B=300) the user can set increment to 4 (1+3) which would give a perfect relation between the strata sample sizes in each iteration, since according to PsN:s algorithm strata A will always get 100/400=1/4 of the samples and strata B 300/400=3/4 of the samples and $X_{desired\_total}$ would always be a multiple of 4. If the user leaves increment to the default = the number of strata, then the results would still be acceptable. In every other iteration

the relation would be perfect ($X_{desired\_total}$ is a multiple of 4). In the rest of the iterations the actual relation would deviate slightly, e.g. if $X_{desired\_total}$=10 then X_A= round(10*100/400)=3 and X_B= round(10*300/400)=8 giving X_{actual_total}=3+8=11, but the larger the total sample size is the smaller the deviation will be.

The user can also choose to set start_size to manipulate the sample sizes, but it is recommended not to set this option to anything other than a multiple of increment.

Technical overview of algorithm

1. If NONMEM6 is used, then the iofv option to PsN is set automatically. If NONMEM7 is used, no extra settings are needed.
2. PsN checks that the requirements on the options are fulfilled (see list of options above).
3. Look-up critical ofv if not given on command-line, then make sure sign is +.
4. Unless option simdata is given, or both table_reduced and table_full are given, simulate the simulation model with a random number seed in $SIM which set by PsN.
5. If -reduced_model is specified, PsN will add a $TABLE to the first $PROBLEM with ID <stratify_on> FIRSTONLY NOAPPEND ONEHEADER NOPRINT FILE=strata.tab. PsN does not check that it is possible to request <stratify_on> in $TABLE, so it is the responsibility of the user to either have it in $INPUT or define it in $PK/$PRED/$ERROR.
6. If -reduced_model is not specified but -full_model is, then PsN will add a $TABLE to the first $PROBLEM with ID <stratify_on> FIRSTONLY NOAPPEND ONEHEADER NOPRINT FILE=strata.tab.
7. In both the reduced and full model, set $DATA to the simulated data file.
8. Estimate the reduced model unless -table_reduced is specified.
9. Estimate the full model unless -table_full is specified.
10. Extract iofv-values from full and reduced iofv-tables, and subtract to create delta-iofv-vector.
11. Stratify delta-iofv-table based on strata.tab/-table_stratify.
12. Loop over total sample size starting on -start_size and incrementing with -increment in each step and continuing as long as total sample size does not exceed -max_size. For each total sample size and each strata, compute the number of samples to draw from this strata using the formula
    Z=round(total_sample_size*strata_individuals_in_full_data_set/individuals_in_full_data_set) . Draw Z samples from the strata, randomly with replacement. Repeat n_bootstrap times, sum delta_iofv for the samples from all strata to generate n_bootstrap delta_ofv:s. Compute percentage of delta_ofv > critical_ofv=power. In output table, write line with total_sample_size, power, sample_sizes from each strata. Halt if power exceeds target_power % three times in a row.