USER GUIDE FOR MENDELIAN

VERSION 3.0

MENDELIAN SAMPLING VARIANCE TEST

May 22, 2018

A.-M. Tyrisevä[1], W.F. Fikse[2], and M.H. Lidauer[1]

[1] Animal Genetics, Natural Resources Institute Finland (Luke), 31600 Jokioinen, Finland. Correspondence: anna-maria.tyriseva@luke.fi

[2] Växa Sverige, Box 288, S-75105 Uppsala, Sweden

Contents

1 Introduction

The purpose of this user guide is to give information on the format of the data and the instruction files, execution of the program as well as the related technical information. For more detailed information on the data edits and the validation method, see Appendix 8 and [12].

The program estimates within-year genetic variances and tests for a possible trend and outliers of the estimated variances. An empirical 95% confidence interval for the trend is obtained by bootstrapping 1000 samples with replacement within year classes and fitting a weighted regression model by using number of animals in the year classes as weights. The trend is expressed as a percentage relative to the genetic variance. Results from bootstrapping are also used to test for outliers among the within-year genetic variance estimates. Tolerated thresholds are fitted for both the trend and the outlier tests to detect only those cases that have a practical impact.

The outlier test in the program version 3.0 has changed compared to that in version 2.6 and in [12], see the Appendix 8.2.3. Further, the standard output has been extended: besides printing yearly Mendelian sampling (MS) means, the program prints also yearly MS means expressed in unit of average genetic standard deviation as well as the average MS mean across the years.

2 Getting started

2.1 Compilation

The Mendelian program is written in Fortran95. It is distributed as pre-compiled executable files only. It has been compiled as 64 bit versions with GNU (gfortran) and INTEL FORTRAN (ifort) compilers for Linux and as 32 and 64 bit versions with INTEL FORTRAN compiler for Windows. Also debugging versions are available. The latter are adviced to be used only for debugging purposes since they are notably slower than the optimized versions. If there is any needs for other versions, they will be provided.

- Optimized versions:
  - Linux, 64 bit ifort: `Mendelian3_0`
  - Linux, 64 bit gfortran: `Mendelian3_0.gnu`
  - Windows, 32 bit ifort: `Mendelian3_0_32.exe`
  - Windows, 64 bit ifort: `Mendelian3_0.exe`
- Versions for debugging:
  - Linux, 64 bit ifort: `Mendelian3_0.debug`
  - Linux, 64 bit gfortran: `Mendelian3_0-debug.gnu`
  - Windows, 32 bit ifort: `Mendelian3_0-debug_32.exe`
  - Windows, 64 bit ifort: `Mendelian3_0-debug.exe`

2.2 Installation

In Linux, unzip the file using the command:

```
unzip mendelian.zip
```

that will create a directory called `mendelian`. In Windows, the archive can be opended directly or, for example, using 7-zip program (`http://www.7-zip.org/`). The directory comprises the above mentioned executables, one example data set with the instruction and output files as well as R and SAS codes for plotting the results.

3 Data file

The data file is given in free format, all information for one animal is given on one line. The line consists of the following fields in the given order:

1. Identity for the animal a
2. Identity for the animal's sire s
3. Identity for the animal's dam s
4. Birth year of the animal
5. Estimated breeding value (EBV_{a_i}) of the animal and trait i , where i=1, N
6. EBV_{s_i} of the sire of the animal, where i=1, N
7. EBV_{d_i} of the dam of the animal, where i=1, N
8. Reliability of the animal's EBV ($r^2_{a_i}$), where i=1, N
9. $r^2_{s_i}$ of the sire's EBV, where i=1, N
10. $r^2_{d_i}$ of the dam's EBV, where i=1, N

- Serial number of the traits used later in the notes refers to i=1, N

Identities of the animals can be either character strings such as international identities or integers. `No spaces or forward slash signs (/) are allowed in the identities due to free format!` The program works incorrectly in this kind of situation. Maximum length of the identities is 30 characters. Animals with missing parental information can exist in the datafile, even though such animals will be excluded from the analyses. Code for missing parent is:

- Negative integer
- From one up to 30 zeroes
- International identity with zeroes after the breed-country-sex code, e.g. HOLCANM000000000000000

The birth year is expressed as a four-digit integer YYYY. The default time interval of the analysis is the last 12 years fullfilling the editing rules specified in the Appendix 8.2.5. Estimated breeding values and their reliabilites are coded as real values. A code for missing EBV must be -9998. or any larger negative real number. Reliabilities should be expressed between 0 to 1. A code for missing reliability must be zero or negative number. A maximum of 99 traits can be included in the datafile.

4 Instruction file

The instruction file comprises five rows giving the following information in the given order:

1. Name of the data file
   - e.g., cows.dat or /home/ejo39/2013/protein/red/cows.dat
   - Maximum length is 1024 characters
2. Number of traits in the data file
3. Space separated list of traits to be analyzed
   - If only some of the traits are analyzed, give their serial numbers
   - If all the traits are analyzed, you can give 0 instead of a sequence of 1,2,...,N
4. Space separated list of the *names* of the traits to be analyzed
   - Maximum length of the name is 15 characters
5. Space separated list of the most recent birth year included and how many years are analyzed.
   - The year is expressed as a four-digit integer YYYY
   - Default is 12 years
   - At least 8 years must be included
   - It is not necassary to define the number of years, if the default is used

Example:

Consider a case, where a datafile bulls.dat contains information on five traits (milk, protein, fat, scc, clinical mastitis) from 2000 to 2011. Given we would like to analyze the traits number two and five, the format of the instruction file is:

```
bulls.dat
5
2 5
protein clinmast
2011 12
```

Given we would like to analyze all the traits, the format of the instruction file looks like:

```
bulls.dat
```

```
5
0
milk protein fat scc clinmast
2011
```

Because the default 12 years are included in the analysis, only the most recent birth year was defined.

5 Execution of the program and generated output files

The program can be executed by typing the command prompt:

```
Mendelian3_0 < name.msv > name.log
```

The name of the program is naturally according to the choice of the version the user has intended to use. Most of the results will be printed on screen, therefore redirecting them to a file (specified here as name.log) is sensible. Three additional files will be created: *trname*.dat, *trname*.out, and *trname*.summary, where *trname* is the name of the analyzed trait specified by the instruction file. The file *trname*.dat contains animals in the analysis for *trname*, providing the following information for each: new integer id, birth year, MS term, PEV of the MS term, d, d × MS^2, d × PEV. For more information on the variables, see Appendix.

The file *trname*.out is intended to be used as an input file for R or SAS. On the first row, *trname*.out shows the result of the trend test. T refers either to the statistically non-significant result or to the significant result that is within the tolerated threshold, and F to the statistically significant result outside the tolerated threshold. The average genetic variance is shown on the second row. The next part comprises columns for the years in the analysis, size of the year classes, estimates of the within-year genetic variances, 95% empirical confidence intervals (CI) as well as the tolerated thresholds of the outlier test (order: lower and upper CI, lower and upper tolerated thresholds). The last column shows the final test results of the outlier test. Ok refers to the test result, where the average genetic variance either lies within the CI or is outside but within the tolerated threshold. Out refers to the test result that is outside the CI and exceeds the tolerated threshold.

The R and SAS codes needed for plotting yearly variances are provided in the package, giving all necessary information for a succesful execution. Line color is green, if the trend in genetic variance is statistically non-significant or significant but within the tolerated threshold of $\pm 2\%$, red otherwise. Statistically significant outlier years exceeding the tolerated thresholds have been marked with "out". The 95% empirical CI, tolerance thresholds as well as the average genetic variance line are plotted as well.

The file *trname*.summary is created for the Interbull Centre to easily create an overall summary of the results that countries have sent for the validation. A one row file consists of the following information:

- *Trname*
- Statistical significance of the trend test (NS refers to non-significant and SS to statistically significant test result)
- Magnitude of a trend in the genetic variance

- No of statistical outliers exceeding the tolerated threshold
- Total number of animals in the analysis
- Number of years included
- Number of year classes with records under 50% of the average year class
- Number of year classes with less than 10 records after excluding animals with too low a MS reliability value

6 Error and warning messages

The program gives an error message, if opening or reading of the instruction file or the data file fails or the given values are outside the boundaries, e.g. less than 8 years are included in the analysis or more than 99 traits are found in the datafile.

Warning messages are given, if a) in some birth year class the number of animals with observations is less than 50% of the average class size in the testing period, or b) after checking MS reliabilities of animals, number of records in some birth year class is below 10, or c) even below two in which case the program also terminates. In cases from a) to c), the program suggests changing the time window of the analysis or exclude the outermost year(s). Low heritability traits are recommended to be analyzed with bulls only.

7 Example

One data file with related instruction and output files is provided for the training purposes. The data contains one simulated trait.

8 Appendix: Validation of consistency of Mendelian sampling variance

A.-M. Tyrisevä[1], W. F. Fikse[2], E. A. Mäntysaari[1] J. Jakobsen[3], G. P. Aamand[4], J. Dürr[5], M. H. Lidauer[1]

[1]Animal Genetics, Natural Resources Institute Finland (Luke), Jokioinen, Finland
[2]Växa Sverige, Uppsala, Sweden
[3]Norwegian Association of Sheep and Goat Breeders, Norway
[4]NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark
[4]Council on Dairy Cattle Breeding, Bowie, MD, USA

8.1 Background

Reliable estimation of genetic merit of dairy bulls coming from different populations and production environments is fundamental to ensure unbiased international evaluations. Several studies have demonstrated that the biased genetic trends and heterogeneity in genetic variance in national evaluations affect also multi-trait across country evaluation (MACE). All countries participating in the international evaluations are required to validate their national evaluations for biased genetic trends, but homogeneity of variance across years has not been tested. However, under- or overestimation of genetic variance in some country affects the spread of breeding values on the other country scales, which can significantly affect the ranking of top bulls. National evaluation centers and Interbull therefore need a validation method to detect all significant heterogeneity in variance that impede reliable ranking of bulls in the international sire evaluation.

Based on Sullivan's original idea to calculate withing-year genetic variances [9], Fikse et al. proposed a modified method (IB4) to estimate within-year genetic variances and a statistical test [1, 2]. The procedure was tested on field data sets, but some inconclusive results were obtained and it was not implemented. Later, Lidauer et al. [3] developed a full model sampling method (FMS) to estimate within-year genetic variances. IB4 and FMS differ in the way the prediction error variance of Mendelian sampling is estimated, but give relatively similar results [3]. However, FMS requires simulation of new observations according to the model used in the national evaluation system. Therefore, it is not easy to implement in a scheme with a wide variety of national evaluation models.

A research project was set up to further study the IB4 method and to develop a suitable test statistics. The results have been presented in three papers [10–12]. Simulations were performed to study IB4 under different scenarios, and the effects of inbreeding, data size, quality and type of data (cows/bulls, magnitude of h^2, level of EBV and MS reliabilities) have been studied as well.

The new, proposed validation procedure is based on the original IB4 method, but a new statistical test, comprising tests for a trend and outliers, was developed. The new validation procedure has been tested with several field data sets. Based on the experiences obtained from these tests, the method was fine-tuned and is ready for use.

8.2 Validation procedure

8.2.1 Estimation of genetic variance

Within-year genetic variance $\sigma^2_{u_i}$ is estimated by utilizing Mendelian sampling and its prediction error variance, as shown by Sullivan [9]:

$$\sigma^2_{u_i} = \frac{\sum_{k=1}^{q_i} d_k \hat{m}_k^2}{q_i - \sum_{k=1}^{q_i} d_k PEV(\hat{m}_k)}, \tag{1}$$

where q_i is the number of animals in year i, d_k is the inverse of the proportion of genetic variance not explained by the known parents of animal k, $\hat{m}_k^2$ is the squared estimated Mendelian sampling deviation of animal k, and $PEV(\hat{m}_k)$ is the prediction error variance of Mendelian sampling deviation. As the exact values for $PEV(m_k)$ are difficult to obtain, Fikse et al. [1] proposed to approximate it following Misztal et al. [6, 7].

8.2.2 Statistical test for a trend

After obtaining within-year genetic variances for the most recent years fulfilling the data selection terms (defined below), the existence of a possible trend is tested by fitting a weighted linear regression of within-year genetic variances ($\hat{\sigma}^2_{u_i}$) on year x_i:

$$\hat{\sigma}^2_{u_i} = \alpha + \beta x_i + e_i, \tag{2}$$

where α and β are regression coefficients and e_i residual terms. The number of animals in each year class is used as the weight. The trend is expressed as a percentage relative to genetic variance $(\beta/\bar{\sigma}^2_u \times 100\%)$, where $\bar{\sigma}^2_u$ is the estimated genetic variance averaged over years. An empirical 95% confidence interval for the trend (β) is calculated by bootstrapping 1000 samples with replacement within year classes. For each bootstrap sample, the sampled animals are used to calculate yearly variances, after which the above regression model is fitted and the sample $\hat{\beta}$ is estimated. A 95% confidence interval is obtained by defining the 0.025 and 0.975 quantiles for the bootstrapped $\hat{\beta}$. If the confidence interval does not include zero, the trend is considered to deviate statistically significant from zero.

8.2.3 Statistical test for outliers

To detect years with possible outlier estimates of genetic variance that do not fit the linear trend model, and to find indications of a possible non-linear trend, an outlier test is applied. Within-year genetic variance estimates obtained from the 1000 bootstrap samples are used to detect outliers. An empirical 95% confidence interval for each within-year estimate of genetic variance is obtained by finding the 0.025 and 0.975 quantiles of the bootstrapped data. Further, a Bonferroni correction for the N independent tests is applied with N being the total number of years included in the analyses. If the confidence interval does not include the average genetic variance for some year, then the estimate of variance for this year is considered a statistical outlier.

8.2.4 Tolerated thresholds for the trend and the outlier tests

Field data sets used for testing can be very large, comprising hundreds of thousands of animals in a single year class. Hence, the statistical power is high to detect very small deviations from a zero trend in genetic variance. The same applies for outliers. Therefore, tolerance values are needed both for the trend and the outlier tests to enable detecting only those cases that have a real practical impact.

Based on the simulations presented by Tyrisevä et al. [11], an estimated linear trend, which lies within $\pm 2\%$ of the average estimated genetic variance, is suggested as a limit for acceptance.

For testing single outliers, a tolerated interval of $\hat{\sigma}^2_{u_i} \pm 0.10\bar{\sigma}^2_u$ of the average estimated genetic variance is suggested. This view is based on considering a 5% hypothetical standard error for the estimated variances and defining estimates that deviate more than two times the standard error (i.e., 10%) as failed outliers. The equality corresponds roughly to variances estimated from 800 observations (an approximated estimate of the standard error of $\hat{\sigma}^2_{u_i}$ is $\sqrt{2/(n-1)}\bar{\sigma}^2_u$).

8.2.5 Data editing rules for the validation

To obtain comparable results, a default time period of 12 most recent birth year classes should be covered. In each birth year class of the period, the number of animals with observations must be at least 50% of the size of the average birth year class in the testing period.

The test can be performed either for bulls or cows. For bulls, the same data edits are applied as outlined in the Interbull's code of practice, item 5.1.3 [8]. For cows, no specific data edits are needed. Animal's birth year, EBVs for the animals and their parents, as well as the estimates of the EBV reliabilities are required. The approximated reliabilities should be of high quality. Only animals with complete parental information are included in the analyses. Further, only animals with the MS reliability higher than 0.1 will be considered by the program. The limit was set to avoid biased, inflated estimates due to numerical instability when approximated MS reliabilities are close to zero [12]. For the same reason, evaluations for low heritability traits such as clinical mastitis are recommended to be validated with bull data only.

Tyrisevä et al. [12] showed that inbreeding can have an effect on the estimates of the within-year genetic variances. However, it should be accounted for both in the estimation of breeding values, approximation of reliabilities and in the estimation of genetic variances. However, the effect of not accounting for inbreeding was found to be tolerable (i.e. smaller than the 2% threshold for trend in genetic variance). Given that inbreeding is not considered in many national evaluation models, it is proposed that inbreeding is not accounted for in the estimation of genetic variances either.

9 Acknowledgements

Authors thank Prof. Ignacy Misztal for providing some modules for the program [4, 5].

References

[1] WF Fikse, L Klei, Z Liu, and PG Sullivan. Procedure for validation of trends in genetic variance. *Interbull Bull*, 31:30–36, 2003.

[2] WF Fikse, Z Liu, and PG Sullivan. Tolerance values for validation of trends in genetic variances over time. *Interbull Bull*, 33:200–203, 2005.

[3] M Lidauer, K Vuori, I Strandén, and EA Mäntysaari. Experiences with interbull test iv: estimation of genetic variance. *Interbull Bull*, 37:69–72, 2007.

[4] I Misztal. Blupf90 family of programs. http://nce.ads.uga.edu/wiki/doku.php. Accessed January 24, 2017.

[5] I Misztal. Programs. http://nce.ads.uga.edu/~ignacy/programs.html. Accessed January 24, 2017.

[6] I Misztal, TJ Lawlor, TH Short, and GR Wiggans. Continuous genetic evaluation of holsteins for type. *J Dairy Sci*, 74:2001–2009, 1991.

[7] I Misztal and GR Wiggans. Approximation of prediction error variance in large-scale animal models. *J Dairy Sci*, 71(Suppl.2):27–32, 1988.

[8] International Bull Evaluation Service. Code of practice. http://www.interbull.org/ib/codeofpractice. Accessed January 24, 2017.

[9] PG Sullivan. Reml estimation of heterogeneous sire (co)variances for mace. *Interbull Bull*, 22:146–148, 1999.

[10] A-M Tyrisevä, EA Mäntysaari, F Fikse, and MH Lidauer. Simulation study on Mendelian sampling variance tests. *Interbull Bull*, 44:57–61, 2011.

[11] A-M Tyrisevä, EA Mäntysaari, J Jakobsen, GP Aamand, J Dürr, WF Fikse, and MH Lidauer. Validation of consistency of Mendelian sampling variance in national evaluation models. *Interbull Bull*, 46:97–102, 2012.

[12] A-M Tyrisevä, EA Mäntysaari, J Jakobsen, GP Aamand, J Dürr, WF Fikse, and MH Lidauer. Validation of consistency of Mendelian sampling variance. *J Dairy Sci*, 101:2187–2198, 2018.