USER GUIDE

FOR

MENDELIAN SAMPLING VARIANCE TEST

VERSION 2.5

December 9, 2013

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

[1] Biotechnology and Food Research, Biometrical Genetics, MTT Agrifood Research Finland, 31600 Jokioinen, Finland. Correspondence: anna-maria.tyriseva@mtt.fi

[2] Department of Animal Breeding and Genetics, SLU, Box 7023, S-75007 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.

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 data with 1000 case resampling within year classes and fitting a weighted regression model by using number of animals in the year classes as weights. The trend is calculated as a percentage change in genetic variance. The estimated residuals from the bootstrapped data are used to study possible outliers that do not fit the model. A biological significance threshold is fitted for both the trend and the outlier tests. The test is considered failed, if the trend or the outlier test, or both of them, fail the statistical test and exceed the biological significance threshold.

2 Getting started

2.1 Compilation

The 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: `Mendelian2_5`
  - Linux, 64 bit gfortran: `Mendelian2_5.gnu`
  - Windows, 32 bit ifort: `Mendelian2_5_32.exe`
  - Windows, 64 bit ifort: `Mendelian2_5.exe`
- Versions for debugging:
  - Linux, 64 bit ifort: `Mendelian2_5.debug`
  - Linux, 64 bit gfortran: `Mendelian2_5-debug.gnu`
  - Windows, 32 bit ifort: `Mendelian2_5-debug_32.exe`
  - Windows, 64 bit ifort: `Mendelian2_5-debug.exe`

2.2 Installation

Unzip the file in Linux 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 of 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 given on one line. The line consists of the following fields in the given order:

1. Animal identity
2. Sire identity
3. Dam identity
4. Birth year of the animal
5. Estimated breeding value ($\mathrm{EBV}_{\mathrm{a}_\mathrm{i}}$) of the animal, where i=1, N
6. $\mathrm{EBV}_{\mathrm{s}_\mathrm{i}}$ of the sire, where i=1, N
7. $\mathrm{EBV}_{\mathrm{d}_\mathrm{i}}$ of the dam, where i=1, N
8. Reliability of the animal's EBV ($\mathrm{r}^2_{\mathrm{a}_\mathrm{i}}$), where i=1, N
9. $\mathrm{r}^2_{\mathrm{s}_\mathrm{i}}$ of the sire's EBV, where i=1, N
10. $\mathrm{r}^2_{\mathrm{d}_\mathrm{i}}$ 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 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 rules specified in the Appendix 7.2.6. Estimated breeding values and their reliabilites are coded as real values. A code for missing EBV must be -9998. or any smaller value. Reliabilities should be expressed between 0 to 1. A code for missing reliability must be zero or any smaller value. A maximum of 99 traits can be included in the datafile.

4 Instruction file

The instruction file comprises of 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
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
   - 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 info on five traits (milk, protein, fat, scc, clinical mastitis) from 2000 to 2011. Given we would like to analyze the trait 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:

```
Mendelian2_5 < 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. Two additional files will be created: *trname*.dat and *trname*.out, 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 $\times$ MS^2, d $\times$ 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 after fitting a biological significance threshold. T refers to the passed and F to the failed trend test. The next part comprises of four columns, showing years in the analysis, size of the year classes, estimates of the genetic variances as well as the result of the outlier test after fitting a biological significance threshold. The R and SAS codes needed for plotting yearly variances is provided in the package, giving all necessary information for a succesful execution. Line color is green for passed tests, red otherwise. Years that are both statistical and biological outliers have been marked with "out".

6 Example

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

7 Appendix: Validation of consistency of Mendelian sampling variance

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

[1]MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics, Jokioinen, Finland
[2]Interbull Centre, Department of Animal Breeding and Genetics, SLU, Uppsala, Sweden
[3]NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark
[4]Dept. Animal Breeding and Genetics, SLU, Uppsala, Sweden

7.1 Background

Sullivan [6] derived an equation to calculate within-year genetic variances. Since then, there have been several reports of detected trends in within-year genetic variances. Under- or overestimation of genetic variance in some country affects the spread of breeding values on 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 trends that impede reliable ranking of bulls in the international sire evaluation.

Based on Sullivan's idea, 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 many countries/traits failed the test 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 they estimate the prediction error variance of MS deviations, but give relatively similar results [3]. No test statistics was developed for the FMS.

A research project was set up to further study both the methods and to develop suitable test statistics. The results have been presented in two papers [7,8]. Simulations were performed to study IB4 and FMS 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.

Based on the experiences obtained so far, FMS has been found to be robust, whereas IB4 slightly sensitive to the quality of the data. However, FMS requires simulation of new observations according to the model used in the national evaluation system, and it is followed by the national evaluation using the simulated observations. Therefore, it is not easily implemented in a scheme with wide varieties of national evaluation models. The new, proposed validation procedure is based on the original IB4 method, but a new statistical test, comprising of tests for a trend and outliers, was developed. Simulations were performed to quantify the effect of a biased mean or a biased variance on true and estimated breeding values, to define the acceptable level of bias in trend. The new validation procedure has been tested with field data sets from the initial pilot study, and new data sets provided by the Nordic Cattle Genetic Evaluation (NAV). Based on the experiences obtained from these tests, the method was fine-tuned and is ready for a new pilot study.

7.2 Validation procedure

7.2.1 Estimation of genetic variance

Within-year genetic variance $\sigma^2_{u_i}$ is estimated according to Fikse et al. [2]:

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

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 the Mendelian sampling deviation approximated according to Fikse et al. [1].

7.2.2 Statistical test for a trend

After obtaining within-year genetic variances for the most recent years fulfilling the data selection terms (defined below), an existence of a possible trend is tested. The trend is defined as a percentage change in genetic variance $(\beta/\bar{\sigma}^2_u \times 100\%)$, where $\bar{\sigma}^2_u$ is the estimated average genetic variance, and β is a regression coefficient from a weighted regression model for $\sigma^2_{u_i}$ with number of animals in the year classes used as weights:

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

in which α and β are the regression coefficients, x_i N years and e_i residual terms. An empirical 95% confidence interval for a trend (β) is calculated by bootstrapping data with 1000 case re-sampling within year classes. For each bootstrapped sample the above regression model is fitted and the trend is calculated. A 95% confidence interval is obtained by defining 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.

7.2.3 Statistical test for outliers

Residual terms obtained from the regression model fitted for the 1000 bootstrapped samples are used to detect possible outlier years that do not fit the model. A 95% confidence interval is obtained by defining 0.025 and 0.975 quantiles for the bootstrapped residuals. Further, a Bonferroni correction for the N independent tests is applied. If the confidence interval does not include zero for some year, an estimate of variance for this year is considered to be a statistical outlier.

7.2.4 Biological significance thresholds

Field data sets used for testing can be very large, comprising of hundreds of thousands of animals in a single year class, for example data sets for Holstein cows. This kind of data has power to detect the tiniest deviations from a zero trend as statistically significant, without any practical impact. In addition, a small deviation between a within-year genetic variance estimate and its predicted value from the linear regression will result in a statistically significant outlier, even if the regression line was in practice a straight line with observations fitted in its very vicinity. Therefore, a biological significance threshold is needed both for a trend and outliers to detect only those cases that have a real practical impact.

Based on the simulations presented by Tyrisevä et al. [8], an estimated linear trend, which lies within $\pm 2\%$ of the average estimated genetic variance, is suggested as a limit for acceptance. Further, when the original IB4 method was developed, Fikse performed simulations that also supported using a threshold of 2%.

For testing the outliers, a biological acceptance interval of $\bar{\sigma}_u^2 \pm 0.10\bar{\sigma}_u^2$ of the average estimated genetic variance is suggested. The motivation for this is based on considering a 5% hypothetical standard error for the estimated variances and on defining estimates that deviate more than two times the standard error as biological outliers (i.e. 10%). The equality corresponds roughly to the variances estimated from 600 observations ($\sqrt{2/(n-1)}$). The threshold was found to be sensible, based on all available field data that has been tested so far.

7.2.5 Definition of the failed test

We suggest that the validation test is considered failed, when the trend or the outlier test, or both of them, fail the statistical test $\mathbf{and}$ exceed the biological significance thresholds specified above.

7.2.6 Data used for testing

As a default, a time period of 12 most recent birth year classes should be covered. In each birth year class of that period, the number of animals with observations must be at least 50% of the average yearly size of the animals 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. 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 quality of the approximated reliabilities should be considered. Only animals with complete parental information are included in the analyses. Further, only animals with 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; a characteristic that was not detected when analyzed under the more robust FMS method [8]. 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. [8] 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), and 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.

So far, the test has been developed for animal models only, but is in principle applicable for sire models too.

7.3 Experiences from the initial pilots and NAV data sets

The new validation procedure has already been tested in-depth. The countries of the initial pilot study that were tested again were: Australia (AUS), Canada (CAN), Germany (DEU), France (FRA), Italy (ITA) and the Netherlands (NLD). They provided data sets for milk, protein and fat yields in Holstein. FRA sent data for Montbéliard also. FRA and ITA sent cow data sets and AUS and DEU bull data sets edited in accordance with the editing rules by Interbull. CAN and NLD sent both cow and bull data sets. In these countries, bulls of the data sets consisted of both domestic bulls and bulls used in the international evaluation. The initial pilot covered primarily years from the late 1980's to the turn of the 21st century. Based on the experiences, the new validation procedure works well.

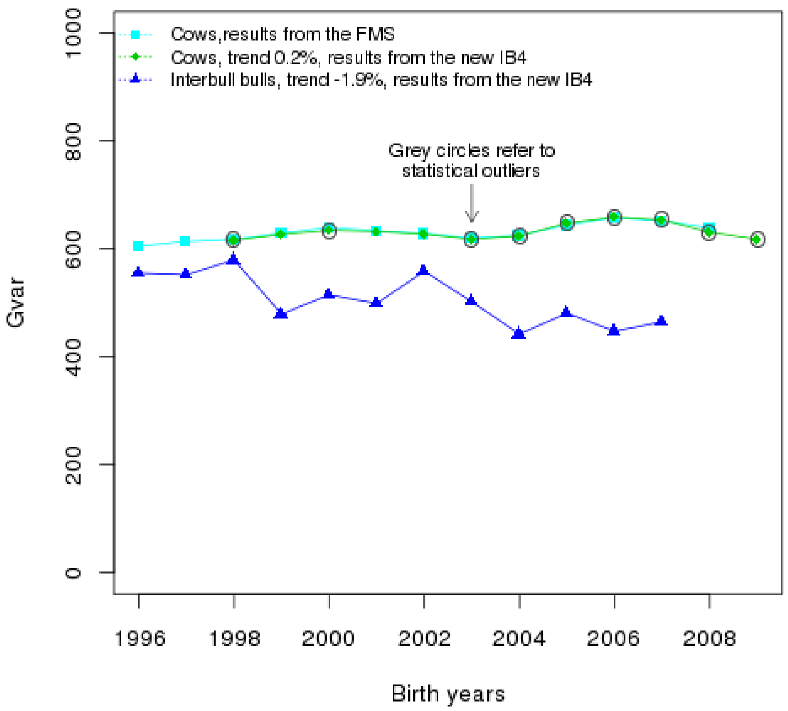

Figure 1: Within-year genetic variances of the combined three-lactation protein yield in Nordic Holstein cows and bulls. Results for the cows are obtained both from the new validation procedure and from the FMS method. Trend was statistically significant in both bulls and cows, but it did not reach a biological significance threshold of 2%. Statistical outliers in cows were within the biological outlier limits. Thus, both bulls and cows passed the validation test.

NAV provided data for a second pilot. They consisted of cow and bull data sets for Holstein, RDC and Jersey, covering birth year classes from 1998 to 2009 for cows and from 1996 to 2007 for bulls. Traits analyzed were protein yield, somatic cell count and clinical mastitis. NAV kindly gave permission to distribute some of the test results. Figure 1 shows within-year genetic variances for combined three-lactation protein yield in Holstein. The average class size in cows was over 200 000 animals, whereas in bulls it was 480, thus giving a good example of the behavior of two different kind of data sets. For both cows and bulls, there was a statistically significant trend, but with different magnitude. The trend in cows was very small, only 0.2%, whereas that in bulls was close to the limit of the biological significance threshold, being -1.9%. There were no statistical outliers in the bull data set, but due to very large class size, nine of the years were considered statistical outliers in cows. After considering the biological significance threshold, no outliers existed in cows either. As a final test result, both bulls and cows should pass the test. For cows, estimates of within-year genetic variances from the FMS method were also available. Data used in the FMS analysis covered two extra years and somewhat more animals, but the results were in a very good agreement between the methods.

8 Acknowledgements

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

References

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

[2] W F Fikse, Z Liu, and P G 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 E A 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.

[5] I. Misztal. Programs. http://nce.ads.uga.edu/~ignacy/programs.html.

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

[7] A-M Tyrisevä, E A Mäntysaari, F Fikse, and M H Lidauer. Simulation study on mendelian sampling variance tests. *Interbull Bull*, 44:57–61, 2011.

[8] A-M Tyrisevä, E A Mäntysaari, J Jakobsen, G P Aamand, J Dürr, W F Fikse, and M H Lidauer. Validation of consistency of mendelian sampling variance in national evaluation models. *Interbull Bull*, 46:97–102, 2012.