EAAP 68th ANNUAL MEETING Tallinn, Estonia 2017

«Development of milk data control system for reducing the bias of genomic evaluation for the Russian Holstein breed»

Rukin I, Pantiukh K, Rysina M, Grouzdev D

I Gene LLC, MSU

Knyazeva M, Sheglov M, ARSRIB

Introduction

Project: national GEBV evaluation system
Performers: Moscow State University, All-Russian Scientific Research Institute of Breeding and I Gene company

Model: ***MT ssGBLUP AM***

Traits: Milk production traits - Milk (kg), fat and protein(%, kg)
Breed: Holstein (75% and more by pedigree and genome)

Extremely important factors:

1. High-quality milk records data;
2. General pedigree information;
3. Presence of genomic information.

Introduction

Not all herds in Russia use ICAR certified methods
of milk records collection.
It’s necessary to **control TD milk data.**

<u>Milk data control system:</u>

1. Outlier control;
2. Gestation length control;
3. Herd data variability control;
4. Amount of TD per lactation control;
5. Herd reliability control;
6. Lactation reliability control.

Raw data	
Category	**Number**
Animals	**2 438 733**
Lactations	**6 517 123**
Herds	**1 057**
Regions	**44**
TD milk yield	**74 738 833**
TD milk fat	**68 545 716**
TD milk protein	**49 023 002**

1. Outlier control

Remove rough errors
(incorrected records, rough errors during data transfer etc.)

Trait	Confidence interval
TD milk yield, L	From 1 to 50
TD milk fat, %	From 2,5 to 6
TD milk protein, %	From 2,5 to 6
Lactation number	From 1 to 10

Results (milk protein as example)

Trait	Before control	Deleted data	Deleted data, %
TD milk protein	49 023 002	428 031	**0,87**

2. Gestation length control

Control of calving date

For each lactation:

1. Gestation length (GL) count (calving date – mating date)
2. GL analysis (remove lactation data if GL != 280±20 days)

Results (milk protein as example)

Trait	Before control	Deleted data	Deleted data, %
Lactation, number	6 503 012	190 934	**2,93**
TD milk protein	48 594 971	1 276 168	**2,62**

3. Herd data variability control

Excluding herds, which "copy-past" data

For each herd and each trait (3174 groups):

1. Data grouping (N subgroups:

same TD date, same TD week, same TD month)

2. Variability analysis in each subgroup:

if var(subgroup) for >50% data = 0 - remove all **herd** data

If var(subgroup) for >40% data = 0 - remove all **trait** data

Results (milk protein as example)

Trait	Before control	Deleted data	Deleted data, %
Herd	1 058	54	**5,1**
Herd (only protein)	1 004	22	**2**
TD milk protein	48 594 971	1 961 036	**4,14**

4. Amount of TD per lactation control

Exclude lactations, that have not enough information

For each lactation:

1. DIM amount analyzing:

If amount of DIM (milk yield) < 5 – remove all data

If amount of DIM (fat or protein) < 5 - remove trait data

Results (milk protein as example)

Trait	Before control	Deleted data	Deleted data, %
Lactation, number	6 127 678	678 064	**11,06**
TD milk protein	45 357 767	1 315 697	**2,9**

5. Herd reliability control

Two-step approach with in-lactation and in-herd analysis

Mean absolute error (MAE) calculation for each lactation
Traits: milk yield, milk fat, milk protein

↓

MAE distribution analysis for each herd
Traits: milk fat, milk protein

5.1 MAE calculation

<u>Good example (on one lactation, protein, %):</u>

1. 9 DIM in lactation;
2. 305d lactation curve calculation (Wilmink et al.);
3. Mean absolute error (M) calculation:

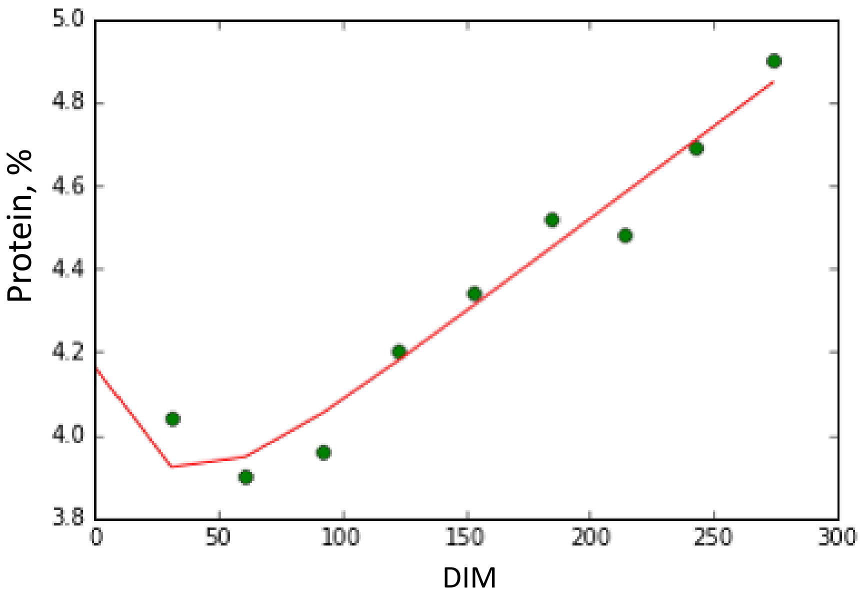
Count the $\mathrm{M} = \sum_{t=1}^{9} \left| \frac{A(t) - F(t)}{A(t)} \right|$, where:

$A(t)$ – actual DIM;

$F(t)$ – forecast DIM.

M = **0,025**

Normal lactation (protein, %)



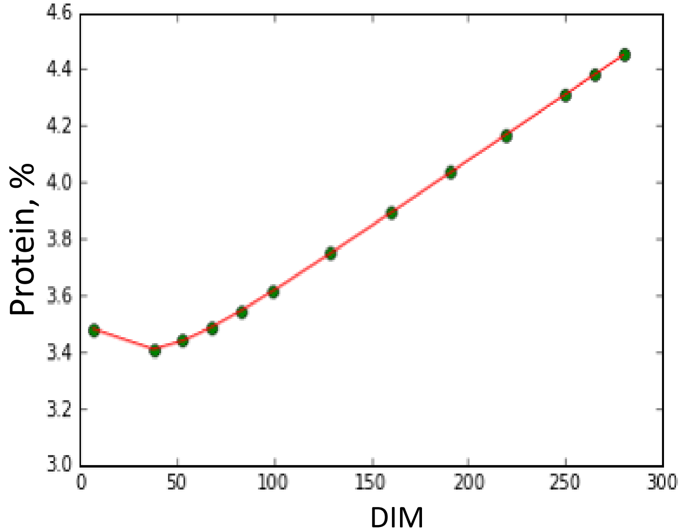
5.1 MAE calculation

Bad example №1:

$\mathrm{M} = \mathbf{0}$

Possible reasons:

DIM records were **generated artificially**, same $\boldsymbol{M}$ for **copy-past** of one value for all DIM in lactation

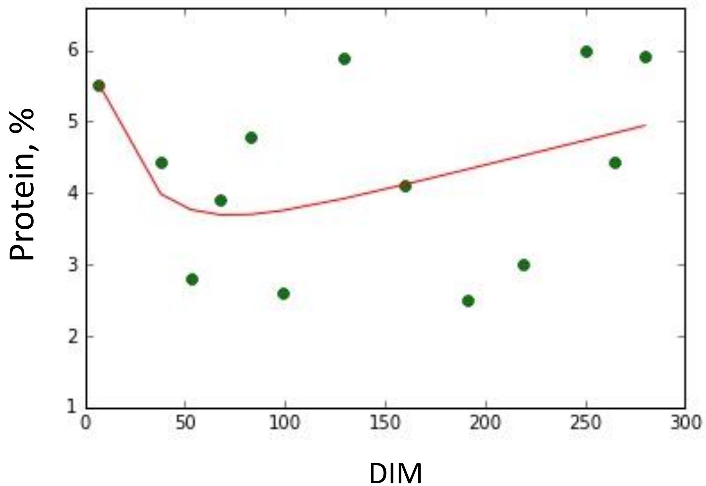


Bad example №2:

$\mathrm{M} = \mathbf{0{,}248}$

Possible reasons:

Errors in data transfer, when TD data go from milk lab to farmer and than to N-GES center



5.2 MAE distribution analysis

For each lactation: M(MAE) calculation

MAE distribution analysis in each herd

Theory: Lactations mean absolute error distribution for each herd must have properties of normal distribution.

Aim: Compare the mean M-values distribution of each herd with reference normal distribution

Method: **2-sample Kolmogorov–Smirnov test (K-S test)** (scipy.stats.ks_2samp)

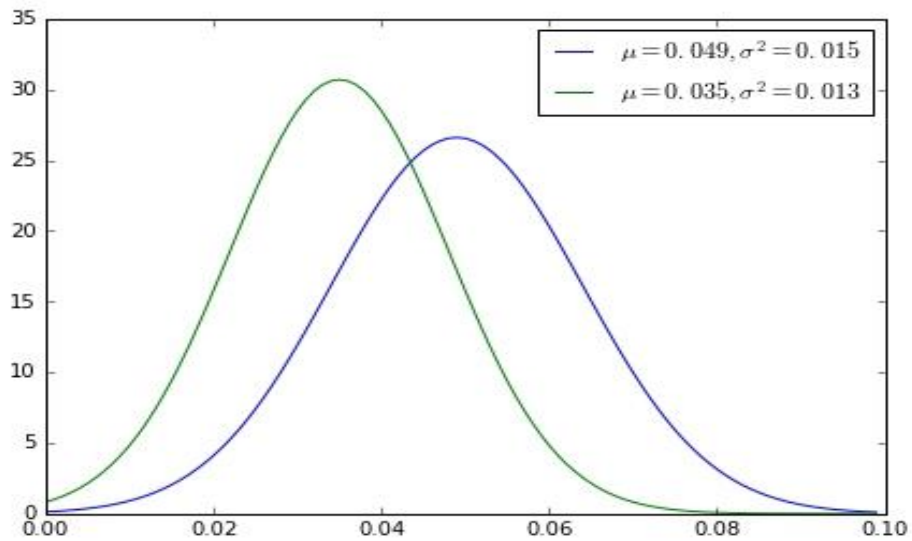
5.2 Reference normal distribution

Data: high-quality TD records from Leningrad region, Russia

Herds: **59** Lactations: **349 931**

Traits: milk fat, % and milk protein,%

Trait	Reference normal distribution
Milk fat, %	$\boldsymbol{N(0,049;\ 0,000225)}$
Milk protein, %	$\boldsymbol{N(0,035;\ 0,000169)}$



5.2 MAE distribution analysis

Exclude herds, which copy-past data in one lactation

For each herd and 2 traits (1964 groups):

1. MAE calculation for each lactation in herd;
2. K-S test for all lactations MAE distribution in herd:

   count D for each herd

   ($0 \leq D \leq 1$, where:

   0 – distribution is fully coincides with reference

   1 – distribution is fully not coincides with reference)

If D for milk fat is >0,91 – remove all milk fat data of herd

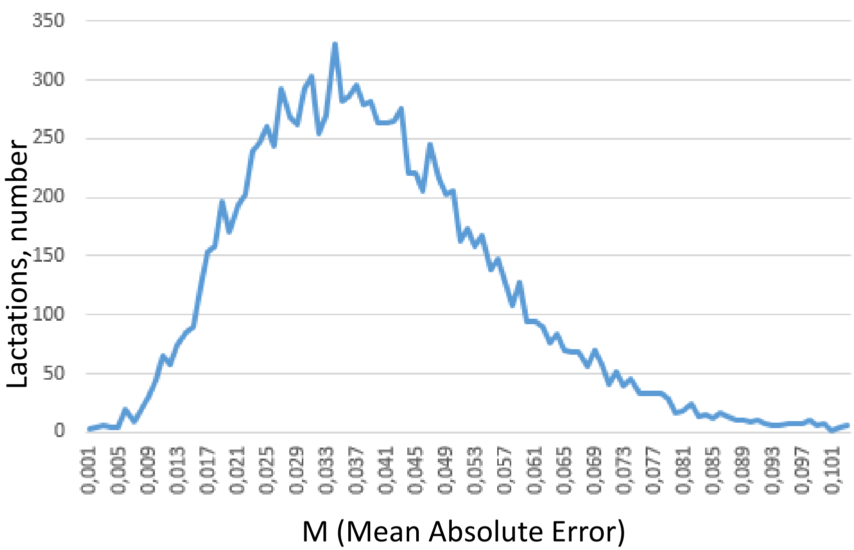
If D for milk protein is >0,8 – remove all milk protein data of herd

5.2 MAE distribution analysis

Results (milk protein)

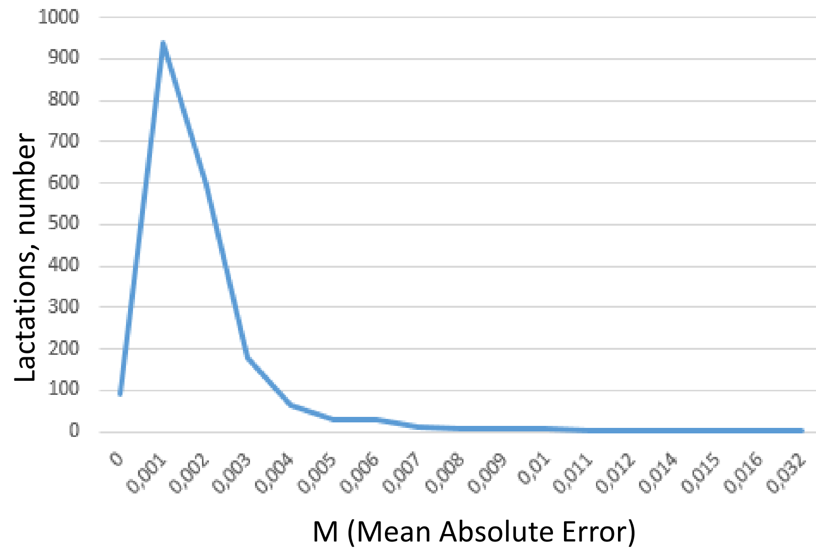
Herd 1 (**D = 0,01**)

Protein data didn't remove



Herd 2 (**D = 0,96**)

All protein data removed



Trait	Before analyze	Deleted data	Deleted data, %
TD protein	44 042 070	3 748 279	**8,51**

6. Lactation reliability control

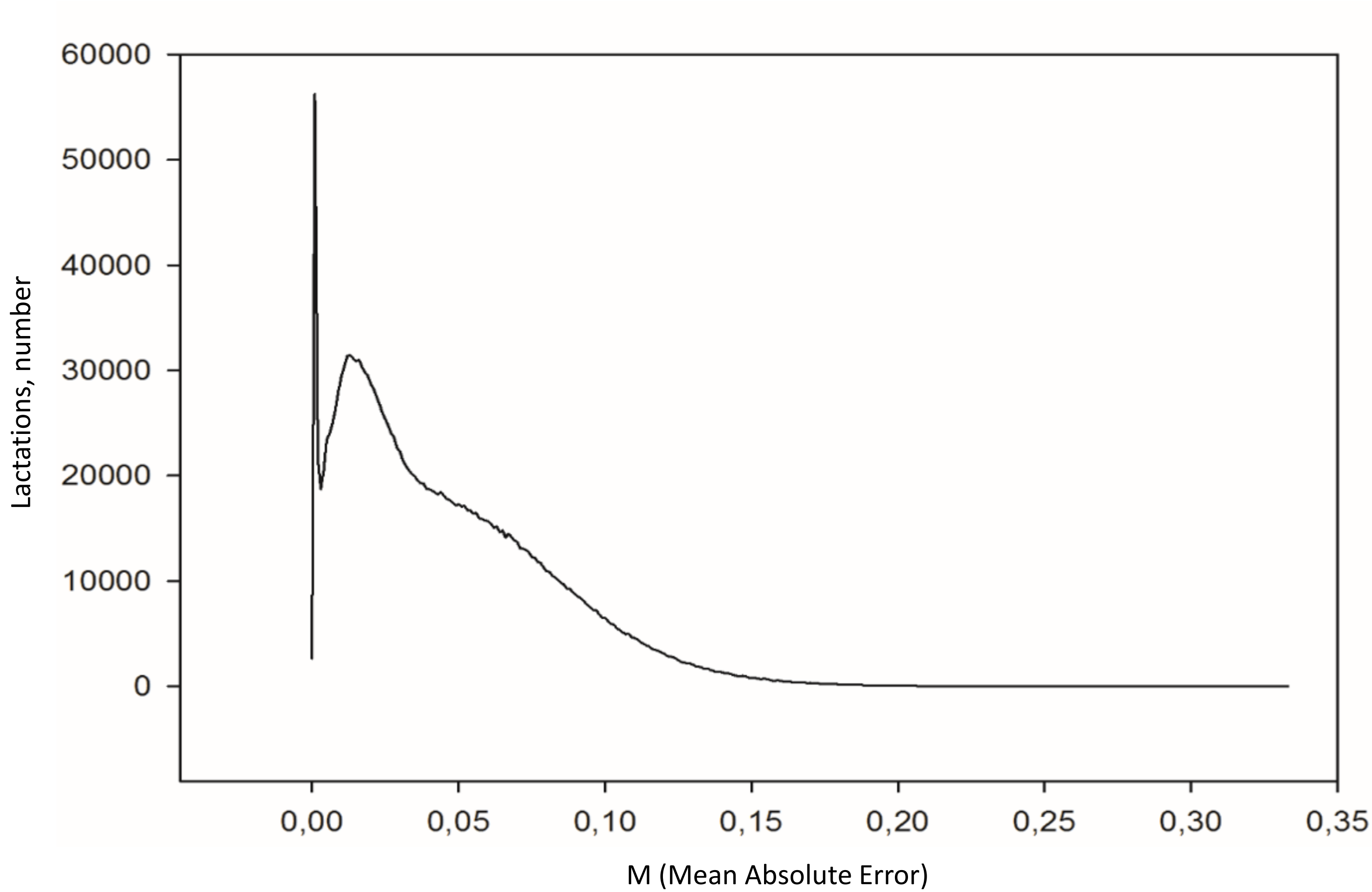
For each lactation: $\mathrm{M(MAE)}$ calculation. **DONE!**

For each lactation after control №5:
remove all outliers according to M-value

M-value for trait	Confidence interval
305d milk yield, L	From 0,011 to 2,19
305d milk fat, %	From 0,004 to 0,17
305d milk protein, %	От 0,004 до 0,17

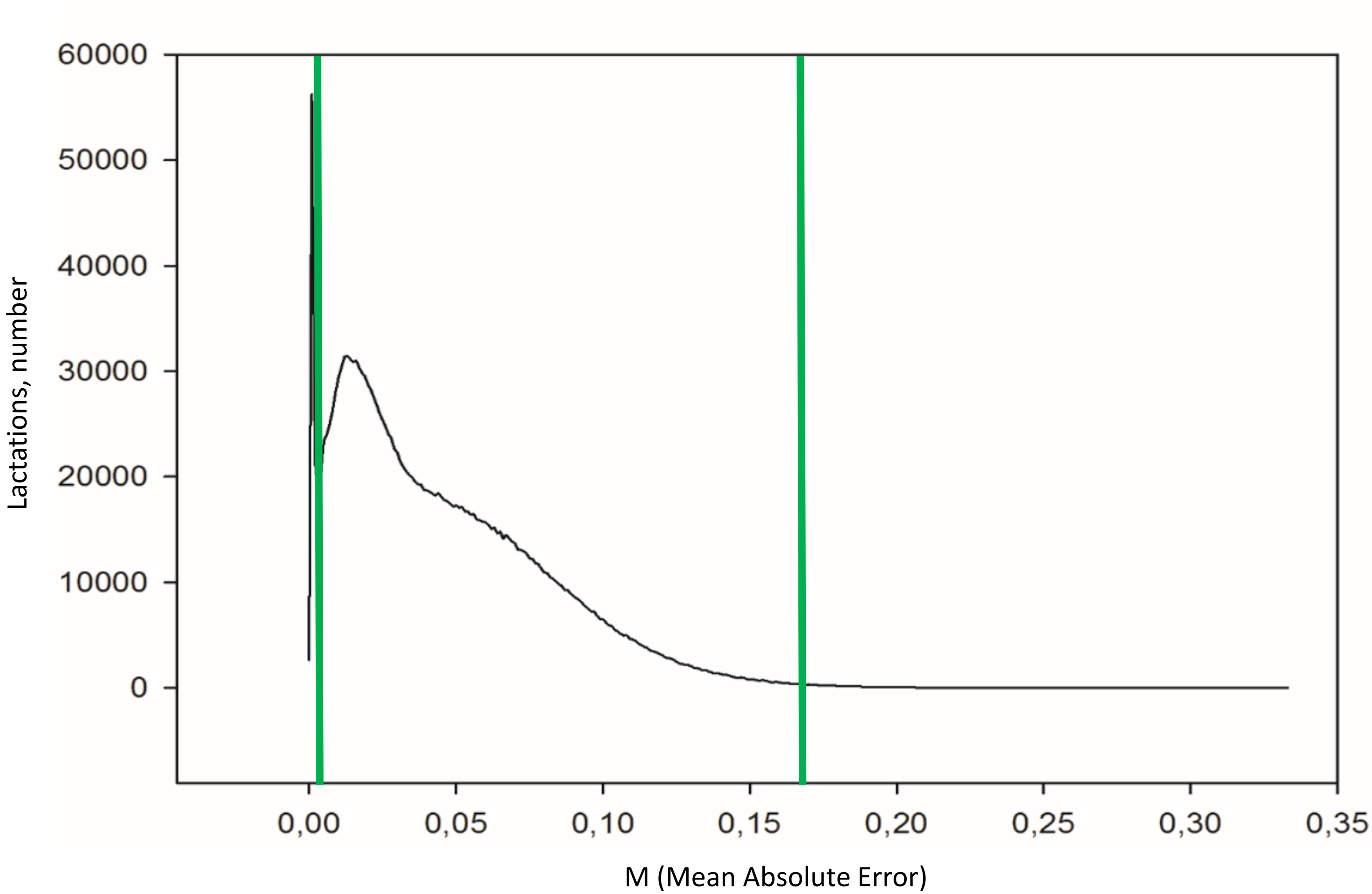
6. Lactation reliability control

Lactations M-distribution before control (protein,%)



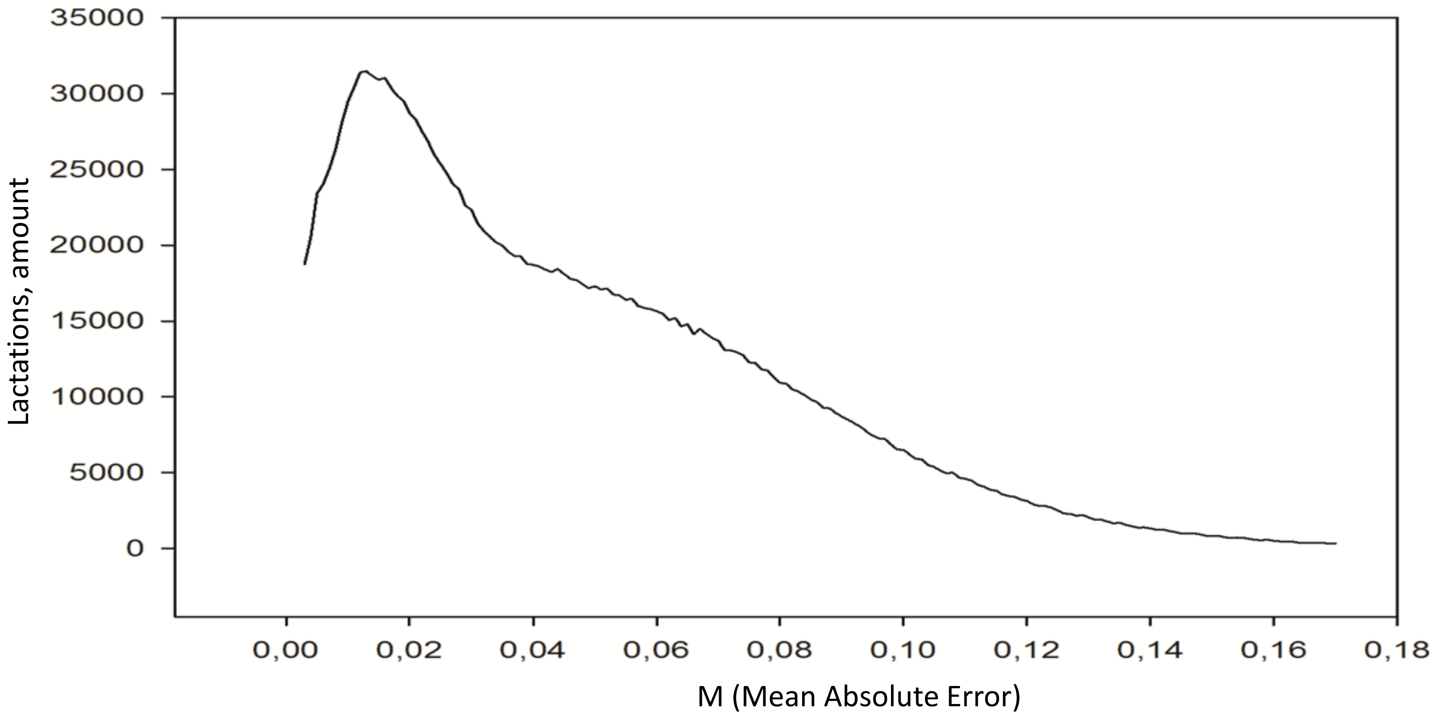
6. Lactation reliability control

Lactations M-distribution before control (protein,%)



6. Lactation reliability control

Lactations M-distribution after control (protein,%)



Trait	Before 6 control	Deleted data	Deleted data, %
TD protein	40 293 791	1 721 507	**4,27**

Results

In average, **20,37% of milk data** was removed.
High-quality milk production database from 44 regions was created for estimation of breeding values.

Results for each control step, in %

Category	Raw data	№1	№2	№3	№4	№5	№6	Total remove, %	Final data
Animals	**2 438 733**	**0,22**	**2,94**	**2,92**	**11,07**	**5,88**	**0,63**	**19,20**	**2 001 385**
Lactations	**6 517 123**	**0,22**	**2,94**	**2,92**	**11,07**	**7,76**	**1,07**	**25,97**	**5 021 595**
Herds	**1 057**	**-**	**-**	**5,1**	**-**	**-**	**-**	**5,1**	**1 003**
TD milk yield	74 738 833	0,49	2,11	2,79	1,67	-	**14,42**	21,48	**59 532 813**
TD milk fat	68 545 716	**0,88**	2,17	4,00	2,00	3,87	3,40	16,32	**58 067 441**
TD milk protein	49 023 002	0,87	**2,63**	**4,14**	**2,90**	**8,51**	4,27	**23,33**	**38 572 284**

Glad to answer any questions

Moscow State University, I Gene LLC
<u>Contacts:</u> Rukin Ilya
E-mail: rukin@i-gene.ru
Telephone: +7-926-710-01-52

Thank you for attention