THE PROGRAM "Rabbit"

Agustín Blasco

Animal Science and Technology Institute. Universitat Politècnica de València.

Apartado 22012. Valencia 46071. España

ablasco@dca.upv.es

ABOUT THE AUTHORS

The programs where originally written by the Egyptian geneticist **Wagdy Mekkawy**. They were later modified to make them friendlier by the Cuban geneticist **Dianelys González-Peña**. Professor **Agustin Blasco** designed the process and took part in programming them, all suggestions and reports of errors found when working with the programs should be addressed to him.

INTRODUCTION

The use of Bayesian statistics has increased spectacularly in all scientific fields since the numeric problems that prevented its used were solved at the end of the 90's. Its popularity has been due in some cases to the possibility of integrating prior information in the experiments analyzed and in some other cases to its ability for solving complex problems; for example, determining the uncertainty in nested models, or the comparisons between non-nested models, for which classical statistics has no even approximate solutions.

Classical statistics is still used in simple problems like mean comparisons between treatments. In these cases both Bayesian and classical statistics give similar results, but the description of the uncertainty is different. Bayesian statistics gives a new approach to the description of the uncertainty. Classical statistics is prepared to analyze one trait, the one for which the experiment was designed, giving significant differences when the difference between treatments is higher than the value for which the experiment was designed. In practice many traits are analyzed in the same experiment, giving often significant differences when they are irrelevant or non significant differences when they are substantial. Sometimes there is no experimental design (using field data, for example), and results are often misinterpreted, particularly P-values and non significant differences. Bayesian statistics offers a much more intuitive approach, based in probabilities.

The result of Bayesian statistics is a function called "posterior" used to find the probabilities the will describe the uncertainty about the differences or the parameters we are estimating. Multiple integrals, often impossible to calculate even using approximated methods, are needed to calculate these probabilities, but the recent use of numeric methods called "Monte Carlo Markov Chains" (MCMC) permits to estimate samples of the posterior distributions. Long chains of numbers, random samples of the posterior distributions, are used now for Bayesian inference. **Rabbit** makes Bayesian inferences without publishing the chains, but if they are needed for particular inferences, the program **Bunny** offers the chains for each difference or parameter estimated. The programs **Compare** and **Bayes** help in managing these chains for inferences.

WHAT Rabbit DOES

Rabbit makes Bayesian inferences on a statistical model with several effects and covariates. It is designed for analysis of experiments using the advantages of Bayesian statistics. It is the equivalent to GLM or MIXED programs of SAS, but (at present) it does not include interactions. It allows including in the model one effect of the type known in classical statistics as a "random effect". It permits to analyze several traits with the same model. It also permits missing values in traits or covariates. The programs **Bunny**, **Compare** and **Bayes** can be used to obtain intermediate results (like the MCMC chains) or for particular inferences.

HOW TO ANALYZE THE DATA

The model

The more general model is

$$y = m + N + T + b \cdot x + p + e$$

Where *y* are the data, *m* is the general mean, *N* one or more 'noise' effects, *T* is one or more 'treatments', *x* one or more covariates, *p* is a random effect and e is the residual.

EXAMPLE

We are going to analyze perirrenal fat (PRF) and intramuscular fat of rabbits loin (IMF) coming from five generations of selection for growth rate (*Gen*) and two sexes (*Sex*), at constant liveweight (*LW*) and muscular pH (*pH*). The data were taken at different seasons (*Season*) and different parities (*OP*). We took two rabbits per litter, thus we have added a litter random effect (*litter*). The model is

$$PF = m + Season + OP + Gen + Sex + b_1 \cdot LW + b_2 \cdot pH + litter + e$$

General mean: m
Noise effects: Season, OP
Treatments: Gen, Sex
Covariates: LW, pH
Random effect: litter
Residual: e

Random effects

They are effects in which we are not interested. We assume that repeating the experiment they will be different at random. For example, if we are examining the effect of a treatment on litter size and we have several litters per dam, when repeating the experiment we would have different dams and we are not interested in the particular value of the effect of each dam. The same applies to the litter effect in the trait perirenal

fat; the effect of the dam on two full sibs is a random effect, since repeating the experiment the dams will be different.

Noise effects

They are effects in which we are not interested. We would like that all rabbits would have been measured in the same season and parity, but it was not possible and we need to correct these effects. By including these corrections, all data come *as if* they were measured in the same season and parity. Of course, season and parity has a different effect on each animal, and we correct for an average seasonal and an average parity effects instead of the real effect on the animal, but it is important that we have data enough in each level of these noise effects to obtain an accurate estimate of these average effects. **Rabbit *will consider these noise effects, but will not perform Bayesian inferences on them***, since they are not a matter of interest.

Treatments

They are the effects in which we are interested. In our example we would like to know whether there are differences between sexes, and to know the means of the traits in each generation. Rabbit gives inferences on means of each level of each treatment and also inferences on comparisons between levels of each treatment. These comparisons can be differences between levels or ratios between them.

Covariates

Covariates are effects that are linearly related to the trait analyzed. They are often used as 'noise' effects, but sometimes they have interest in themselves, thus **Rabbit** analyzes them as treatments. Covariates can be discontinuous (as litter size) or continuous (as liveweight), but even continuous covariates can be converted in discontinuous by classifying then in different levels, thus the characteristic of a covariate is not to be continuous or not, but the linear relationship with the trait.

When shall I put a 'noise' effect in a model?

When it makes sense. The researcher should know when from her biological knowledge of the problem; statistics is a tool, it cannot substitute thinking. ***The only way in which is should NOT be included or not is attending to whether it is significant or not.*** Significance only means "sample large enough" and N.S. only means "too small sample", and both are not related to the size of the effect. With large samples everything is significant, and with small samples N.S. is common. Moreover, if an effect does not actually have influence on the trait and we put it, results will be

approximately the same, and our sample will not show an effect that should not appear. Normally accuracy will improve because the residual variance will be lower and the loss of degrees of freedom will be minimal. Experiments have few noise effects, and filed data have many data in each effect. If doubting about whether to include or not an effect the best way is to perform both analyses, with and without the effect, and see the consistency of the results.

When shall I put a random effect in the model?

Results with or without random effect are similar in general, with some exceptions. Uncertainty ("s.e." in classical statistics) is better estimated when there is a random effect and it is included in the model. Effects are considered random when they cannot be corrected properly as fixed effects because there are few data in each level of the effect. For example:

- Repeated measurements on individuals: several lactations or several litters. In this case the dam is a random effect for litter size or for milk production.
- Common environment: if we take data of full sibs, the common litter environment is a random effect.

When having few data per level the correction performed by a random effect is rather light, but if we have a substantial amount of data per level, the correction of random and fixed effects are similar. However if a fixed effect has few data per level, the estimate of this effect is inaccurate and the results for the treatments can be seriously affected, thus it is better in these cases to take this effect as random. For example, herd-year-season effect is usually considered fixed in dairy cattle, but when the farms are small and there are few data per farm-year-season, it can be considered as a random effect.

Shall I be worried if the data are not normally distributed?

Rabbit has been programmed under the supposition of data normally distributed. However, if this is not the case it can be also used, depending on sample size and on how far from normality are the data. As Ronald Fisher (one of the founders of classical statistics) stressed, even if the data are not normal, the sample means can have a normal distribution. For example, the trait number of dead rabbits in a litter follows a distribution in which the value 0 is the most frequent, 1 is less common, 2 is rare and 3 or more is rather exceptional. This is not a Normal distribution, but the average of number of dead rabbits per litter in a sample of 30 litters of a breed can be 0.8, and if we repeat the experiment it can be 0.9, repeating again 0.7, and many samples can be normally distributed around the true value (say, 0.8). Of course, if we take one litter per

sample instead of 30, we will repeat the distribution of number of dead rabbits which is not normal, but if the sample is large enough, the distribution tends to normality. The same occurs in Bayesian statistics, in which all posterior distributions tend to normality when the sample is big enough, depending on how far are the data from normality. I general, results are robust to lack of normality and transformations, often difficult for interpretation, are not usually needed.

HOW TO WORK WITH THE PROGRAMS

The program Rabbit

DATA FILE

TYPE

Data should be in a text file of the type **File.txt**

NAME

The data file name should not contain white spaces, stresses or symbols different from the English alphabet

CONTENT

The effects and data ***should appear in the following order***:

Random effect, Noise effects, Treatments, Covariates, Traits

The first line is for the names of effects and traits. Names should not contain white spaces, stresses or symbols different from the English alphabet. ***The program will only keep the first three characters of each name***; thus it is convenient to use short names like OP instead of Order_of_Parity, and avoid names as Trait1, Trait2, Trait3, since the program only will keep “Tra”. ***Traits should have at least three characters*** (for example, pH is an inappropriate name for a trait); this is not needed for effects or covariates.

Levels of random effect, noise effects and treatments should be identified with numbers from 1 to n. ***Letters or numbers not starting in 1 should not be used to identify levels.***

The program writes results with a maximum of six figures and four decimals, thus the scale of the traits should be adequate (for example, do not write the weight of cows in milligrams). ***A maximum of ten digits can be used for traits or covariates***. If the data file was generated from an Excel file or

other programs, it may happen that more than some variables have more than ten digits, it should be checked.

Missing values are permitted in traits and covariates. A missing value should be a number that could be identified later as different from all numbers in the data base (for example, 99999). It should have less than 12 digits. Negative numbers can be used.

EXAMPLE

We are going to analyze perirrenal fat (PRF) and intramuscular fat of rabbits loin (IMF) coming from five generations of selection for growth rate (Gen) and two sexes (Sex), at constant liveweight (LW) and muscular pH. The data were taken at different seasons (Season) and different parities (OP). The model is

$$PF = Season + OP + Gen + Sex + b_1 \cdot LW + b_2 \cdot pH + e$$

Notice that we have not included any random effect

The data file will be called **Fat.txt** and should be like this:

Season	OP	Gen	Sex	PV	pH	IMF	GPP
1	3	1	1	1790	5.64	1.089	9.20
1	1	1	2	1795	5.45	1.118	12.40
2	2	1	1	1505	99999	0.939	3.10
1	2	1	1	1870	5.64	1.229	13.82
1	2	2	2	1725	5.51	1.068	7.01
2	1	2	1	1570	5.44	1.439	8.20
1	1	2	1	1780	5.47	1.359	9.70
1	1	2	1	1675	5.45	1.066	7.81

Notice that there is a missing value 99999 in pH

If having a random effect, as in the model of page 3 (litter effect), the data file should be:

Litter	Season	OP	Gen	Sex	PV	pH	IMF	GPP
1	1	3	1	1	1790	5.64	1.089	9.20
1	1	1	1	2	1795	5.45	1.118	12.40
2	2	2	1	1	1505	99999	0.939	3.10
2	1	2	1	1	1870	5.64	1.229	13.82
2	1	2	2	2	1725	5.51	1.068	7.01
3	2	1	2	1	1570	5.44	1.439	8.20
4	1	1	2	1	1780	5.47	1.359	9.70
4	1	1	2	1	1675	5.45	1.066	7.81

PROCEDURE

The program will ask questions. It is prepared to use by default parameters that should be adequate for almost every analysis. The programs **Bunny** and **Bayes** permit a more detailed analysis of each trait with several options.

EXAMPLE

Enter the name of the file with its extensión.txt

Fat.txt

Enter the total number of traits

2

Enter the number of 'noise' effects

2

Enter the number of 'treatments'

2

Comparisons between treatment levels: Enter 1=DIFFERENCE, or 0=RATIO

1

Enter the number of covariates (0,1,2,…)

2

Has the data file missing values? (*1=Yes, 0=No*)

1

Please enter the missing value

99999

Enter the probability for HPD interval (for example, 0.95)

0.95

Enter the probability for the guaranteed value k (for example, 0.80)

0.80

Do you want to calculate probabilities of relevant values for some traits? (*1=Yes, 0=No*)

1

Do you want to calculate the probability of intervals [a,b]for some traits? (1=Yes, 0=No)

0

TRAIT IMF

Do you want to calculate probabilities of relevant values for this trait? (*1=Yes, 0=No*)

1

Enter r for P<r or P>r

0.02

TRAIT GPP

Do you want to calculate probabilities of relevant values for this trait? (*1=Yes, 0=No*)

0

EXIT FILES

A file *Summary* for each trait (in our example, **SummaryIMF.txt** and **SummaryPRF.txt**), with

- A description of number of records and missing values
- A description of the model, with noise effects, treatments and covariates
- The variance of the error and the DIC of the model
- The characteristics of the MCMC chain
- The limits for the mean and the effects

EXAMPLE

SummaryIMF.txt

```
Trait                                  =  IMF

 Data file                              =  Grasa.txt
 Number of rows in the data file        =        502
 Rows with missing values in covariates =          5
 Other missing data                     =          8
 Number of data analysed                =        489

 Number of 'NOISE' effects      =  2
 NOISE  effects                 =  Est   OP
 Levels of each effect          =  2     3

 Number of 'TREATMENTS'         =  2
 TREAMENTS                      =  Gen   Sex
 Levels of each effect          =  5     2

 Number of covariates           =  2
 COVARIATES                     =  PV    PH

Number of random effects    =  0

 Ve =      0.0228
 DIC =   1422.9236

 Chain length   =      60000
 Burn-in        =      10000
 Saved length   =       5000
 lag            =         10

 Bounds for the mean and the effects
  -2.82085404967600
   2.82085404967600
```

A file *Results* for each trait (in our example, **ResultsIMF.txt** and **ResultsPRF.txt**), with

- Bayesian analyses for Mean, Residual standard deviation (RSD), coefficient of variation of the trait (CV= 100 x mean/sd) and for each covariate.
- Bayesian analyses for the mean of each treatment
- Bayesian analyses of all contrasts (or ratios) between levels of treatments.
- Residual variance, variance of the random effect (when appropriate) and DIC of the model

Bayesian analysis gives the following results:

- **Mean** and **median** of the posterior distribution of the parameter being estimated (means, CV, RDS and contrast or ratios). Both are estimates of the parameter and should be similar if the sample is large enough. The Mean minimizes the expectation of the square errors, and the Median the expectation of the absolute value of the errors. In practice, usually they will have almost the same value.
- **HPD** is the shortest confidence interval[1], the one allowing the highest precision, with a probability that we determine (usually, 95%). It is not necessarily symmetric around the estimate (the **mean** or the **median** of the distribution); for example, a 95% confidence interval of a correlation coefficient of 0.9 can be [0.65, 0.95], avoiding wrong expressions of the type 0.9±0.2 sometimes found in papers. Nevertheless, in practice, for means and contrasts, it will be normally symmetric around the estimated value.
- **Effect** is the estimate of the level effect referred to the mean (i.e., the sum of all of them is null). The program gives also its probability of being higher or lower than zero.
- **Probability of the covariate or the contrast being higher than zero** or (if it is a ratio) higher than 1. This substitutes all hypothesis tests with the advantage of giving the actual probability of the difference between levels of a treatment.
- **Probability of relevance**. It is the probability of the contrast being higher than a value that we consider to be relevant from an economic or biological point of view. This is the value that would be used for designing an experiment, finding the size of the sample to get significant values when they will be higher than this relevant value. The relevant value is independent of the experiment, and it is often found using economical arguments as the

[1] Bayesians prefer to call it “credibility interval” or “high posterior density”.

minimum value from which a decision should be taken; for example, a difference between levels of a treatment of one piglet in al litter or 0.1 of index of conversion are usually considered to be relevant to choose one of them, whereas lower quantities are not. Sometimes it is difficult to find it (for example in enzyme activities or results from a meat panel test). In these cases it is recommended to use a fraction of the standard deviation of the trait (most productive traits in animal production have a relevant value between ½ and ⅓ of their standard deviation). Another solution is to use ratios instead of contrasts; for example, it can be determined the probability of a level of a treatment to be a 10% higher than other one. Rabbit recalculates relevant values when the contrast is negative or the ratio lower than 1; for example, if the relevant value of a contrast is 0.2 and the contrast is negative, the new relevant value is -0.2; if the relevant value for a ratio is 1.10 and the ratio is lower than one, the new relevant value is 0.90.

- **Probability of similitude**. It is the probability of the absolute value of a contrast being lower than a relevant value. A high probability of similitude (for example, a 96%) means that the difference between levels of a treatment is null for practical purposes, or it is irrelevant from a biological point of view. If this probability is intermediate (say, 40%), the experiment has no data enough to decide whether there are differences between levels of a treatment or not. This allows distinguishing between "*there is no difference between levels of a treatment*" and "*I do not know whether there is a difference between levels of a treatment*", something that the classical result "N.S." does not permit. It should be noticed that there are always differences between treatments, since a difference of exactly 0.000…..has a null probability; levels of treatments will be always different whatever small is this difference. Our objective is not to find these small differences (otherwise the samples had to be extremely large), but to find when the differences between levels are small enough to become irrelevant for practical purposes or biological interpretations ([2]).
- **Guaranteed value k**. It is the value of the interval [k, +∞) with a probability determined by the user. Sometimes it may be interesting to guarantee a minimum value with some probability. A difference between levels of a treatment can be important, but according to a high HPD, the true value can be irrelevant. The value **k** guarantees that the difference between levels will

[2] If the relevant value r is extremely small, the probability of similitude (the probability of the true value being between –r and +r) can be also extremely small and it can be not well estimated.

have *at least* this value with a probability determined by the user (for example, 80% or 95%). The lowest bound of the HPD interval does not guarantee this. In the case of a negative difference between levels, k is the limit of the interval (-∞,k]. If the contrast is very inaccurate, the value k is not useful. For example, a difference between index of conversion of two foods A and B is 0.3 with k= 0.2; in this case k is a useful value, guaranteeing that at least the difference will be 0.2. If k=-0.1 it is not useful, since we guarantee the interval [-0.1, +∞), which means that food A can be better than B or conversely B can be better than A, we do not know. ***Rabbit only gives the value k when it is useful.***

- **MCse**: It is called "Monte Carlo standard error". It is a s.e. originated by the numerical method used to analyze the data. It is not important in our case because **Rabbit** produces very small MCse for the type of problems that analyzes. If for some reason it is large it would be better to use the program **Bunny** that allows changing the parameters of the numerical methods.
- **Geweke test**: It is a test that detects cases in which the numerical methods do not work properly. In our case this should never happen, but sometimes referees ask for it. It should be lower than 2 in absolute value.

For **Mean**, **RSD** and **CV**, the program only gives **Mean**, **median** and **HPD**
For **covariates** the program also gives **P>0** or, for ratios, **P>1**, **MCSe** and **Geweke test**
For **Means of the treatments** the program also gives the **guaranteed value k**
For **contrasts** and **ratios** the program gives the full set of analyses
When a random effect is included, **RabbirR** also gives the variance of the random effect and the residual variance.

EXAMPLE

ResultsIMF.txt

```
_____________________

 Results for trait IMF
_____________________

_____________________

 Results for trait IMF
_____________________

 ================================================
```

```
 RESULTS OF       MEAN
 ================================================

Mean   =    1.1642
Median =    1.1647

 HPD 0.95 = [   1.1119 ,    1.2231  ]

 ================================================
 RESULTS OF        RSD
 ================================================

Mean   =    0.1509
Median =    0.1508

::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::

 ================================================
 RESULTS OF       PV
 ================================================

Mean   =   -0.0075
Median =   -0.0076

 (P>0 )= 0.13
 (P<0 )= 0.87

::::::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::::::

 ================================================
 RESULTS OF  Gen1
 ================================================

Mean   =    1.1516
Median =    1.1519


  HPD 0.95 = [   0.9992 ,    1.3050  ]

Effect=       0.0126
(P>0) = 0.58
(P<0) = 0.42

 GUARANTEED VALUE k
Probability of the interval [k,inf) = 0.80
 k for the interval [k,+inf)           =    1.0874

MCse =   0.001699
Geweke Z-Score =  0.41

 ================================================
 RESULTS OF  Gen2
 ================================================
```

```
Mean   =    1.1907
Median =    1.1911

::::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::::

 ================================================
 RESULTS OF  Gen1 - Gen2
 ================================================

Mean   =   -0.0391
Median =   -0.0393

 (P>0 )= 0.24
 (P<0 )= 0.76

HPD 0.95 = [  -0.1503 ,    0.0708  ]

 PROBABILITY OF RELEVANCE

 Relevant value
r =   -0.0200

(P>  -0.0200) = 0.15
(P<  -0.0200) = 0.85

 PROBABILITY OF SIMILITUDE

Probability of interval [  -0.0200 ,    0.0200] = 0.22

MCse =   0.001151
Geweke Z-Score =  0.50

::::::::::::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::::::::::::
::::::::::::::::::::::::::::::::::::::::::::::
```

The program Bunny

DATA FILE

The same as in **Rabbit**

PROCEDURE

The program will ask questions as in **Rabbit**. The program will ask the user whether she wants to change prior bounds or the numerical procedure (MCMC) parameters.

EXIT FILES

- A file *Summary* for each trait, as in **Rabbit**
- A file *Descr* for each trait, with the chains for general description of the trait: Overall mean, Residual standard deviation, Coefficient of variation and covariates.
- A file *Means* for each trait, with the chains of the means of each treatment level.

EXAMPLE

DescrIMF.txt

IMF	ITER	MEAN	RSD	CV	PV	PH
	1	1.1849	0.1556	13.13	-0.0134	0.0003
	2	1.1491	0.1497	13.02	-0.0183	0.0003
	3	1.1645	0.1462	12.56	0.0003	0.0002
	4	1.1477	0.1545	13.46	0.0000	0.0002
	5	1.1367	0.1536	13.52	-0.0063	0.0003
	6	1.1892	0.1490	12.53	-0.0159	0.0003
	7	1.1738	0.1501	12.79	-0.0048	0.0002
	8	1.1439	0.1574	13.76	-0.0096	0.0003
	9	1.1801	0.1509	12.78	-0.0084	0.0003
	10	1.1689	0.1514	12.95	0.0018	0.0003

MeansIMF.txt

IMF	ITER	Gen1	Gen2	Gen3	Gen4	Gen5	Sex1	Sex2
	1	1.1495	1.2159	1.2392	1.2020	1.1654	1.1022	1.1983
	2	1.0888	1.1788	1.1772	1.1653	1.1427	1.0813	1.1674
	3	1.1728	1.1920	1.2100	1.2019	1.1296	1.0893	1.1642
	4	1.1194	1.1810	1.2303	1.1550	1.1438	1.0284	1.1477
	5	1.0583	1.1491	1.1661	1.1928	1.1186	1.0568	1.1430
	6	1.2335	1.2287	1.2260	1.2368	1.1563	1.0983	1.2051
	7	1.2351	1.1796	1.2345	1.1917	1.1382	1.1252	1.1786
	8	1.1478	1.1609	1.1514	1.2003	1.1187	1.0881	1.1535
	9	1.2005	1.1957	1.2057	1.2375	1.1532	1.1084	1.1885
	10	1.1935	1.1867	1.1962	1.2229	1.1361	1.1023	1.1671

The program Compare

DATA FILE

A file with chains of the means of the treatments levels. It can be a *Means* file coming from **Bunny.** The first column should contain the number of the iteration, and the fist row the headings. Only the first three characters of the heading will be considered. *The number of levels of a treatment should be reasonable*; for example, 9 levels already means 36 contrasts!

PROCEDURE

The program will ask questions as in **Rabbit**.

EXAMPLE

We will prepare ratios for the file Means.txt, which has 2 treatments (Gen and Sex) with 5 and 2 levels respectively

Enter File name with its extensión.txt

MeansIMF.txt

Enter the number of Treatments

2

Enter the number of levels of Treatment 1

5

Enter the number of levels of Treatment 2

2

Enter 1=DIFFERENCE or 0=RATIO

0

EXIT FILE

A file *Contr* with all contrasts between levels of a treatment, or a file *Ratio* with all ratios between levels of a treatment.

EXAMPLE

RatioIMF.txt

ITER	Gen1/Gen2	Gen1/Gen3	Gen1/Gen4	Gen1/Gen5	Gen2/Gen3	Gen2/Gen4	Gen2/Gen5	Gen3/Gen4	Gen3/Gen5	Gen4/Gen5	Sex1/Sex2
1	0.5947	0.5916	0.6168	0.6569	0.9948	1.0373	1.1046	1.0427	1.1105	1.0650	0.8972
2	0.7765	0.7792	0.7966	0.8291	1.0035	1.0259	1.0677	1.0224	1.0640	1.0408	0.9340
3	0.8155	0.7832	0.8331	0.8543	0.9604	1.0215	1.0476	1.0637	1.0908	1.0255	0.8998
4	0.9298	0.9256	0.9297	0.9812	0.9955	0.9999	1.0553	1.0044	1.0601	1.0554	0.8917

The program Bayes

DATA FILE

A file Ready for Bayesian analysis. It can be a *Contr* or a *Ratio* file coming from the program **Compare**, or a *Means* or *Descr* file coming from **Bunny.** The first column should contain the number of the iteration, and the fist row the headings

The first name of the headings is for the iterations, for example ITER and shouldnot have blank spaces or characters different from letters or numbers.

PROCEDURE

The program will ask questions as in **Rabbit**. The program allows having different relevant values, probabilities for HPD intervals or k guaranteed values for each parameter mean or contrast.

EXAMPLE

Enter the name of the file with its extensión.txt

ContrIMF.txt

Enter the number of chains to be analyzed

11

Are the chains RATIOS between levels (1=Yes, 0=No)

0

Do you want to calculate probabilities of relevant values? (1=Yes, 0=No)

1

Do you have the same relevant value for all chains? (1=Yes, 0=No)

1

Enter r for P<r

0.1

Do you have the same HPD interval for all chains? (1=Yes, 0=No)

1

Enter the probability for the HPD interval (for example 0.95)

0.95

Do you have the same probability for the intervals [k,inf) or (-inf,k] for all chains? 1=Yes, 0=No)

1

Enter the probability for the interval [k,inf) or (-inf,k] (for example 0.95)

```
        0.95
Do you want to calculate the probability of intervals [a,b]? (1=Yes, 0=No)
        0
Enter the number of columns for each name of the chain
        13
```

EXIT FILE

A file *Results* as in the program **Rabbit**.

AN EASY INTRODUCTION TO BAYESIAN STATISTICS WITH MCMC

I recommend downloading the lecture notes of a course I have given in the universities of Wisconsin, Edinburgh, Padova, Sao Paulo, Lavras, Nacional de Uruguay, Politècnica de València and in the French Institut National de la Recherche Agronomique. The can be found in http://acteon.webs.upv.es/material_docente.htm

APPENDIX I.

Significance is often irrelevant and P-values misleading, what can we do?

Paper presented at the III Iberoamerican meeting of Biometry (Barcelona, 2011).

1. Introduction

Most statistical problems in agriculture and even in biological sciences are related to estimation and accuracy of estimation and not to hypothesis testing. However, the extended practice of performing hypothesis testing, mainly because they are included in software outputs, leads often to confusion. In this paper the common use of significance, P-values, standard errors and confidence intervals is discussed. A Bayesian alternative that permits to present results in a more intuitive way, facilitating the discussion of results, is presented. In this paper only very simple examples, like comparison between two treatments or accuracy of a correlation, are presented, but generalization to more complex analyses is immediate.

2. Significance is often irrelevant

A well designed experiment should give significant results when the difference between two treatments is higher than a quantity that the experimenter considers to be “relevant”. In well designed experiments the power of the test is also considered and

n.s. will mean that there is no difference between treatments, with a determined risk of being wrong.

In agriculture, typically many traits are measured in a single experiment. What happens when a hypothesis test is applied to traits that are not the one used to design the experiment? Signification can appear when the differences between treatments are irrelevant (which will not be a problem if the scientist pays attention to the irrelevance of the differences instead stressing the signification), but n.s. can appear when the differences are relevant. In this case, as the significance level and the power of the test were not calculated for this trait, we should say “we do not know whether there are differences between treatments or not”, which is a poor result for possible important differences between traits.

Something worse can happen; we can find a small difference between treatments and an n.s. beside it, giving the false impression that there are not differences between treatments. For example, a difference in litter size of 0.1 piglets (n.s.) is irrelevant, but the confidence interval can be [-1.5, 1.7], implying that we may have a big difference between treatments and we do not know which treatment is the better one. However, typically in agricultural papers, the discussion will be centred in the small 0.1 estimate and its n.s.

Something even worse can happen, finding a relevant difference between treatments, significant, but with a confidence interval (C.I.) high enough to include irrelevant values. For example, finding a difference of 1.1 piglets (more than 1 piglet is an economically relevant difference), with a C.I. of [0.3, 1.9], which means that it is possible that the difference between treatments is actually irrelevant. People have the false impression that the true value should be around the middle of the interval, which is not the case. Repeating many times the experiment we will obtain a different C.I. in each repetition, of which 95% of them will contain the true value, but we do not know where, it can be near the centre or near one side. As we have actually only one interval, we utter that this is “one of the good ones” with the hope of being right at least a 95% of the times we utter this. In our example, the true value can be 0.5 or other irrelevant value. However, typically again, all the discussion will be about the estimation 1.1 and the ‘*’ of significance obtained.

We can also obtain significant differences just by hazard. In agriculture often many traits are measured and the statistical models contain several effects with many levels.

This means that significant results will appear by chance. There is the false impression that a significance level of a 5% leads to a 5% of probability of being wrong. As the significance levels are determined before the experiment starts, it cannot be the probability of being wrong, which depends o the sample size and the characteristics of the trait. The result of the test is like a death sentence: “yes” or “not”. If the answer is “yes” we will *behave* as if the probability of finding differences between treatments is 100%, hoping to be right at least a 95% of times along our career. It may happens that we were right a 99% of times along our career, but *the test does not give any rule of measurement about how much evidence we have from our results*, it only gives a rule of behaviour. Because of that, it is strictly incorrect to speak about higher or lower levels of significance and put one, two or three starts depending on our actual results.

Frequently in practice, and always when field data are used, there is no experimental design or the power of the test has not been calculated, leading to significances or n.s. that have a difficult interpretation. When data come from complex or expensive experiments, facilities are scarce and n.s. is a frequent result. In these cases experimenters try to give some meaning to their estimates using nonsense like that “there is a trend in the results”. A P-value of 0.10 does not show a “trend” but a level of ignorance, the actual difference can be very high and not just “a trend”.

The problem lays in the question to be solved. This question is not “are there differences between treatments”? We do need an experiment for this, the answer is “yes”, if the sample is big enough everything will de significantly different, and if the sample is small enough no differences will be found. Because of that, it is nonsense to include effects in a model if they are significant and remove them if they are not. If an effect is n.s. but the difference between levels is high we cannot be sure that it is not having a strong influence in the results. Nowadays model selection using AIC, BIC, DIC and other criteria are used to avoid models too highly parametrised, but the problem then is that it is not clear which is the meaning of a difference between models of say, 4 points. Crossvalidation techniques are not free from criticism, since modes can fit well in some areas of the parametric space and poorly in other areas, they do not provide a choice criterion.

2. P-values are misleading

P-value is the probability of obtaining the difference between treatments that we actually obtained or a higher one. If it is very low either we got an exceptional sample or it is not true that the treatments are equal. The problem is how much exceptional is,

say, a 4%. Again, *this is not the probability of being wrong when we say the treatments are equal*. It is not the significance level, because if we repeat the experiment this value will change, and we take conclusions not only from our sample but from the hypothetical repetitions of our experiment. Thus, *how much evidence* gives a 4% about the treatments being different? For example, if we get a P-value of 5% and the true value of the difference between treatments is the same as our sample (it should not be far away if our estimate is good), when repeating the experiment all samples will be around the true value and half of them will be n.s. When researchers obtain a value of 5% they see that a 5 is small when compared with a 100 and they do not think that repeating the experiment half of the times they can find n.s. results (all of this is independent on the power of the test). We know that a small value gives more evidence than a larger one, but again *we do not have any scale of measurement of evidence*. Of course, a P-value of 0.0001 gives a lot of evidence, but statistics has not been created to find the difference between a mouse and an elephant, but to decide whether results that may be due to chance are in fact due to chance or not.

3. What to do

Using Bayesian statistics results can be better understood since the uncertainty is expressed in terms of probability. Here we have some examples:

Substituting hypothesis tests for difference between treatments A and B by $P(A-B|\mathbf{y})>0$ (where **y** are the data). This is not a hypothesis test, there is nothing like significance here. But it substitutes it with the advantage that gives the actual probability of the treatments being different. If this is 93%, the researcher decides whether this is enough for the trait he examines.

Finding the probability of a difference between treatments being higher or lower than a relevant value. A relevant difference R has an economical or biological meaning under which both treatments are equal for practical purposes. This relevant value is defined when designing experiments to find significant differences when they are higher or lower than R.

Finding the probability of similitude of treatments. We can find the probability of the absolute value of treatments being higher or lower than a relevant value R. This allows, for practical purposes, to distinguish between “there is no difference” from “I do not know whether the treatments are different”.

Finding the minimum value with a probability of 95%, or other probability. Sometimes we want to find which is *at least* the difference between treatments with some fixed uncertainty. We can calculate the value *k* for the intervals [*k*, +∞) or (–∞, *k*] of the marginal posterior distribution of the difference between treatments containing a 95% of probability. We choose on or other interval depending on this difference being positive or negative.

Estimating ratios instead of differences. Sometimes may be useful to express the result as ratios, for example when it is not clear which can be the relevant value R because of the type of trait analysed. We do not know how relevant is three points of liver flavour in a meat panel test score, but we understand how much is to be a 13% higher. Using MCMC it is trivial to find the marginal posterior probability of a ratio and then to estimate, for example, the probability of the ratio of being higher than a 10%.

Estimating the shortest interval having a probability of 95%. This is the classical HPD95% Bayesian solution. This interval is not symmetrical around the estimate (for example a correlation coefficient can have an estimate of 0.9 and a HPD95% of [1, 0.6]. This solution is more intuitive than giving, as usual, for example an estimate of 0.9 with s.e. of 0.2.

Estimation is an area of knowledge in which statistics gives solutions for practical purposes, but model selection is still an open area. Most statistical problems in agriculture are estimation problems or can be converted into estimation problems for practical purposes. Bayesian techniques offer an easier way to understand results and permit to show uncertainty in many ways that are useful for taking decisions.

APPENDIX II.

Technical characteristics of the programs

Rabbit is programmed in FORTRAN95. It has not been intended to optimize computing, since it is not needed for the type of problems to which **Rabbit** is addressed. Experiments have always reasonably small reduced data bases.

Rabbit does not estimate (at present) interactions. The characteristics of the models that **Rabbit** solves are:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Wc} + \boldsymbol{Zp} + \boldsymbol{e}$$

Whre $\boldsymbol{y}$ are the data, $\boldsymbol{b}$ all the effects (mean, Noise and treatments), $\boldsymbol{c}$ the covariates, $\boldsymbol{e}$ the residuals and $\boldsymbol{X}$, $\boldsymbol{W}$ the design matrixes.

$$\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}, \sigma_e^2 \sim N\left(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2\right)$$

$$\boldsymbol{b} \mid inf, sup \sim U\left(inf, sup\right)$$

$$\boldsymbol{p} \mid \boldsymbol{Z}, \sigma_p^2 \sim N\left(\boldsymbol{0}, \boldsymbol{I}\sigma_p^2\right)$$

$$\boldsymbol{c} \mid \sim U\left(-\infty, \infty\right)$$

$$\sigma_p^2 \sim U\left(\boldsymbol{0}, \infty\right)$$

$$\sigma_e^2 \sim U\left(\boldsymbol{0}, \infty\right)$$

where inf, sup are different bounds for the mean and levels of all effects. I<f they are not given by the user, their values are $\pm 10s$, where s is the standard deviation of the sample. This is formally incorrect, because Bayesian analyses should not use the sample to define prior information, but in practice it is completely irrelevant, because if they are higher values they will be outliers or the distribution of the data too far from normality (in this case the program should not be used). Improper prior distributions are used for covariates and residual variance, since bounds are difficult to determine *a priori.* Even being improper prior distributions, posterior distributions are always proper for these cases.

The program uses Gibbs sampling to estimate the marginal posteriors. As default it takes 60,000 iterations with a burn-in period of 10,000 and it keeps one each 10 iterations, which means that 5,000 samples are used to estimate each distribution. This is more than enough for the type of problems that **Rabbit** solves. It could be argued that samples can be directly taken from the marginal posterior distributions, and that only 5,000 iterations without burn-in and without lag between sampling would give the same result, but as computation is cheap and quick this becomes irrelevant.