Available online at www.sciencedirect.com
SCIENCE DIRECT®
Animal Feed Science and Technology
129 (2006) 1–11
ANIMAL FEED SCIENCE AND TECHNOLOGY
www.elsevier.com/locate/anifeedsci

Editorial

Some experimental design and statistical criteria for analysis of studies in manuscripts submitted for consideration for publication

Abstract

This editorial discusses some statistical principles that may be useful in guiding authors of manuscripts submitted to Animal Feed Science and Technology (AFST) for consideration for publication. The editorial also discusses some common experimental designs and statistical models used by AFST authors, including some of the most frequent problems that cause misunderstandings and disputes among authors, reviewers and editors. The editorial is not meant as an exhaustive treatise on all possible experimental design and statistical issues that could arise, but as a general guide to correctly interpret results that are not meant to limit anyone's imagination, as there are many ways to evaluate and examine biological responses.

Keywords: Experimental design; Statistics

1. Introduction

As a subject, statistics sometimes seems, for many animal and plant scientists, to be a necessary evil that is required by scientific journals as a prerequisite to consideration of a manuscript for publication. However, an appropriate statistical model properly applied to data allows experimental results to be interpreted accurately and completely. In contrast, inappropriate or incomplete statistical designs can lead to inaccurate, or incomplete, conclusions that may lead to misinterpretation of the results of the study. As the cost of research,

Abbreviations: AFST, Animal Feed Science and Technology; SAS, statistical analysis system; SPSS, statistical package for the social sciences

0377-8401/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.anifeedsci.2006.05.011

particularly animal research, continues to escalate in terms of money and paperwork, it is the responsibility of all experimentalists to extract as many appropriately supported conclusions as possible from their research.

A common characteristic of manuscripts submitted for consideration for publication in Animal Feed Science and Technology (AFST) is that virtually no authors complete their own statistical analysis by writing a statistical model and analyzing their data parameter by parameter, which was the case as recently as 30 years ago. Virtually all authors utilize a 'statistical package' that only requires a 'model statement' and a data array to create pages of output that purport to show the probability that various response parameters differ by applied treatment. There are numerous problems with this approach, the primary one being that it is now the rare exception when authors actually understand the statistical model that they used. In addition, authors are hostage to the statistical software package that they chose, and can be unaware of errors within it that are creating inaccurate statistical implications. It is also clear that the statistical models employed often tend to be those that are within the available statistical software package and that the statistical models available within that software package may or may not be the most appropriate to the experimental objectives.

The Co Editors in Chief take the issue of appropriate statistical analysis of experimental results very seriously and it is AFST policy to have two animal scientists with a substantial statistical background on the AFST editorial board at all times.

The objective of this editorial is to provide guidelines on experimental design and statistical criteria that are deemed acceptable and/or desirable and/or required for studies in manuscripts submitted for consideration for publication in Animal Feed Science and Technology. However, these are guidelines and each manuscript will be considered on its individual merits.

2. General areas that are often overlooked

2.1. Use a statistical model

It is highly desirable that a statistical model actually be used. In general, tables that simply contain a matrix of means by treatment and parameter, all with an S.E. in parenthesis beside it, do not represent a statistical model. While it occasionally may be suitable, it seldom leads to an organized ability to make statements relative to differences among values due to treatment within parameter.

In general, data should not be presented in tables without having been statistically analyzed and those statistical results should be presented with the data in the table. This may seem intuitive, but in fact presentation of data without statistical analysis occurs more frequently than might be anticipated in submitted manuscripts. The most frequent area where data columns without statistical analysis are presented (and sometimes published) is in the definition of the biomasses, materials and/or diets that were used in the study. Indeed such tables often contain very little information on the number of samples collected and/or assayed, in spite of their representing the materials upon which the study was completed (and the hypothesis being predicated upon). A general rule is that if you want to be able to

make statements in the 'Results' section that, for example, 'the NDF level of grass A was higher than grass B', then you need to provide statistical support for the statement. Thus in the case of biomasses or materials used in the study, there should be enough samples collected in a suitable experimental design to support a statistical analysis. Relative to diets, you might wish to say 'the starch level of diet A was higher than diet B, which did not differ from diets C and D'. If so, diet samples should be collected on a similar schedule as the response parameters in order to facilitate appropriate statistical analysis, unless the diets were prepared in advance of the study in quantities sufficient for the entire study in which case samples collected on the same schedule as the animal response parameters can be used as the statistical replicate.

2.2. Describe your statistical model

One of the most frequent issues that causes recommendations of manuscript rejection, and/or seriously lengthens the review process, is a failure of authors to clearly and completely describe their statistical model(s). As a result, reviewers cannot discern the meaning of statistical implications and become frustrated. That the statistical model(s) described should be completely consistent with the experimental design described in the 'Materials and methods' section, as well as the tabular presentation of results, seems intuitive, but it is often not the case.

The best description of a statistical model should describe the model in words as well as presenting it as an equation. While the latter is not an AFST requirement, providing it often prevents miscommunication with reviewers and editors that prefer to see the model presented as an equation. When several models have been tested and rejected, this should generally be indicated in the 'Results' section. As it is widely accepted that non-significant interaction terms can be omitted from the model finally adopted, it is wise to inform reviewers and editors of the process that was followed to arrive at the model used.

2.3. Define your accepted probability levels

It is wise to clearly define the probability (*i.e.*, P) values that you are prepared to accept as indicative of statistical significance at the end of the statistical methods sub-section within the 'Materials and methods' section (*e.g.*, 'differences among means with $P<0.05$ were accepted as representing statistically significant differences'). In addition, if P values slightly higher than 0.05 are accepted as trends or tendencies to differences, then this also should be stated as, for example, 'differences among means with $0.05<P<0.10$ were accepted as representing tendencies to differences'. Note that the AFST convention is that the word 'significant', without qualification, may only be used if $P<0.05$ (and restricting use of the word 'significant' to statistical effects generally prevents confusion). However, AFST will accept 'P=', rather than 'P<' or 'P>', but in this case P values should be expressed to three decimal places. In addition, P values less than 0.001 should be expressed as '<0.001' rather than the actual value. This approach is useful both in informing readers of the statistical basis for the interpretations, as well as obviating the need for repeated use of ($P<0.05$) in the 'Results' section.

2.4. Interpret within your statistical model

In general, authors should be prepared to live within their chosen experimental design and statistical model. Again, although seemingly intuitive, this is often not the case. Always be certain that the statistical model is consistent with your objectives and that your statements relative to treatment differences are statistically fully supported. Statements based upon visual appraisal of data (*e.g.*, 'there was a numerical difference between treatment A and B') are generally a short road to trouble with reviewers and editors.

3. The experimental unit

There is controversy amongst statisticians and biologists as to the most appropriate experimental unit when, for example, several animals are fed together within a pen or paddock. Broadly speaking, these positions can be divided into a view that the most appropriate experimental unit is the smallest unit upon which a treatment can be applied (generally the group of animals) *versus* a view that the most appropriate experimental unit is the smallest unit upon which a measurement can be made (generally the animal). The following two paragraphs describe these alternate approaches, and most readers will disagree with something that is written in at least one of them.

If the experimentalist takes the view that the experimental unit is the group of animals to which the treatment was applied, for example in the case of groups of animals fed in pens or paddocks, the experimental unit for feed intake would be pen, if they were group fed, or individual animal if they were individually fed and the different diets allocated at random to animals within the pen. However, different experimental units can be used within the same study if different treatment factors have been applied at different scales. An example is where diets have been fed to pens of animals within which individual randomly selected animals were treated with an animal specific treatment. Here the experimental unit for comparing main effects of diets would be pens and, for comparing treatments and their interaction with diet, the experimental units would be individual animals. In the simpler situation for diets that are fed to groups of animals in pens, and all animals are treated similarly, then the statistical advantage gained by measuring responses for groups of animals within pen, rather than one animal, is to achieve a more precise estimate of the value for that pen. This often results in lower among pen variation, which is the appropriate error term against which to compare diet effects in this design. If blood samples were collected in the same study, for example at several times during the measurement period from individual animals, then that model could include individual (*i.e.*, time specific) blood parameters as the experimental unit if 'time' was included as a repeated measure in the statistical analysis.

However, if the experimentalist takes the view that the experimental unit is the smallest unit upon which a measurement was made, for example in the case of groups of animals fed in pens or paddocks, the experimental unit for feed intake would be pen, if they were group fed, or individual animal if they were individually fed. However, different experimental units could be used within the same study. For example, feed intake would be statistically analyzed with pen as the experimental unit if the animals were group fed, gain could be statistically analyzed with animal as the experimental unit (within pen) if animals were weighed separately, but feed/gain ratio would be statistically analyzed with pen as the

experimental since one of the units in the ratio was estimated on a pen basis. The statistical advantage of this approach is that the variability in response among the animals within a group can be used in assessing statistical differences due to the applied treatment, and the biological advantage is that differences in responses among different identifiable groups of animals within the group can be assessed. If blood samples were collected in the same study, for example at several times during the measurement period, then that model could include individual animal blood parameters as the experimental unit if 'time' was included as a repeated measure in the statistical analysis.

In neither of these approaches would a situation with two pens, or groups, each with a single treatment likely to be deemed acceptable by reviewers or editors, unless the area under study was very new and/or the differences due to the applied treatment likely to be overwhelming (*i.e.*, the experimentalist is asking questions relative to extent of a known response rather than if a response occurred). The two approaches to animals fed in groups, outlined in the two proceeding paragraphs, are both considered acceptable statistical approaches in manuscripts submitted for publication in AFST, if supported by appropriate statistical models.

4. The use of 'time' in a statistical model

For 'time' to be included in the statistical model, there should be a stated, and supported, expectation of, and interest in, a pattern with time, rather than using, for example, daily or weekly values as repeated measures over an experimental period of a fixed time period. Thus, in general, the use of 'time' as an effect in a statistical model is best restricted to situations where patterns over time, or interactions of main effects and time, are of interest. For example, changes in diurnal patterns in blood or rumen parameters to a dietary challenge *per se*, or relative to interactions with a main effect, are strong candidates for inclusion of 'time' as an effect in the statistical model. In contrast, disappearance or appearance of a fraction, or compound, in an *in vitro* assay, for example, is much better handled by use of a predictive model, such as: $y = a + b\,e^{-kt}$ to describe disappearance of a feed fraction from an in situ incubation. This approach allows the pattern of disappearance or appearance to be described, and parameters, or functions of the parameters (*e.g.*, $t_{1/2}$ or mean retention time), can be statistically analyzed for treatment differences, where appropriate.

Repeated measures in time should, with very few exceptions, not be handled by repeated statistical analysis of each set of values, by time, as this increases the risk of Type II statistical error (*i.e.*, accepting that a difference occurred when it did not), while failing to handle the data as a single study.

5. Normal distribution

That variables are normally distributed around their means is generally true and usually taken for granted by animal scientists. However, if there is a reason to believe that it may not be true, it should be tested. Such tests are available in several statistical packages and variables can be transformed if they are found to not be normally distributed. Variables such as pH and fractional rate constants may, in spite of what may seem logical, be normally

distributed, whereas particle size, for example, has a log-normal distribution. The general rule with non-normality is to consult a statistician.

When presenting data of statistical tests completed on transformed values, it may be desirable to report both the results of the analysis (*e.g.*, log means ± S.E.M.) and the untransformed means so that the reader does not have to re-calculate them. As generalized linear models (GLM) are widely available, an alternative approach is to use Poisson or Binomial errors, where appropriate, *versus* log transformations.

6. The use of covariate adjustment

Covariate adjustments are, in general, used less frequently than in the past, even though they can be a very powerful means to reduce variation and decrease treatment P values. However, covariates should, in general, be the same parameters as are being adjusted and should reflect a sufficiently long collection time so as to be accurate. For example, pre-experiment milk fat concentration would likely only be used to adjust milk fat concentration covariately within the experiment, and not milk protein concentration for example, and should represent a time period broadly equivalent to the length of the measurement period within the experiment. Thus experiments could frequently have only some response parameters covariately adjusted, and it is critical that this is clearly explained, and the covariates defined, in the statistical methods section of the manuscript.

7. Definition of statistical replicates

It is not uncommon that authors confuse laboratory replicates with experimental replicates. Experimental replicates represent different samples of the material in question, rather than repeated analysis of sub-samples of the same material. While a laboratory replicate is often associated with a chemical assay, it can also relate to a biological assay, such as an *in vitro* fermenter, a mini-silo or even a ruminally cannulated cow. For example, an experimentalist collects a large sample of whole crop maize from a field and divides it into four sub-samples, each of which is treated with a different enzyme prior to ensiling each in five mini-silos per treatment. At mini-silo opening, the materials in the mini-silos are dried and analyzed in triplicate for various compounds. In this case, the parameter assays and the mini-silos are both laboratory replicates and so no meaningful statistical analysis is possible. However, if the large sample of whole crop maize had been randomly divided into 20 sub-samples, and five of each had been treated with each enzyme and ensiled singly in mini-silos, then the mini-silos are statistical replicates.

8. Some common designs

8.1. Factorial designs

Factorial designs are very common in experiments within manuscripts submitted for publication to AFST. Such designs can be very powerful ways to examine interactions of

several key parameters, and they are in fact designed to do just that. However, authors are cautioned that factorial designs with more than three factors can be extremely difficult to present to readers in a cogent fashion, and even more difficult to interpret, particularly in cases where the highest level interaction is statistically significant. Since one of the main reasons for using factorial designs is to examine interactions, it is critical that authors actually do so. Thus, for example, a 3×2 factorial should generally be presented in tables as six data columns with either the second factor means listed within the first factor, or *vice versa*. In general, main effect means should not be pooled since it is always possible for readers to calculate main effect means from the individual means within factors, but not possible to calculate the individual means within factors from pooled means. There should also be a column to express an S.E. or S.E.M. and, in the case of the 3×2 factorial, columns to show the P values for each main effect and the interaction. In cases where the interaction is deemed statistically significant, then a secondary statistical test should be used to separate the efficacy of factor 2 within factor 1 (or *vice versa*).

8.2. Latin squares

Latin square designs can be very powerful way to allow rapid assessment of a response to application of a treatment. However, they are susceptible to poor estimation of least square means if there are missing cells. In general, single 2×2 Latin squares with two animals (or experimental units) are not sufficiently powerful to detect statistical differences, although repeated 2×2 Latin squares can, sometimes, overcome this problem. In general, 3×3 Latin squares have similar problems to 2×2 squares, but have the additional limitation that it is not possible to have each treatment follow every other treatment, thereby creating the risk of unequal treatment carryover. Of the Latin square designs used with single animals (or experimental units) per cell, 4×4 designs are by far the most powerful as they can be set up to ensure that all treatments follow all other treatments (and this is the only 4×4 design that should ever be used) in order to equalize treatment carryover effects. However, Latin square designs should generally not be used if substantial carryover effects (relative to the anticipated treatment difference) are likely and are, in general, most sensitive when the individual experimental units (*e.g.*, animals) selected are as similar as possible. Latin squares also have, in general, the limitation that it is not possible to examine period by treatment interactions and so should generally not be used if this is a likely possibility. This raises the issue of appropriate period length, as the chances of period interacting with treatment are reduced as treatment length declines. However, shorter period lengths may not allow full expression of a treatment effect. In general, treatment lengths should not greatly exceed the time anticipated for the treatment to be expressed.

Latin squares can also be used with pens or paddocks of animals. In this case even 2×2 Latin squares can be very powerful if there are numerous animals per pen or paddock, or experimental units, within the pen or paddock (see Section 3). In this case, period length is less of an issue as 'period×treatment' interactions can be examined.

8.3. Separating treatment responses within response parameters

In general, multiple comparison tests are overused within manuscripts submitted for publication to AFST, and are often inconsistent with the stated objective. If the stated objective

is to, for example, examine the impact of four different silage additives on fermentation characteristics of a silage, then the best statistical design will likely be to statistically compare each to the control. However, if the objective is to identify the best silage additive, then each treatment could be compared to each other treatment by a multiple comparison test. However, authors are cautioned that this latter approach often makes discussion of the results difficult, as it becomes necessary to compare each additive to each other additive in the 'Discussion' section. Often it is more sensible to identify the comparisons that one wants to make, and then use orthogonal contrasts to focus on them. This generally results in tables that are more easily interpreted by readers (and reviewers). Tables of numbers with multiple overlapping letters (a, b, c, d, e) superscripts are generally very difficult for reviewers and readers to interpret.

A very common design within manuscripts submitted for publication to AFST is to examine incremental additions of a nutrient, or compound, to a diet or a material in order to identify an optimal level and/or response surface. In such cases, multiple comparison tests are unlikely to be appropriate and polynomial contrasts are generally the most effective statistical test to identify the form of a response to an increasing (or decreasing) level of a treatment. However, when describing a quadratic impact in the text, it is critical that its form be defined. For example, only a significant quadratic effect could lead to a statement such as: 'values were highest/lowest at the intermediate addition level', whereas significant linear and quadratic effects could lead to a statement such as: 'values increased/decreased at an increasing/decreasing rate'.

When using incremental addition of a feedstuff to a ration it is critical that its addition not be confounded with changes in the nutrient profile of the diets.

9. Use of power tests

The use of power tests has become much more common in recent years. Such tests are useful in determining the number of experimental replicates required to detect a desired numerical difference between, or among, treatments prior to initiating an experiment. As such, they can be useful in allocating enough, but not too many, experimental units to meet a goal. In addition, power tests can be useful after an experiment has been completed to examine, and perhaps justify, the number of experimental units allocated to the experiment.

However, statistical analyses of experimental results are merely the means to an end, where 'the end' is to quantify, or understand, the impact of an experimental treatment. It would be unfortunate if statistical considerations, such as power tests, were to result in animal and plant biologists abandoning some fields of experimentation simply because pre-experiment use of power tests suggests that their ability to detect a biologically important difference within individual experiments is low relative to the practical availability of experimental units. In fact, true confidence in quantifying or understanding the impact of an experimental treatment generally comes from examining a body of literature and, in this regard, individual experiments that show no statistical difference due to a treatment can contribute to demonstrating that this impact does in fact occur (or not) when other scientists complete a meta analysis that includes many experiments in a particular area.

Indeed the Co Editors in Chief of AFST encourage authors to submit manuscripts that examine the impact of treatments, or strategies, using appropriate statistical procedures to combine results of numerous published experiments. As our science results in more and more publications of individual experiments, the possibility to contribute to knowledge by 'mining' this body of literature increases.

10. Biological, practical and statistical treatment differences

Manuscripts published in AFST generally focus on identifying a biologically real and/or practically important difference due to imposition of a treatment. Statistical models are useful in reaching that goal. However, authors often seem to forget that numerical treatment differences, that may have biological and/or practical importance, in the absence of statistical support have little more value than statistically significant differences that are so small as to have no conceivable biological or practical value. While power tests, as discussed in Section 9, can be useful in allocating sufficient experimental units to detect a desired numerical difference as statistically significant, they can also prevent allocation of so many experimental units that many trivial numerical differences between treatments reach statistical significance. Then again, because any experiment measures numerous response parameters, it will never occur that there is the same 'ideal' number of experimental units for each response parameter, at least relative to an ability to statistically detect a desired numerical difference.

Ultimately, authors are encouraged to first weigh the importance of the numerical difference between treatments from a biological and/or practical perspective before stressing the importance of a statistically significant difference. Highlighted tabular and figure based 'treatment differences' in manuscript 'Results' sections should generally focus on response parameters that are statistically impacted by treatment and are of sufficient magnitude to be judged to be biologically and/or practically important. Assessment of statistical, biological and practical significance are all subjective to a substantial degree, but all should be weighed before highlighting the difference in the 'Results' section, or building a response mechanism around it in the 'Discussion' section.

11. Use of statistical packages and the statistical options within them

There are several statistical packages that AFST authors frequently utilize, although the most common are the statistical analysis system (SAS), general statistics (GENSTAT), statistical program for the social sciences (SPSS), Minitab and Microsoft EXCEL. The Editors in Chief of AFST take no position on which of these is most desirable in any particular situation, preferring to leave that decision to authors. Indeed, all these packages have strengths and weaknesses and any may provide the most appropriate statistical models for specific circumstances.

However, within these statistical packages there are different general procedures that can be used. The oldest is simply analysis of variance (ANOVA), but general linearized means (GLM), mixed (MIXED) and REML procedures are now all used by scientists that

submit manuscripts for consideration for publication in AFST. Once again, the Co Editors in Chief of AFST take no position on the most appropriate model that is desirable, again preferring to leave that decision to authors. Indeed in many cases of traditional experimental designs, including those discussed in Section 8, where the experiments are balanced and there are no missing values, the ANOVA, GLM and MIXED/REML model analysis outputs will be similar. However, in more complicated statistical models, particularly those that use time as a statistical effect, have missing values and/or animal within group as the statistical unit, MIXED/REML model analysis generally provides a more accurate assessment of true statistical differences due to treatments.

12. Use of several statistical models on the same data set

It is evident that many experimental designs and data sets may have more than one model that is accepted as valid by some statisticians. This often, but not always, relates to available statistical models among statistical packages. In addition, regression analysis between response parameters, or between response parameters and treatments, may suggest significant differences between treatments not suggested by ANOVA, GLM or MIXED/REML procedures. This can lead to 'model shopping' after completion of an experiment to find the statistical model that supports statistical differences among treatments and/or may result in different least square means. While 'model shopping', in this context, could be interpreted as being scientifically fraudulent, it is often extremely difficult to differentiate a quest for the truth from an attempt to support a belief. If in doubt, the Co Editors in Chief of AFST suggest that authors consider presenting the contrasting statistical analyses of the same experiment in the manuscript, and deal with those different interpretations in the 'Discussion' section. While this should, in general, be the exception to the rule, it is certainly an acceptable approach.

13. Final comments

Animal Feed Science and Technology is a scientific journal with a wide international readership as well as a wide variety in the types of experiments that are published within it. In addition, the statistical expertise of authors, reviewers, editors and readers varies widely making it very important that all statistical methods used within manuscripts submitted for consideration for publication are appropriate to the experimental design and are clearly and completely described in the text.

There are many statistical models available, even for the same statistical design, and generally the least complex model that accurately describes the data leading to biologically correct conclusions is the most desirable. Stated differently, the Co Editors in Chief of AFST are suggesting that more complex statistical models are generally only desirable if they are necessary to meet the stated objective or most fully interpret the study. If authors wish to obtain an AFST evaluation of the potential acceptability of a statistical design and/or model prior to initiating an experiment, they are encouraged to contact any of the Co Editors in Chief for an opinion.

It is not possible that a single editorial can deal with all possible combinations of statistical designs and statistical models that can be utilized by all authors in all experiments included in all manuscripts submitted for consideration for publication in AFST. Thus the purpose of this editorial was to outline general statistical principles, as well as highlight some of the more common issues that cause authors, reviewers and editors to disconnect during the review process.

Acknowledgements

The authors express their sincere appreciation to Dr. James Craigon (University of Nottingham) and Dr. James Fadel (UC Davis) for their careful examination of a draft of this editorial from a statistical perspective, and their very carefully considered comments which resulted in significant modification to several areas of the text. The authors also thank Dr. Ed DePeters (UC Davis) for his careful examination of a draft of this editorial from a biological perspective.

P.H. Robinson *
Department of Animal Science,
University of California,
Davis, CA 95616-8521, USA

J. Wiseman [1]
Division of Agricultural and Environmental Sciences,
School of Biosciences, University of Nottingham,
Sutton Bonington Campus, Loughborough,
Leics LE12 5RD, UK

P. Udén [1]
Swedish University of Agricultural Sciences,
Uppsala, Sweden

G. Mateos [1]
Departamento de Produccion Animal,
Universitaria Politécnica de Madrid,
Cuidad Universitaria, Madrid 28040, Spain

* Corresponding author. Tel.: +1 530 754 7565; fax: +1 530 752 0175.
E-mail addresses: phrobinson@ucdavis.edu
(P.H. Robinson),
julian.wiseman@nottingham.ac.uk
(J. Wiseman),
peter.uden@huv.slu.se
(P. Udén),
gonzalo.gmateos@upm.es
(G. Mateos)
[1] Co-corresponding authors.