Makerere University Journal of Agricultural and Environmental Sciences
Vol. 12 (1). pp. 23 - 38, 2023
© Makerere University 2023
ISSN 1563-3721
eISSN: 2958-4795

# Common mistakes and misuse of statistics in agricultural experiments and guidelines on how to avoid them - A commentary

Charamba, V.[1*], Lutaaya, E.[1] and Shihepo, S.[1]

[1]Department of Animal Production, Agribusiness and Economics,
University of Namibia, Private Bag 13188, Windhoek, Namibia

[*]Corresponding author: vcharamba@unam.na

## Abstract

The main objective of this commentary was to highlight some of the common mistakes and misuse of statistics in biological experimental designs, as well as possible ways of addressing these misconceptions. More often than not, researchers seek for guidance on their data analysis after data collection is completed, upon which the Biostatisticians detect poorly conceived experimental designs. Poorly conceived research often results in data analysis and/or results which do not correspond with experimental designs. The advent of ready to use statistical software seems to have both positive and negative benefits to agricultural research. Although it enables efficient data manipulation, processing and analysis, researchers with limited statistical experience are more likely to misuse statistical procedures, and this might lead to erroneous decisions. The validity, reliability and usability of biological research findings is dependent on adherence to statistical protocols and codes of conduct in the designing of experiments, analysis and interpretation of data as well as the conclusions made from the results. Therefore, researchers with limited statistical knowledge are encouraged to seek guidance from Biostatisticians right from the conception stage of their experiments. Researchers need to use relevant statistical analysis and interpretation protocols for their research results to be scientifically valid and usable.

Key words: Data analysis, experimental designs, hypotheses, interactions, *p*-value, statistical errors

## Introduction

During the course of our careers, as biostatisticians in an academic institution, we have observed with dismay investigators, colleagues, undergraduates and postgraduates struggle to analyse their research data with poor appreciation of the statistical rigor required to make valid inferences. We have also encountered numerous proceedings and papers where the statistical inferences and conclusions are questionable given the methodologies and experimental protocols, for example, lack of adherence to the basics of randomization, blocking, inadequate replication and inappropriate statistical analysis approaches. We wish to share these with the readers in this commentary, with the hope that highlighting some of the common flaws we have encountered, will reduce their occurrence and improve the quality of research and make better use of available resources invested to improve our understanding of biological phenomena.

The validity of results from an experiment requires adherence to scientific protocols and codes of conduct. Therefore, it is of ultimate importance that scientists pay attention to the protocols and codes of conduct from problem formulation, setting up of hypotheses, experimental design, data collection and analysis up until the stage of results interpretation. Scientists often apply analysis methods that are not in concordance with the type of data collected as well as the design used for the study. Oftentimes, this may result in scientists making errors when interpreting their research findings, compromising the inference from the results. Researchers often fail to seek assistance at the conception phase of their research experiments, often resulting in common mistakes such as collection of inadequate data due to improper designs. Problematic experiments may hinge from inadequate replication, failure to block or consider clustered populations, and the oftentimes application of wrong experimental designs. Different study designs have strengths and weaknesses (Grabowski, 2016a). The selected design should be the most appropriate to address the study objectives and study materials or resources (Mbotwa *et al*., 2017).

Over the years, the ready availability of easy-to-use and free open source statistical software has resulted in researchers with limited statistical background increasingly misusing or failing to use statistical procedures accurately. Currently, scientists tend to adopt any statistical method that produces a $p$-value, which can be used to draw conclusion on the hypothesis being tested with little thought for compatibility between the study objectives, research design, types of data collected and data analysis. Therefore, understanding the consequences of using an analysis method that is not at par with the study design and the type of data collected; and how they can be avoided is of great importance to the scientific community. This commentary is aimed at highlighting some of the common mistakes and consequences of oversights in the design, and analysis of data in scientific experiments including the lack of use and/or misuse of proper statistical procedures, and offers guidelines on how to avoid them.

## *Common mistakes in experimental design*

### *Poor planning at the conceptual stage of the experiment*

The reliability and validity of scientific findings require a rigorous process from the beginning of the study or experimental design. This may include the definition, number and selection of experimental units, setting up hypotheses, data collection, data analysis methods and interpretation of results. Investigators tend to use inappropriate experimental designs due to lack of insight on efficient designs to accomplish study objectives. Quite often, researchers with limited statistical knowledge bring their data to statisticians for assistance with data analysis, only after the experiment or data collection has been completed. Since some of these researchers have limited knowledge of the statistical design that was applied, they expect the statistician to help them decide and understand the experimental layout and deduce the best statistical analysis procedure from the data. The researchers may also have limited understanding of the hypotheses they are testing, which may invalidate the statistical tests (Zar, 2010), because hypotheses should be stated before collecting the data to avoid subjectivity.

Poor statistical consideration at the conception stage of the experiment seems to be the biggest and most common mistake that scientists have to be aware of so that they avoid reporting and interpreting wrong findings or invalid conclusions. Proper planning enables one to identify the possible sources of variation or confounding effects in the proposed study and choose designs that are more efficient for an enhanced quality of the data to collect. If one has limited statistical and experimental design knowledge, it is important that they engage relevant expertise at the conception stage of the experiment or study. Experimental consideration of issues such as the design structure and treatment structure to be used, the number of replicates as well as random allocation of experimental units to treatments, should be clearly determined at the onset of the experiment for quality data collection (Montgomery, 2017). The design structure refers to the grouping of experimental units into homogeneous groups where necessary and examples are the completely randomized design (CRD) and randomized complete block design (RCBD); while the treatment structure refers to the set of treatment combinations that the investigator intends to study and examples are one-way treatment structure or factorial treatment structure.

Select statistical textbooks (e.g. Hays, 1994; Quinn and Keough, 2002; Gelman, 2012) provide user-friendly guidelines for designing experiments and should be consulted prior to having a meaningful discussion with a statistician on the choice of appropriate designs, especially for those with limited training in Statistics. Failure to identify and account for all the possible sources of variation may result in the use of an inappropriate experimental design and this might in turn result in use of wrong statistical models, which could result in erroneous inferences and invalid conclusions.

*Inadequate replication*

In any experimental design, the number of replicates is a critical decision as it directly affects the validity and reliability of conclusions to be made from the study findings. In designed experiments, replication allows accurate estimation of the experimental error (to reduce and quantify uncertainty) (NSDU, 2023). Through increased replication, an increasingly more precise estimate of the treatment means can be obtained. Replication improves the sensitivity of statistical tests for comparing means. Common mistakes or oversights in scientific studies include incorrect sample-size estimation methods, which might result in under-replication or over-replication.

Determination of the number of replicates required for a study depends on the study design, the minimum size of difference that is desired for detection, the variance, the power of a test or certainty with which the difference is detected, the level of significance and the type of statistical test being performed (Santhoshkumar, 2016; Sharifi, 2017). Under-replication could result in imprecise estimates and lack of statistical power. In addition, a small sample will be unethical as it might put subjects or participants to inferior treatments when results will not be generalizable in the end (Mbotwa *et al*., 2017). On the other hand, over-replication will be a waste of resources and unethical if animal, plant or human subjects are at risk of exposure to weak treatments. As a result, it is imperative that proper replication size determination procedures are utilized in scientific studies. Experiments have been observed to be more on the under-replication side than over-replication (Nelson and Rawlings, 1983), possibly due to time or resource constraints. For example, it is quite common to see a research experiment with only two replicates, which will be inadequate.

Different computations have been proposed and can be applied to determine adequate replication. Kaps and Lamberson (2004), and Montgomery (2017) provide useful guides to the determination of the proper number of replicates. However, some scholars propose that a level of replication that gives an error degrees of freedom (df) of at least 12 is usually considered desirable (EPPO Bulletin, 2012; Santhoshkumar, 2016; Sharifi, 2017) as error df below 12 leads to rapid increase in the critical F values, resulting in decrease in the statistical power of the F-test, thus making it difficult to detect true differences. Some scholars argue that the number of replications should provide at least 10 to 15 degrees of freedom for computing the experimental error variances (Jayaraman, 2000; Mullan, 2021). The more the df for the error, the more powerful the statistical test based on the F-ratios from the analysis of variance (ANOVA). Computer software and online calculators are also readily available via search engines to assist in determining the adequate replication for experiments.

*Inadequate randomization procedures*

Another common mistake in biological research is inadequate randomization in experimental designs. Randomisation in scientific studies removes bias in selection of experimental units and allocation of experimental subjects to treatments; balances the groups with respect to many known and unknown confounding variables, and forms the basis of statistical tests (Suresh, 2011) as it allows the error terms to be independent. Experimental units in one group should not differ in any systematic way from experimental units in another group, as this will bias research results, leading to erroneous inference after data analysis. For example, if older animals are allocated to one treatment, the outcome of the experiment may be influenced by this imbalance. Failure to use proper randomisation procedures could cause certain treatments to be favoured or hampered due to their position in the experimental layout, leading to differences in the precision for different comparisons (Nelson and Rawlings, 1983) and overestimation treatment effects by up to 40% compared to properly randomized trials (Schulz and Grimes, 2002). Although Generalized Linear Models (GLM) are sometimes adjusted for covariance imbalance at the analysis stage, the interpretation of results with the adjusted covariance structure might be difficult as this might result in unanticipated interaction effects (Suresh, 2011). Randomisation might be considered at the stage of assigning treatments to experimental units but thereafter, spurious correlation might be introduced at a later stage of the experiment. For example, in a feeding trial, individual animals might be allocated to experimental units at random, but if animals under one treatment (say diet) are to be housed in one pen, spurious correlation might be introduced if the individual animal is considered the experimental unit, and not the pen or cage. Furthermore, there are studies in which the experiment is a chain of individual steps, which are stand-alone trials. For example, hydroponic fodder might be grown according to an experimental design in a greenhouse and then a second stage is included where the fodder is allocated to different ensiling methods using another experimental design. A good design would incorporate provisions for error control at both stages of the experiment. However, for easier analysis, if the main objective is to compare ensiling methods, then the fodder can be grown under uniform conditions.

*Blocking*

The precision of the data from an experiment depends to a large extent on the experimental technique used to collect the data and the ability to factor in all possible sources of variation including confounding variables. If one knows beforehand that there are some nuisance factors that might affect the results, either due to experimental units or environmental conditions that are not homogeneous, one should try to design the experiment in such a way that the disturbance from the nuisance factors is minimized. Blocking is where by the experimental units are first grouped into similar units, or sets, by making use of prior information about the experimental units, such that homogeneous experimental units are categorized together, to reduce the contribution of the nuisance factors to the experimental error.

Under uniform treatment, the variability in the responses among experimental units in one block will be less than the variation among the blocks, making conclusions more precise as separate conclusions can be made for each block. However, agricultural studies are sometimes poorly blocked, either by omission or commission. Nelson and Rawlings (1983) noted that many "agronomists" do not recognise the value of effective blocking and they block to provide replication and not to control the experimental error. Researchers should try to use uniform experimental subjects/ units if they are available and use appropriate blocking procedures if homogeneous experimental units are not at their disposal.

*Inappropriate analysis methods*

It is a very common and unfortunate practice to inappropriately use readily available statistical software, get *p*-values and conclude, regardless of whether the statistical analysis method is appropriate or not (Grabowski, 2016b). It is also not uncommon in biological research to use inappropriate data analysis methods and err in drawing conclusions even if the experiment was properly designed. In this regard, Mukaka and Moulton (2016) emphasized the need for consistency between experimental designs and analysis methods and discussed how contradictory conclusions can be arrived at depending on a chosen analysis methods. Discrepancies often lead to erroneous conclusions which may not just be unacceptable to the scientific community, but may put animal, plant and human subjects at risk if decisions are to be based on such findings, and thus are unethical.

*Mis-specification of the model*

(i) *Ignoring random effects due to repeated measures designs*

At times, biological researchers are compelled to employ repeated measures designs which involve multiple measurements taken on each subject at different time points.

Repeated measurements are often considered when fewer experimental units are available for experimentation, when the researcher wants to increase the efficiency and sensitivity of the experiment and when the researcher wants to observe changes in experimental subjects over time (Shaughnessy *et al.*, 2015). There are different types of repeated measures designs. Firstly, each experimental unit is subjected to one treatment and measurements are recorded multiple times. For example, a feeding trial where the experimental subjects are given one type of feed over time and the weights are recorded repeatedly during the feeding trial. Secondly, repeated measures might be in the form of cross-over Latin Square designs where the experimental subjects are exposed to all the treatment conditions at different time points. In the second design, each subject functions as an experimental block. These are very common in animal experimentation when the numbers of animals are limited. Each animal receives all treatments at different time periods with a short time interval for adaptation and to ensure the effects of the previous treatment do not affect the current treatment. The second type has an advantage that fewer subjects are required for experimentation and thus cheaper. However, at times there are challenges encountered in analyzing such type of data.

The appropriate statistical approach to analyzing repeated measures data includes the paired-sample t-tests, general linear mixed models or the repeated measures. The linear mixed effects control for correlated errors emanating from data that were collected from the same experimental units at different time points, which might result in muddled statistical inference if not accounted for. However, common practices for analyzing such biological data sometimes ignore these random effects and is probably the biggest issue in inference in experimental biology (Walker, 2020). Other challenges in analyzing such type of data include considering treatments as repeated measures even if the units were subjected to the same treatments over time. In addition, significant effects for the repeated measures does not mean the treatments are significantly different. Littell *et al.* (1998) have elaborated on the appropriate statistical procedures for analyzing data from repeated measures designs.

(ii) *Ignoring random effects due to sub-sampling*

Biological observations are usually taken from the experimental units themselves. However, sometimes it might be impossible to measure the entire experimental unit (Kaps and Lamberson, 2004) or more observations are desired or convenient. This sometimes leads to researchers sampling from the experimental units, resulting in a phenomenon termed sub-sampling. Erroneous inference may be arrived at if researchers utilize ordinary ANOVA models with a single error term for analysis, as these assume that the subsamples within a given experimental unit are independent. The appropriate ANOVA model should have a sampling error in addition to the usual experimental error (MSE) to control for correlations among measurements sub-sampled from the same experimental unit. In the case of a completely randomized design with sub-sampling, the appropriate statistical model for analysis will be:

$$Y_{ijk} = \mu + \tau_i + \varepsilon_{ij} + \delta_{ijk}$$

Where: $i = 1 \ldots t$, $j = 1 \ldots n$, $k = 1 \ldots m$

$\mu$ is the overall mean, $\tau_i$ is the fixed treatment effect, $\varepsilon_{ij}$ is the random experimental error for the j<sup></sup>th experimental unit and $\delta_{ijk}$ the random effect for the $k$th sub-sample of the $j$th experimental unit of the $i$th treatment. The $\varepsilon_{ij}$ and $\delta_{ijk}$ are independent random effects that are normally distributed with mean 0 and variances $\sigma_\varepsilon^2$ and $\sigma_\delta^2$, respectively.

For example, in a study to evaluate the effect of substitution of soybean with four levels of dietary *Spirulina platensis* (T1 - 0%, T2 - 5%, T3 - 10% and T4 - 15%) to grower diets on the internal organ sizes of chickens, an investigator may place multiple birds per cage and have six replicates per treatment; hence the cage is the experimental unit. In order to compare the internal organs, if chickens are randomly sampled in each cage for slaughter, then that constitutes sub-sampling and the individual chickens are not replicates but rather subsamples (pseudo replicates). The total variability should be disaggregated in the treatment variability, random error and sampling error such that the one-way ANOVA is as shown in Table 1.

Table 1. ANOVA Table with sub-sampling

| Source | DF | SS | MS | E[MS] |
|---|---|---|---|---|
| Treatment | $t$ -1 | $SS_{Trt}$ | $MS_{Trt}$ | $\sigma_\delta^2 + c_1\sigma_\varepsilon^2 + \frac{\sum_i m_i\tau_i^2}{t-1}$ |
| Error | $\sum_i n_i - t$ | $SS_E$ | $MS_E$ | $\sigma_\delta^2 + c_1\sigma_\varepsilon^2$ |
| Sampling | $N - \sum_i n_i.$ | $SS_S$ | $MS_S$ | $\sigma_\delta^2$ |
| Total | $N$-1 | $SS_{Total}$ | | |

Similar problems are encountered where laboratory samples are done in duplicates or triplicates where more than one laboratory test is performed from the same sample and the results taken as replicates instead of being averaged out. Analysing subsamples as if they were replicates creates false degrees of freedom (df) and may lead to erroneous inferences.

### *Misinterpretation of p-values*

The *p*-value indicates the strength of evidence against the null hypothesis. Investigators sometimes may have inadequate replication of experiments or the size of the difference between means to be detected for a given sample size may be small, or the variance of the sampled population may be large, or indeed there is limited evidence for rejection of the null hypothesis, with the consequence that *p*-values may not reach the preset significance level (usually $\alpha = .05$) and would be reported as trends ($0.05 < P < 0.1$) (Kim and Bang, 2016; Benjamin *et al.*, 2018; Resnick, 2019). We contend that with proper prior planning of an experiment under guidance of a statistician to ensure adequate replication, sample size and uniformity of experimental units, the ambiguity of reporting p-values can be avoided, so that either the results are reported as significant ($P < 0.05$) or not significant ($P > 0.05$). Furthermore, researchers should bear in mind that statistical significance does not equate to biological significance (Quinn and Keough, 2002).

### *Wrong interpretations of results in the presence of significant interactions*

Interaction effects refers to the combined effects of two or more factors on the dependent variable and occur when the effects of one factor are dependent on the levels of the other factor. Interaction effects are common in regression analysis and ANOVA. Interaction effects can be detected by plotting an interaction graph, for example, when the factor levels on the "*x*"-axis are quantitative, in a line graph displaying fitted values on the "*y*"-axis; Factor A is on the x-axis while various lines show Factor B (or vice versa). Parallel lines indicate the absence of interaction effects while deviation from parallelism suggests the presence of interaction effects. Researchers should perform statistical hypothesis testing to detect interaction, as non-parallel lines can be an indication of random sampling error (Stevens, 1999) and hypothesis testing separates real effects from random noise. For example, in a study to evaluate the effect of substitution of soybean with different levels of dietary *Spirulina platensis* grower diets on the nutrient composition of two indigenous chicken breeds (P- Koekoek and Boschveld). If the factor level combinations of breed and *Spirulina platensis* levels have been replicated, then it is possible to include an interaction term in the model and analyze for its significance, that is if the effects of levels of *Spirulina platensis* depend on breed. The results are in Figure 1. In Figure 1(a), the lines for Boschveld and P- Koekoek appear parallel, suggesting the absence of interaction effects on ash content while in Fig. 1(b) there are suggestions that there might be significant breed and *Spirulina platensis* levels interaction on the fecal neutral detergent fibre (NDF) as the lines intersect. To complement these results, the interaction *p*-values for ash and NDF were 0.65 and 0.003, respectively (with 0.09 and 0.55 Cohen effect sizes, respectively) (Cohen, 2013).

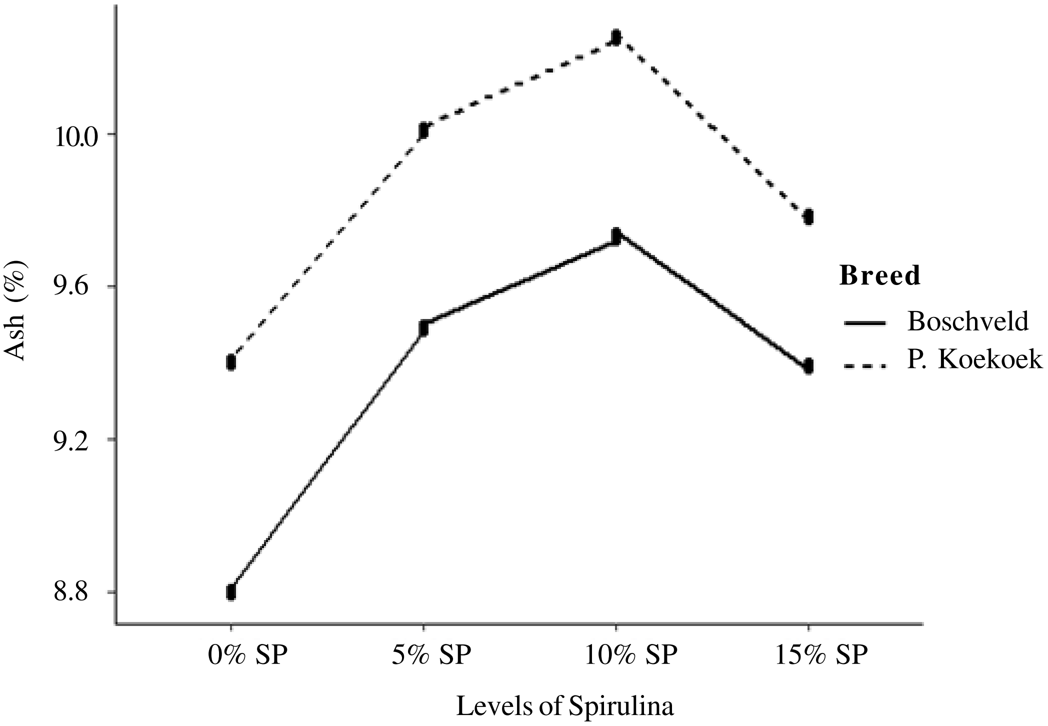

Figure 1(a). Effects of breed and *Spirulina* levels on ash content.

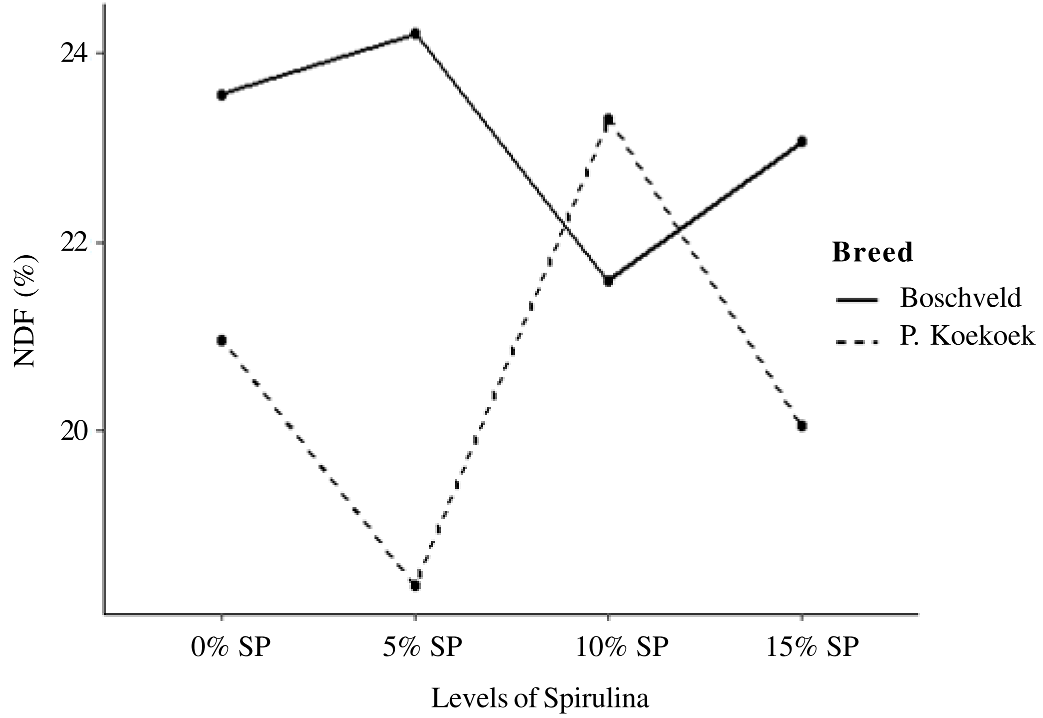

Figure 1(b). Effects of breed and *Spirulina* levels on NDF content.

Interpretation of results in the presence of interaction effects seems to be a challenge in many agricultural studies. If interaction effects are not significant, they may be dropped from the model and interpretation of results should dwell on main effects only. If interaction effects are significant, it means comparison for main effects become meaningless and only interaction means are worthy of comparison. When interpreting interaction means, it should be clearly stated how one level of a factor behaves differently with the other levels of the other factor. However, those who compare factor level combination means may be silent about this.

*Failure to perform model diagnostics checks before using the statistical model for inference*

Most statistical procedures for analysis make assumptions on the type of data, the scale of measurement and the underlying distribution. These assumptions on the data determine whether we can use parametric or non-parametric procedures. One common mistake is the application of statistical methods to wrong type of data measurement scales, failure to perform relevant data transformations, and failure to use appropriate non-parametric methods when one should. The validity of most statistical models is based on meeting the underlying model assumptions. The assumptions in the analysis of experimental data using parametric methods include: (1) the relationship between variables should be additive (ANOVA) or linear; (2) data in each comparison group should show a normal or Gaussian distribution; (3) data should exhibit homogeneous variances; (4) observations should be independent; and (5) independent variables should be independent in multiple linear regression (MLR), a violation of which is coined multi-collinearity (Montgomery, 2017; Zar, 2010). If model assumptions are not met, the validity of results can be compromised. It is thus a good and expected practice for researchers to perform model diagnostics checks to verify if the underlying assumptions have been met before using the model for inference and decision making. However, it is a common practice in biological research to apply ordinary data analysis methods and go ahead and make inference without carefully studying the patterns of variation in the data or check if the underlying assumptions have been met and perform necessary adjustments or transformations.

R-squared values, model adequacy *p*-values, the Akaike information criteria (AIC: Akaike, 1974) and the Bayesian Information Criteria (BIC: Schwarz, 1978), the Deviance Information Criteria (DIC: Spiegelhalter *et al.*, 2002), the Widely Applicable Information Criteria (WAIC: Watanabe and Opper, 2010) for Bayesian estimation and the log-likelihood ratios can be utilized to test the overall model adequacy. Graphical plots can be used to visually explore data to check if model assumptions have been met. In regression analysis, scatter plots can be used to explore the pattern of relationship that exist between the variables before selecting the best analytical method (linear, logarithmic, exponential, polynomial, sigmoid, no relationship) or data transformation (square root, reciprocal, logarithmic) that can make the relationship linear. In addition, scatter plots can check highly correlated independent variables in MLR.

Box and whisker plots can be used to examine symmetry and the amount of variance in the data as well as identifying outliers. If the data is highly skewed, then median and interquartile range maybe the best measures of central tendency and dispersion respectively, instead of the mean and standard deviation which are appropriate for symmetric data. Quantile-quantile (QQ) plots of predicted against observed and histogram of residuals can also be used and formal tests such as the Shapiro-Wilkis' test can be used to test for the assumption of normality.

To ensure independence, experimental units should be randomly selected and randomly allocated to experimental treatments. The graphical plot of residuals versus time or observation order and the Durbin-Watson tests can be used to test the independence assumption. The homogeneity of variance assumption can be ascertained by graphically plotting residuals versus predicted values. If there is no obvious pattern, then the variances are stable. Formal tests such as the Levene's test, the Bartlett's test can be employed.

In MLR models, multi-collinearity can be detected through bivariate correlations of explanatory variables that are at least 0.7, very high standard errors for regression coefficients, overall model significance with all model parameters non-significant, large changes in parameter estimates upon variable removal or addition, coefficients with signs that contradict theory, coefficients that differ with wide margins for different samples and high variance inflation. When detected, multi-collinearity can be rectified by removing some of the independent variables that are highly correlated. Alternatively, linearly combine variables that are highly correlated to form new variables that are fewer and not correlated using statistical data reduction procedures such as principal components analysis and factor analysis. Ridge regression and partial least squares regression can also handle multi-collinear variables.

Generally, if the model assumptions have not been met, alternatives include: (1) transformations of the response variable to either make it normal or to make variances homogeneous or to remove or reduce non-additive interactions (Falconer, 1989); (2) if transformations do not work, use non-parametric statistical methods such as the Median test and the Mann-Whitney-Wilcoxon test for two independent samples; the Wilcoxon paired ranks test for two matched samples; the Kruskall-Wallis for $k$ independent samples and the Friedman test for dependent samples. Log transformations can be used when treatment effects are multiplicative rather than additive and when there is a proportional relationship between the mean and the standard deviation. The coefficient of variation (CV) gives an indication of the degree of departure from normality and a useful guide is to transform variables when the CV exceeds 20% (Falconer, 1989). The square root transformation is useful when variances are proportional to sample means and to normalize the Poisson distribution. The arcsine transformation is used for square root of proportions which are the basis of the binomial distribution. The reciprocal transformation is used for data showing a sigmoid curve (e.g. growth curves). The transformation is used when the standard deviations of treatments are proportional to the square of means.

## Conclusion

Correct experiment design relies on the basic principles in statistics of replication, randomization and blocking and these should be borne in mind for any experimental study. However, even for correctly designed and collected data, invalid statistical analysis procedures can lead to erroneous inferences. The advent of readily available statistical software has led to many researchers with limited statistical background wrongly applying some statistical methodologies for analysis as long as they get a *p*-value to use for inference. Though not fully exhaustive, the commentary gives an overview of some of the common pitfalls noted in analysis of data arising from biological research and provides guidelines on the correct procedures to be followed to improve on the quality of statistical analyses and hence inferences. The failure of researchers to properly design research experiments, match the study design, data type and research objectives with the appropriate data analysis method is caused *inter alia* by inadequate training on statistical methods and experimental designs. Furthermore, negligence and lack of coordination between co-investigators, or failure to include co-investigators with statistical knowledge in research teams as well as looking for assistance on data analysis when the experiment is done and data has already been collected. Thus, researchers with limited statistical background are advised to seek advice from statisticians at the conception phase, as it is difficult to remedy poorly designed research. In addition, researchers are advised to adequately replicate and randomize their experiments, and use statistical analysis procedures that best suit their designs and data types. Although the *p*-value is a useful statistical tool in hypothesis testing, researchers are advised to complement their findings with estimation of effect sizes for their results to be more conclusive. It is important to perform model diagnostic checks to check if model assumptions have been met before using the model for inference to avoid dire consequences of making wrong conclusions from wrong models. Researchers are advised to seek training on statistical methods, include co-investigators with statistical background in their research teams and seek statistical guidance at the conception stage of the project, so as to improve the quality of statistical inferences. Refresher courses on key statistical concepts may seem warranted for investigators and to keep abreast with new analytical tools including the widely available open source R software.

## Acknowledgements

Ms. Aino Nambahu and Mr. Samuel Hangula for their project data on *Spirulina platensis* grower diets that was used to illustrate presence of interaction effects.

## References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE transactions on Automatic Control* 19:716–723.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L. and Camerer, C. 2018. Redefine statistical significance. *Nature Human Behaviour* 2:6–10.

Cohen, J. 2013. Statistical power analysis for the behavioral sciences 2nd (ed) Lawrence Erlbaum Associates Hillsdale.

EPPO Bulletin, 2012. Design and analysis of efûcacy evaluation trials. European and Mediterranean Plant Protection Organization, *Eficacy Evaluation of Plant Protection Products* 42: 367–381. https://doi.org/10.1111/epp.2610

Falconer, D., 1989. Introduction to quantitative genetics 3rd (ed.) Longman Scientifical Technical.

Gelman, A. 2012. *The inevitable problems with statistical significance and 95% intervals*, *Statistical Modeling, Causal Inference, and Social Science*, < http://andrewgelman.com/2012/02/02/the-inevitable-problems-with-statistical-significance-and-95-intervals/>

Grabowski, B. 2016a. "P< 0.05" might not mean what you think: American Statistical Association clarifies P values. JNCI*: Journal of the National Cancer Institute* 108.

Grabowski, B. 2016b. Misinterpretation and misuse of P values. misinterpretation-misuse-p-values-research-statistics. URL https://blog.oup.com/2016/09/misinterpretation-misuse-p-values-research-statistics/ (accessed 12.28.21).

Hays, W.L. 1994. Statistics. 5th (ed.). Harcourt Brace, New York.

Jayaraman, K. 2000. A Statistical Manual for Forestry Research, Forestry Research Support Programme for Asia and The Pacific. Food and Agriculture Organization of The United Nations Regional Office for Asia and The Pacific, Bangkok.

Kaps, M. and Lamberson, W. 2004. Biostatistics for Animal Science. CABI Publishing, Wallingford, UK.

Kim, J. and Bang, H. 2016. Three common misuses of P values. *Dental hypotheses* 7:73.

Littell, R.C., Henry, P. and Ammerman, C.B. 1998. Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science* 76:1216–1231.

Mbotwa, J., Singini, I. and Mukaka, M. 2017. Discrepancy between statistical analysis method and study design in medical research: Examples, implications, and potential solutions. *Malawi Medical Journal* 29:63–65.

Montgomery, D.C. 2017. Design and analysis of experiments. John Wiley & Sons.

Mukaka, M. and Moulton, L. 2016. Comparison of empirical study power in sample size calculation approaches for cluster randomized trials with varying cluster sizes andndash; a continuous outcome endpoint. *Medical Statistics* 6.

Mullan, W.M.A., 2021. Calculator for determining the number of replications of an experiment required to achieve the desired degrees of freedom. https://www.dairyscience.info/newcalculators/dof.asp

Nelson, L.A. and Rawlings, J.O. 1983. Ten common misuses of statistics in agronomic research and reporting. *Journal of Agronomic Education* 12:100–105.

NDSU, 2023. Planning experiments, available at: https://www.ndsu.edu/faculty/horsley/Pln_expt.pdf. (Accessed May 2023).

Quinn, G.P. and Keough, M.J. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press; 1st edition (April 2, 2002); ISBN-10: 0521009766; ISBN-13: ý 978-0521009768.

Resnick, B. 2019. 800 Scientists Say it's Time to abandon Statistical Significance. Statistical-significance-p-values-explained. URL https://www.vox.com/latest-news/2019/3/22/18275913/statistical-significance-p-values-explained (Accessed 7.18.21).

Santhoshkumar, A. 2016. Most of scientists tell that the Error degree of freedom for any agricultural design should be >12, Why so? ResearchGate. URL https://www.researchgate.net/post/Most_of_scientists_tell_that_the_Error_degree_of_freedom_for_any_agricultural_design_should_be_12_Why_so/57359b3696b7e4cf3b568231/citation/download. (Accessed 11.14.21).

Schulz, K.F. and Grimes, D.A. 2002. Allocation concealment in randomised trials: defending against deciphering. *The Lancet* 359:614–618.

Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics* pp. 461–464.

Sharifi, R., 2017. Most of scientists tell that the Error degree of freedom for any agricultural design should be >12, Why so? ResearchGate. URL https://www.researchgate.net/post/Most_of_scientists_tell_that_the_Error_degree_of_freedom_for_any_agricultural_design_should_be_12_Why_so (Accessed 11.14.21).

Shaughnessy, J.J., Zechmeister, E.B. and Zechmeister, J.S. 2015. Research Methods in Psychology, 10th (ed.) Mc_Graw Hill Global Education Holdings L.L.C, Online.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology)* 64:583–639.

Suresh, K. 2011. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences* 4: 8.

Walker, J.A. 2020. Elements of Statistical Modeling for Experimental Biology. URL https://www.middleprofessor.com/files/applied-biostatistics_bookdown/_book/Walker-elementary-statistical-modeling-draft.pdf (Accessed 11.14.21).

Watanabe, S. and Opper, M. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11.

Zar, J.H. 2010. Biostatical analysis. 5th (ed), Pearson Education, Inc., Upper Saddle River, New Jersey, USA.