Animal Behaviour xxx (2015) 1–12

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Special Issue: Conservation Behaviour

A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons

László Zsolt Garamszegi*

Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC, Seville, Spain

ARTICLE INFO

Article history:
Received 19 March 2015
Initial acceptance 14 September 2015
Final acceptance 2 November 2015
Available online xxx
MS. number: SI-15-00224

Keywords:
bias
bootstrap
confidence interval
effect size
hierarchical data structure
nonparametric statistics
null hypothesis testing
precision
statistical power
study design

Here, I provide a practical overview on some statistical approaches that are able to handle the constraints that frequently emerge in the study of animal behaviour. When collecting or analysing behavioural data, several sources of limitations, which can raise either uncertainties or biases in the parameter estimates, need to be considered. In particular, these can be issues about (1) limited sample size and missing data, (2) uncertainties about the identity of subjects and the dangers posed by pseudoreplication, (3) large measurement errors resulting from the use of indicator variables with nonperfect reliability or variables with low repeatability, (4) the confounding effect of the within-individual variation of behaviour and (5) phylogenetic nonindependence of data (e.g. when substitute species are used). I suggest some simple analytical solutions to these problems based on existing methodologies and on a consumable language to practitioners. I highlight how randomization and simulation routines, generalized linear mixed models, autocorrelation models, phylogenetic comparative methods and Bayesian statistics can be exploited to overcome the inefficient performance of some conventional statistical approaches with typical behavioural data. To enhance the accessibility of these methodologies, I demonstrate how they can be brought into practice in the R statistical environment, which offers flexible statistical designs. Although the primary motivation behind this discussion was to help animal behaviourists who address questions in relation to conservation, I also hope that researchers working on the evolutionary ecology of behaviour will also find some material useful.

In different scientific disciplines, the investigated research topics and the attributes of the studied subjects set up specific constraints for study design and the statistical analysis of available data. Here I aim to discuss some of difficulties that can typically emerge in the study of animal behaviour and to offer some statistical approaches that can be used to alleviate the limitations embedded in behavioural data. According to the theme of this Special Issue, I will principally focus on issues that emerge in association with species of conservation concern (i.e. those that can be characterized by low or rapidly declining abundance, by high level of specialization to extreme environments, or by having a specific position on the phylogenetic tree). However, as study designs relying on behavioural observations on marked individuals impose some general challenges for the analysis of data independently of the particular research questions, most discussed topics can be viewed more broadly and easily applied to studies with ecological or evolutionary focus. The first three topics discuss concrete problems (limited sample size, the use of surrogate variables when noninvasive studies, unknown identity of subjects) that may typically emerge when working with behavioural data that are constrained for ethical or practical reasons. In the last section, I bring into the focus other approaches (mixed modelling, phylogenetic comparative methods, Bayesian statistics) that could, in general, be more broadly applied in conservation studies.

Given the target audience and the purpose of this paper/journal volume, I provide a superficial overview on a broad array of approaches rather than cover only a few topics with the appropriate statistical deepness (i.e. with complex equations and simulations). This is also because I present nothing new here: all of the methodologies I touch on are already well established in the statistical literature. By maintaining a focus and language that are comprehensible to practitioners, my sole aim is to initiate the spread of a diversity of methodologies that are currently underexploited in the study of animal behaviour and conservation. However, I do emphasize the importance of the scientific foundations for any methodology being implemented in practice. Hence, for the more interested readers, I give pointers to the primary sources that

* Correspondence: L. Z. Garamszegi, Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC, c/Americo Vespucio, s/n, 41092, Seville, Spain.
E-mail address: laszlo.garamszegi@ebd.csic.es.

http://dx.doi.org/10.1016/j.anbehav.2015.11.009
0003-3472/

contain the corresponding mathematical background. For those who wish to try the methodologies with their own data, I provide an electronic supplement that includes several executable statistical scripts written in the R statistical environment (R Development Core Team, 2015) for the simplest scenarios (i.e. those that are covered in the first three sections). For demonstrative purposes, I use illustrations and examples that rely on elementary statistical situations (e.g. correlations, linear regression with a single predictor), but most of the recommended methodology can be easily tailored to more complex statistical designs. Note that this overview is not intended to be exhaustive, it merely reflects the perspective and knowledge of the author. Problems and solutions that are not discussed here are possible.

LIMITED SAMPLE SIZE

Limited sample size is one of the most obvious constraints that confronts animal behaviourists (Taborsky, 2010), especially when working on conservation-related issues (Bradshaw & Brook, 2010; Martinez-Abrain, 2014). For a variety of reasons that arise from the special characteristics of the studied species, in combination with the difficulty of assaying behaviours and ethical policies, it is impossible to acquire an ideal sample that would be representative of the real world. This is a general problem in the study of animal behaviour, but it is particularly important when working with species of conservation concern. These are typically those species that are at low abundance, difficult to observe in nature, impractical or even illegal to capture and unable to be brought into the laboratory for experimentation. Furthermore, most conservation-related questions target population-specific parameters (e.g. abundance, species composition) and their temporal or spatial patterns. These tasks necessitate comparisons across higher group levels with a sample size that is equal to the number of groups being compared. Therefore, effective sample size in conservation studies is severely curtailed, and conservationists occasionally have to work with an extremely small sample size.

This sample-size limitation brings up statistical issues about precision, accuracy and stability (Quinn & Keough, 2002). Low sample size has the statistical consequence that the chances of obtaining a reliable and appropriate estimate of the central tendency (e.g. mean or median), data spread (e.g. variance or standard deviation, shape of the frequency distribution) and the strength of relationship between variables (correlation, between-group differences, regression slopes) are low. Under these circumstances, the ability to tease pattern and noise apart without bias becomes progressively intractable. In a null hypothesis testing (NHT) framework, this problem is typically manifested as limited statistical power signifying that high type II error rates make it very likely that the null hypothesis cannot be rejected even if it is false (Cohen, 1988). More generally, data limitation translates into imprecise parameter estimates meaning that central tendencies can be obtained with very large confidence intervals, which is a considerable shortcoming even in a non-NHT framework (Nakagawa & Cuthill, 2007). In terms of accuracy, some statistical approaches are known to perform badly and provide parameter estimates with a systematic upward or downward bias when supplied with limited data (Bishara & Hittner, 2015; Gorsuch & Lehmann, 2010). A related point is that, because of the strict relationship between the number of parameters and the sample size that can be entered into a statistical model (Bolker, 2007), an observer cannot achieve full control on several potentially confounding variables when data are limited, which can also generate biases. Finally, questions about stability appear via the relative importance of particular data points, as the influence of a single outlier can be drastic in a small sample. Accordingly, small changes in the data can lead to substantially different results, challenging the reliability of the obtained parameter estimate. Note that errors arising from low sample sizes can reach beyond these traditional problems for accuracy and precision, as sign errors and exaggeration errors can also emerge (Gelman, 2015). Furthermore, low statistical power as caused by limited data has consequences for the reproducibility/replicability of results (Button et al., 2013).

The traditional way to circumvent at least some of the above caveats is to use simple statistical methods (such as *t* tests, correlations, Fisher exact test) that have been demonstrated to perform convincingly well when sample sizes are small (Larntz, 1978; Soper, Young, Cave, Lee, & Pearson, 1917; de Winter, 2013). Furthermore, some textbooks recommend the use of nonparametric statistics in such situations (Siegel & Castellan, 1988). However, these approaches offer practical solutions only, as issues about the precision, the role of influential data points and the need for controlling for other variables are treated only partially or remain completely unresolved.

Effect Size Thinking: towards Separating Strength from Precision

When data are limited, several confusions may arise from the NHT-based inference of results (Cohen, 1994; Stephens, Buskirk, & del Rio, 2007). Most of the weaknesses revolve around the fact that small samples inherently incur low statistical power; thus, it is highly likely that effects of small or intermediate magnitude (which could still be of biological importance) remain nonsignificant. Given that the NHT-framework enforces binary decisions about the existence or nonexistence of effects, nonsignificant results are often interpreted as evidence for no biological relationship between the investigated variables. This misleading scientific conclusion is based on too much attention to *P* values, which can generate at least two problems for conservation biology. First, if an effect of a small or intermediate magnitude appears nonsignificant in an NHT-based study and is inferred as being biologically unimportant, such a scientific verdict may lead to an omission of an effect from the practical side as well (e.g. a pollutant has no detected effect, thus no actions are needed against it). This is particularly dangerous if the investigation involves a threatened species that is very hard to study. In that situation, the replication of a given study is not warranted, and the same null results can be repeatedly used as a motivation for a wrong conservation action. Second, nonsignificant results are difficult to publish, and thus often remain in file-drawers and generate publication bias (Møller & Jennions, 2001; Rosenberg, 2005). If policy makers rely on published information for their decisions, they will design their action plans following a biased picture from the published material (i.e. the efficiency of a prevention campaign is overestimated if only supportive studies are getting published). Therefore, drawing strong conclusions with practical importance from small samples and based on significance levels should be avoided. Scientists working with species of conservation concern have a high responsibility to publish their results, even if these are not significant.

Effect size thinking may offer a straightforward alternative to the NHT-based inferential approach (Garamszegi, 2006; Nakagawa, 2004; Nakagawa & Cuthill, 2007; Nakagawa & Santos, 2012; Thompson, 2002). The most important drawback of focusing on *P* values is that they combine statistical power and the magnitude of the underlying effect (in extreme scenarios this leads to the problem that everything will appear significant when sample sizes are very large, but nothing will appear significant when sample sizes are very low). Effect size theorem, on the other hand, separates these properties, as it relies on different metrics to describe the strength of the biological effect and the uncertainty by which it can be measured from the available data. Most biological questions

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

could be tackled through some parameter estimations from statistical outputs, and most of these can be used to calculate unstandardized (e.g. mean, slope of a regression) or standardized (e.g. correlation coefficient, Cohen's *d*, odds ratio) effect sizes that can be used for making judgements about the magnitude of the effect under investigation. In parallel, parameter estimations usually come with a confidence range around the parameter estimate that reflects the limitations of the data and the statistical model, which can be used to make inferences about the precision obtained effect size (Nakagawa & Cuthill, 2007). In fact, because of inherent biases for many statistics at low sample sizes, it becomes essential to derive an appropriate estimate for the confidence interval around the parameter of interest, which can be achieved by various means (Thompson, 2002). Therefore, objective scientific results ideally consist of both effect sizes and the associated confidence intervals, while *P* values and NHT-motivated binary decisions should be prevented. Results in this form can be more appropriately interpreted for conservationist to assess risk factors, while they also preclude publication bias. More discussion on the philosophy of effect size thinking and calculation methods (some of these are adjusted for small sample sizes) can be found in Nakagawa and Cuthill (2007). Fig. 1 demonstrates how a nonsignificant correlation (Fig. 1a) can be translated into an effect size and a confidence interval (Fig. 1b).

Simulation-based Inferences from Small Samples

Another factor that frequently limits the applicability of different statistical tests in combination with small sample size is the inability to convincingly validate certain assumptions that the applied statistics needs to fulfil (Sokal & Rohlf, 1995; Zar, 1996). Most importantly, with severely limited data there is no way to evaluate whether observations are drawn from a normally distributed population, which is a critical assumption for most parametric tests (but see Rasch & Guiard, 2004, as an example for the robustness of some parametric methods to non-normal data). If our sample is large enough, one can reasonably meet this assumption via investigating the distribution of data within the sample. However, in cases of small sample sizes it is impossible to statistically verify assumptions about normality (e.g. none of the known normality tests will be significant because of the power issues discussed above, while graphical methods will also perform deceptively). In this case, distribution-free (i.e. nonparametric) statistical methods seem adequate, as these will liberate the investigator from assumptions regarding the distribution of variables (Siegel & Castellan, 1988). However, the use of nonparametric tests will also hamper constraints in terms of further reducing the statistical power, as information loss occurs when variables are brought onto the rank scale.

To circumvent these problems, the randomization- or simulation-based approximation of confidence ranges for parameter estimates may be useful (Fig. 1). For example, bootstrapping is a resampling technique through which samples are repeatedly drawn with replacement from the original sample (usually with the same sample size) (Efron & Tibshirani, 1993; Manly, 1991). Then, the distribution of the derived metric (e.g. mean, correlation coefficient or other effect size measure) in the bootstrap samples can be inferred as the probability distribution of the same parameter in the population. Therefore, extracting the 95% quantile range of the parameter from a large number of bootstrap samples (usually 1000 or higher) will define the range of values that encompass the 'true' value in the population with 95% probability (i.e. the 95% confidence interval). Applying a similar philosophy, based on the properties of the observed data (e.g. mean and variance) and the parameter estimate from the test statistics, one can simulate a large population of data to imitate how the real world would look if the given parameters were in effect. During data simulation, different scenarios about data distribution can be considered (e.g. the population can be generated based on various distribution functions) and examined in the subsequent sampling and analytic phases. The latter two steps consist of sampling from the simulated world (e.g. following original sampling scheme) and the re-estimation of the parameter of interest from it. After a large number of iterations, the probability distribution of parameters across different simulations can be used to define the 95% confidence range.

A particular difficulty emerges for small samples in terms of the instability of the results. This is because the influence of single data points is relatively high, and thus there is a high probability that a single outlier can mediate an observed relationship. To appropriately investigate whether the obtained results are sensitive to such influential data points, one can perform a jackknife procedure, by which individual data points are excluded one by one, and the parameters of interests are re-estimated in all of these reduced samples (Atkinson, 1985). The stability of results can then be expressed through the range of parameter estimates across the jackknife samples (i.e. if these ranges are relatively wide, one

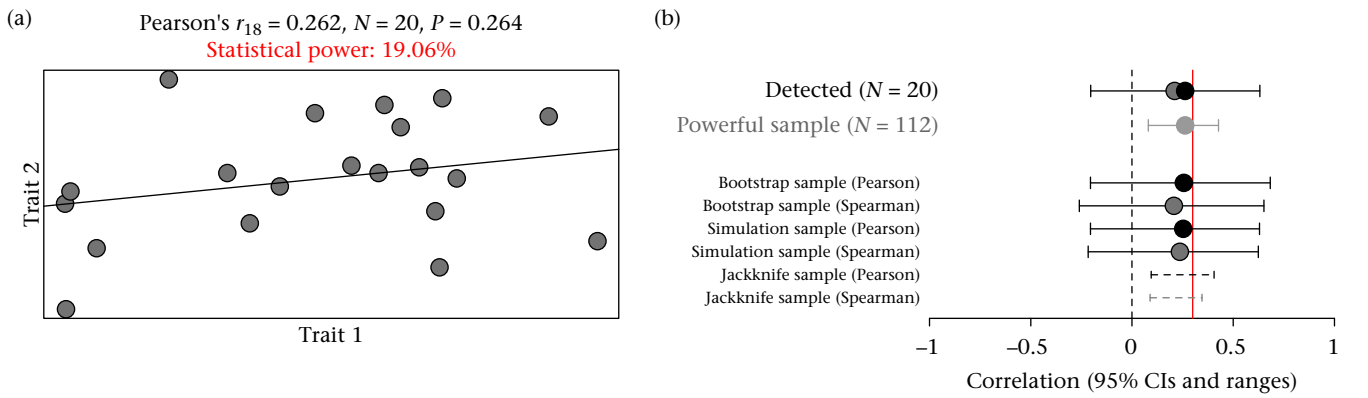

Figure 1. Inferences from small sample sizes. (a) Relationship between two variables showing the corresponding regression line and results from the Pearson correlation test with the post hoc calculated statistical power. (b) Mean effect sizes and confidence intervals obtained by different randomization/simulation procedures (see text) and analytical approaches: e.g. Pearson correlation (black dots); Spearman correlation (grey dots). Shown are the effect sizes calculated from the observed data (N = 20) and the confidence interval for a powerful (80%) sample (N = 112, as determined in a power analysis). The solid vertical red line shows that correlation structure, under which the 'observed' data to be analysed were simulated. The corresponding R codes are available in the Supplementary Material.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

should consider that some particular data points might have remarkable effects on the results and seek biological reasons behind such influential points).

Issues about accuracy, precision and stability are also relevant and need to be investigated for more complex statistical models (such as generalized linear models, mixed models). However, it is also crucial to investigate the additional assumptions that these models make (e.g. independence and homogeneity of residuals, absence of collinearity among predictors, appropriate sample size/number of terms ratio) through model diagnostic exercises (Freckleton, 2011; Gałecki & Burzykowski, 2013; Mundry, 2014; ; Zuur, Ieno, & Elphick, 2010), which potentially embodies additional challenges for small sample size situations. However, it is relatively easy to perform the above simulation and resampling approaches based on the consideration of many parameters provided by such complex models. Moreover, it is very efficient to estimate confidence intervals by applying a parametric bootstrap (Efron & Tibshirani, 1993). This refers to the iterative generation of the response variable based on the prediction of the fitted model, and the subsequent parameter estimations that are gained from the models that have the same combination of predictors and are being refitted to the simulated response in different rounds.

In Fig. 1, I demonstrate how these resampling and simulation-based approximation methods can be applied for the assessment of correlation between two variables with limited sample size. In this demonstration, I also include a comparison between the performance of parametric (Pearson correlation) and nonparametric (Spearman correlation) methods. The underlying R codes are available in the Supplementary Material, which can be easily tailored to other statistical/biological questions. Presenting results in such a way can maximize the information that can be gained from a small sample and would provide wildlife managers with an objective picture about our ability to assess a risk factor from available data. For effective protection programmes, instead of focusing on significance levels, conservation actions should be motivated by effect sizes and the associated confidence intervals as estimated by different resampling/simulation methods.

If the investigator is uncertain about writing statistical codes, s/he can also consult with published simulation studies that investigate the performance of different statistical models under different sample size constraints (Fox, 2008; Martin, Nussey, Wilson, & Réale, 2011; Ruxton, 2006; van de Pol, 2012). These reference materials provide some rules of thumb for different sampling scenarios that can be considered even before running any statistical tests. If these simulation studies showed poor performance in terms of a biased parameter estimate for a given sample size, model outcomes should not be trusted with too much confidence.

NONINVASIVE SAMPLING METHODS: MEASUREMENT ERROR FOR SURROGATE VARIABLES

It is often required to use noninvasive methods for sampling when the model species is to be protected, but this is also becoming preferable practice in the modern study of animal behaviour for ethical reasons (Mench, 2000). Several methods are now available that permit the estimation of certain physiological traits (such as hormone levels, parasite load, heart rate, metabolic status, immunocompetence) through surrogate variables that help reduce or completely eliminate the stress that animals experience during handling and/or capturing (Goymann, Möstl, & Gwinner, 2002; Kersey & Dehnhard, 2014; Lenz, Wells, Pfeiffer, & Sommer, 2009; Lukas et al., 2004; Narayan, 2013; Pereira, Duarte, & Negrao, 2005; Taberlet & Luikart, 1999; Williams, Greenhalgh, & Manning, 2003). Noninvasive methods are also applicable to behavioural variables when assay procedures are designed in a way that causes less suffering to the animals. The spread of such sampling philosophies has obvious benefits for animal welfare and conservation, but using surrogate variables instead of measuring the focal variables directly might have statistical consequences. Neglecting these may lead to wrong biological conclusions and the development of inappropriate protection programmes. Particularly, noninvasive methods can be characterized by increased error rate, as the focal variable can only be estimated by increased uncertainty. This consequence calls for consideration of reliability and measurement errors.

Reliability and Repeatability

The use of a surrogate variable is only advisable if it can be measured with high precision and its variation reliably predicts variation in the focal variable as well, premises that need to be tested in well-designed validation (pilot) studies (Taberlet & Luikart, 1999). For example, using a repeated sampling set-up, the surrogate variable can be assessed multiple times for the same biological entity (e.g. metabolite concentration from multiple droppings, parasite counts from different parts of the body), which can be used to estimate the repeatability of the measured trait. Repeatability has been discussed in detail in the literature and several calculation methods are available for various test situations (Lessells & Boag, 1987; Nakagawa & Schielzeth, 2010; Wolak, Fairbairn, & Paulsen, 2012). In brief, repeatability describes the proportion of the between-subject variance relative to the total variance (typically labelled as phenotypic variance that combines within-subject and between-subject variances). It ranges from 0 to 1, and if it is close to the lower boundary it indicates a role for large error variance (if within-subject variance is coming from the repeated measurement of the same subject), while repeatability estimates around 1 signify high precision.

Furthermore, reliability can be assessed by measuring the focal trait and the candidate surrogate variable in parallel in the same individual and then testing the correlation between them (e.g. Bauer, Palme, Machatschke, Dittami, & Huber, 2008; Lane, 2006; Piggott, 2004; Touma & Palme, 2005). From a statistical perspective, a good indicator should strongly correlate with the focal variable and offer high explanatory power. These properties, expressed as correlation coefficients or the proportion of variance explained, also vary on a scale between 0 and 1, and the validation should provide metrics close to the upper margin. Note that repeatability and reliability are closely related. A surrogate variable cannot correlate strongly with other variables if the surrogate variable has high measurement error and thus low repeatability (in other words, repeatability sets up an upper limit on reliability). Therefore, if it is feasible, in a pilot study based on a smaller sample of individuals (e.g. $N = 20$; Fig. 2a), one can assess the strength of the relationship between the candidate indicator trait and the focal variable. If such a validation proves a strong correlation (e.g. 0.7–0.9), the candidate may be convincingly used as a surrogate variable in future field studies directing more concrete biological questions (note that when repeatability/correlation is high enough, even small samples can provide accurate and precise estimates sensu simulations in Dingemanse & Dochtermann, 2013). In the opposite case, one can conclude that the chosen trait is not a good indicator either because it does not strongly predict the focal trait, or because it cannot be measured with high precision. If the latter reason is suspected, it might be desirable to perform another pilot study with improved assay methodology and calculate repeatability in the hope that the desired predictive value can be regained with better efficiency. In any case, for the appropriate biological interpretations, in the focal analyses, it is necessary to consider how

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

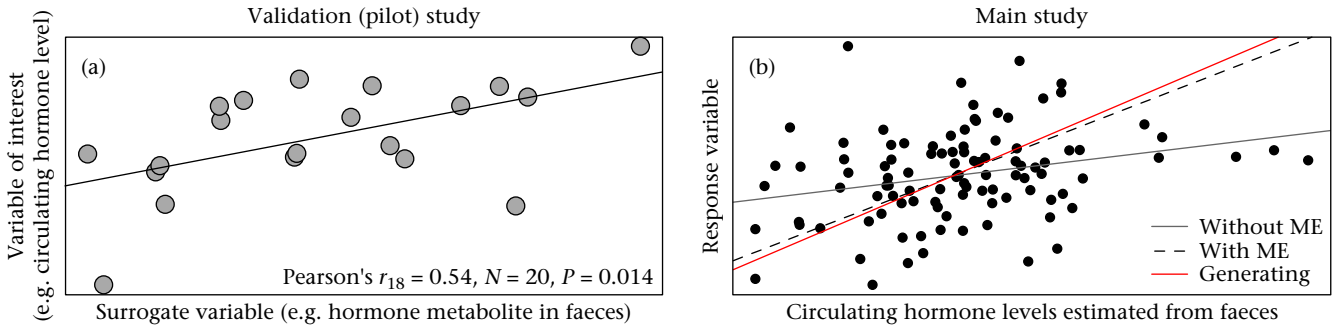

Figure 2. (a) Validation of a surrogate variable and (b) implementation of its reliability in the focal model as measurement error (ME). A hypothetical example when circulating hormone levels are assessed via the concentration of a derivate metabolite in the faecal sample. The validation study shows a strong but not perfect ($r = 0.54^2$) correlation between the surrogate and the focal variable. This estimate of reliability can be used to define measurement error structure (i.e. with the variance proportional to 0.54^2) around the predictors in the regression of interest in (b): the regression line that could be obtained by a standard regression analysis ('Without ME'); an estimate corresponding to the regression analysis that corrects for attenuation bias caused by measurement error ('With ME'); and the regression line based on parameters used to generate the example data ('Generating'). The corresponding R codes are available in the Supplementary Material.

well the surrogate variable explains variation in the variable of interest (e.g. by using measurement error models, as discussed below). Only a validation study can provide a reference for such a correction. Note that there may be instances when, instead of just reflecting another measurable trait, a surrogate variable is used as a proxy for a more composite trait (e.g. measure of fat as an estimate of body condition). In such instances, it is hard to test the correlation between the surrogate and the focal variables, in which case reliability is a question of theory rather than of statistics.

Measurement Error Models

The problem posed by the use of surrogate or indicator variables in terms of not explaining the full amount of variation in the focal variable can be translated as a measurement error problem (Bollen, 1989; Buonaccorsi, 2010; Fuller, 1987). Measurement error introduces uncertainty around parameter estimates even when the sample size is very large. In a univariate case, imprecise measurements will increase the error by which a single datum approximates the true mean of the sample and will shrink our confidence in each measurement (Chesher, 1991; Manisha & Singh, 2001). Such effects occur symmetrically on both sides of the distribution; thus, although measurement error unavoidably increases statistical noise, it does not raise systematic bias when one aims at describing central tendencies or distribution patterns for a single variable. In an NHT framework, this incurs issues about statistical power when the estimated mean of a sample is contrasted against a hypothetical mean. Accordingly, when measurement errors are present via the use of surrogate variables, we are more likely to commit type II statistical errors (failing to reject a false null hypothesis) than in the case of precisely measuring the trait of interest directly at the same sample size. In a non-NHT framework, this can be manifested in the increase of confidence ranges around parameter estimates.

When the relationship between two or more variables is under study, measurement error does not only affect precision in terms of confidence range or statistical significance, but it also leads to inaccurate estimates of correlation coefficients or regression slopes (Chesher, 1991; Fuller, 1987; Judge, Griffiths, Hill, Lutkepohl, & Lee, 1985). Standard statistical models assume that all variables have been measured without error (to be more precise, regression models apply this assumption for the predictor but not necessarily for the response variables). When particular variables or all variables can be approximated with certain error, conventional estimates of correlation coefficients (such as Pearson product–moment correlations) will include a bias towards zero and will undervalue true parameters. In the case of linear regression, such downward bias will be apparent in terms of the underestimation of R^2 and standardized regression coefficients if measurement error is present in the response variable (for nonlinear regression, the problem is more complex). However, if such errors apply to the predictors, unstandardized regression parameters will also be affected. The downward bias on parameter estimates due to measurement errors is known as 'attenuation bias' and necessitates particular statistical approaches that can correct for it. These measurement error models are available for both correlation and regression problems (e.g. Adolph & Hardin, 2007; Adolph & Pickering, 2008; Hansen & Bartoszek, 2012).

I suggest that these measurement error models can be efficiently exploited in noninvasive studies when a surrogate variable with reliability lower than 1 (i.e. in most cases) is used to approximate variation in the focal variable (Fig. 2, Supplementary Material). Imagine that a pilot study revealed that circulating hormone levels (e.g. adrenocorticoid stress hormone) could be reliably predicted by the concentration of a derivate metabolite estimated from faecal samples that could be collected without capturing the animals (e.g. see Lobato et al., 2008). The correlation of these two traits is relatively high, let us say around $r = 0.5–0.8$, while we also have evidence that metabolite concentrations in a faecal sample can be estimated with a high repeatability. Therefore, it is tempting to develop a noninvasive method based on metabolite levels from the faeces and assume that this variable is a good indicator of circulating hormone levels. Under this assumption, now suppose that a large number of samples are collected in parallel to behavioural observations in a study that aims to test for the effect of physiological stress (as exerted by certain stress factors in the environment) on particular behaviours. Then, we relate metabolite concentrations in the faecal samples to the behavioural variable, and we find a modest slope that appears nonsignificant in an NHT framework. Making strong conclusions from these results for the link between stress physiology and behaviour might not be too straightforward in this hypothetical example. However, by implementing the predictive value of the surrogate variable as estimated from the preceding validation study to define the measurement error structure, we might arrive at an improved estimate. We can use the repeatability or reliability estimates (in principle, both can be utilized analogically, as both are measured on the same scale) to simulate variations around each data point. If, given the above example, the surrogate explains 49–64% of the variance in the focal variable, the remaining unexplained variance can be considered as the uncertainty caused by imprecisions of measurements (i.e.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

measurement errors). Hence repeatability or reliability estimates can be used to generate measurement error variance and subsequently to correct for attenuation bias in an appropriate measurement error regression model. Similarly, repeatability (and reliability) estimates can be considered in tests of correlations in order to remove the downward bias that the imprecise relationship between the surrogate and focal variable causes. If known, repeatability/reliability of variables other than the surrogate variable can also be flexibly accounted for in the models. The philosophy of these correction methods is summarized in Fig. 2, while the underlying R codes (for both the correlation and regression problems) are given in the Supplementary Material.

In more general terms, measurement error models can effectively deal with measurement errors in both the response variables and the covariate variables. Therefore, such models are not only considered when working with surrogate traits, but they could also be fruitfully exploited in cases when either the response variables or the predictor variables can be measured with certain error.

UNKNOWN IDENTITY OF SUBJECTS

Given the ethical and practical concerns regarding the capture of rare or endangered species of animals, conservation biologists cannot always directly mark individuals, and in such cases they have to make behavioural observations without knowing the identity of the focal animals. Behaviour is a trait that, unlike many morphological or physiological traits, permits data collection from a distance (using binoculars or video cameras) to some extent. Therefore, if reliable means of identifying animals without tagging them are not possible, then observers could mistakenly enter data from the same individuals more than once, raising issues about pseudoreplication (Hurlbert, 1984). The independence of data is an important assumption of most statistical approaches (Sokal & Rohlf, 1995); thus, the chances of using multiple information on the same individual should be kept at minimum. The typical solution for the problem is to use a distance or temporal threshold (e.g. based on territory size or the length of the breeding season), beyond which it is highly unlikely that the observer encounters the same individual and mistakenly collects pseudoreplicates (if the identity is known, repeated measurements for the same individuals can be effectively used in a mixed modelling framework; see below). However, the use of such criterion values might be impractical for many reasons. For example, it requires some prior knowledge and pilot studies to determine the thresholds for data collection based on biological evidence. Furthermore, it leads to data loss if potentially useful information should be disregarded because they do not fulfil the established criteria for independence (this might be of particular importance if working with a species for which data are already heavily limited).

An alternative approach would to use the entire data set without any threshold criterion and consider the potentially confounding effect of nonindependence at the level of analysis. Below I provide two possible solutions towards this direction.

Randomization Procedures

Data randomization might also be useful to account for the unknown identity of subject animals, as it is easy to develop a routine for the random assignment of identity tags to each observation (Fig. 3). If we have N observations, this can originate from maximum N individuals (if each of them was recorded exactly once), and thus we can associate each datum with an ID label that was randomly taken from a hat containing N labels via a sampling process that allows replacement (i.e. bootstrap; Fig. 3a). If the same identity is accidentally ascribed to two or more observations, these can be treated as being taken from the same individual and thus an individual-specific mean can be calculated across them. Then, to obtain the parameter estimate of interest, these individual-specific values can be submitted to any statistical analyses that rely on the assumption of independence of units. The different steps (i.e. the random assignment of identity tags, the subsequent calculation of individual-specific means and the final statistical analyses) can be repeated several times. This will result in a large number of parameter estimates that corresponds to a large number of different scenarios describing the link between individual animals and observations. As an ultimate step, one would need to apply a multimodel inference across the whole analysis to derive an overall estimate across all possible outcomes (Burnham & Anderson, 2002). As each ID scenario is equally possible (unless we have some biological information to differentiate between them), the inference across all models should consider equal weights; thus, the result would basically encompass the arithmetic mean of the estimates and their 95% quantile range (Fig. 3a). The latter defines the confidence interval around the obtained estimate that arises from uncertainty due to the unknown identity of subjects.

The above procedure can be further developed. For example, the above model makes no assumptions about the size of the source population. It simply considers that the chance of accidentally observing the same individual more than once depends on the size of the sample (i.e. it is high in small samples and low in large samples). However, it is possible to take into account different assumptions about the size of the population from which the data originate. If we suspect that the population is considerably larger than the sample size, the number of ID tags that are used for the above bootstrap sampling can be augmented. Hence, the chance of multiple observations can be reduced even in very small samples. The increase in the population size will lead to a higher confidence around the parameter estimate because it more often results in scenarios in which each observation comes from a different individual. Similarly, if we know that the population size is smaller than the sample size and we are certain that some pseudoreplicate cases must occur in the data, we can use a corresponding number of ID tags for the random assignment process (Fig. 3b). If we have a rough estimate for the size of the population (e.g. it cannot be smaller or larger than a certain number), we can repeat the whole assignment process along the entire range between the considered minimum and maximum population size and perform the multimodel inference across both the randomization rounds and population size situations (Fig. 3c). In fact, if we do not know anything about the true population size, we can rerun these simulations and analyses between the two extreme scenarios: (1) when different observations come from different individuals and (2) when different observations come from the same individual. Note that if the population size is considerably smaller than the sample size, this will increase the risk of acquiring replicates from the same individuals, and thus will also increase uncertainty in the parameter estimate (Fig. 3a–c).

Another development of the above scheme can be achieved based on how the between- and within-individual organization of data is treated. So far, for simplicity, the procedure implemented a step based on the calculation of individual-specific means (once the same ID tags were found for two or more observations). This step inherently assumes that the behavioural trait is highly repeatable and that individuals display the trait in a very consistent way. However, unless we have data to prove otherwise, we need to consider that the repeatability of behaviours is modest at the best (Bell, Hankison, & Laskowski, 2009), and the plasticity of behaviour within individuals may play a confounding role. In such cases, it is misleading to calculate individual-specific values and use them in the statistical analyses (see reasoning in Dingemanse,

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

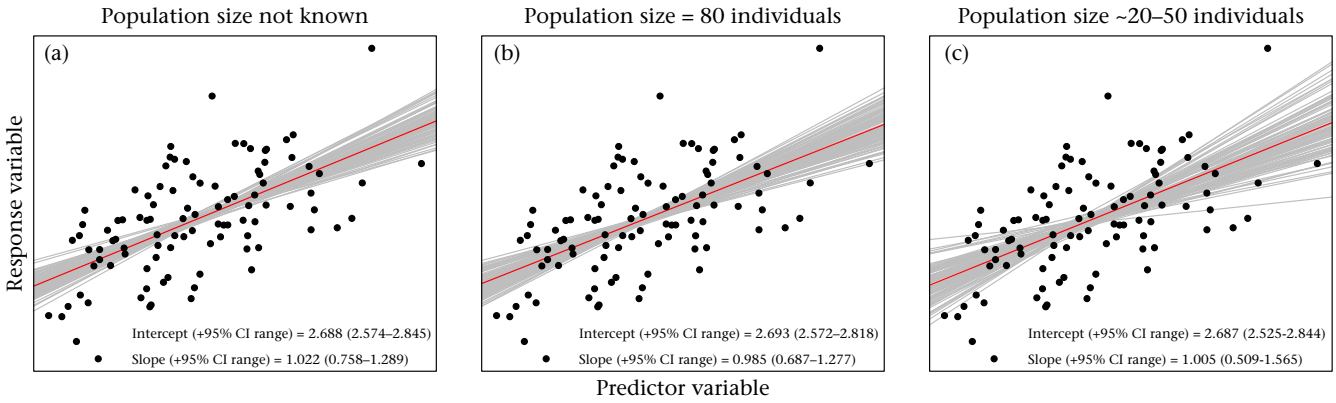

Figure 3. Dealing with the unknown identity of subjects through randomization of individuals' identity tags. Three types of regression results are given based on analysis of the same observation data but by considering different scenarios for the population size and probabilities for accidentally observing the same individuals more than once. (a) The number of ID labels is equal to the sample size and ID labels are reshuffled by multiple replacement sampling procedures to calculate individual-specific means for the correlation. (b–c) The number of ID labels is less than the sample size (as determined by the suspected population size) and ID labels are considered to be linked to the observation points with the between-individual variation. Grey lines shows 100 regression lines that can be obtained by particular sampling scenarios, and across which the mean (red line) and confidence range of the regression parameters can be determined. Note that as population size decreases, from left to right (a–c), the probability that two observations belong to the same individual increases, as does the confidence interval around the parameter estimates. The corresponding R codes are available in the Supplementary Material.

Dochtermann, & Nakagawa, 2012; Garamszegi & Herczeg, 2012). A straightforward alternative would be to enter the raw data with the randomized identity assignments without calculating individual-specific means into a mixed model framework, which can efficiently handle the hierarchical organization of data in the given sample (e.g. the within- and between-individual structure; Snijders & Bosker, 1999). Methods based on within-subject centering can be used to acquire unbiased parameter estimates both at the within- and between-individual levels (van de Pol & Wright, 2009), and this can be the preferred step between the random identity assignment and multimodel inference when the repeatability of the trait is suspected to be a concern.

Autocorrelation Models

The philosophy behind the definition of arbitrary thresholds based on certain spatial/temporal criteria to deal with the unknown identity of subjects can also be applied to define autocorrelation models that can effectively use the entire data set (Zuur, Ieno, & Smith, 2007). When using thresholds, the motivation for data exclusion relies on the assumption that the probability of two data points originating from the same individual is a function of the distance (defined either in space or in time) between the underlying observations. If this assumption holds, the same function can be used to define the expected correlation structure in the entire data: i.e. two adjacent observations are more likely to cause pseudocorrelation because there is a higher chance that they correspond to the same individual. Therefore, the temporal or spatial layout of the observations can be incorporated into the statistical models to characterize this expected autocorrelation.

The benefit of using such autocorrelation models, besides the prevention of data loss, is that it does not necessarily require preceding information about the biological system to determine the most effective threshold level. Instead, the available data can be vigorously analysed to detect autocorrelations (e.g. time series plot, variograms), and experience from such a diagnostic phase can be implemented in the analytical part. For example, temporally structured data (e.g. resulting from the fact that two observations that were made on the same day were probably derived from the same individual) will show that model residuals separated by a narrow time window will have very high correlation (if the focal trait is considerably consistent within individuals; Fig. 4a). Similarly, the same pattern in a spatial scale will drive model residuals to spread in a nonrandom manner (Fig. 5a). In such cases, the appropriate temporal or spatial autocorrelation structure can be used for model building, and improved parameter estimates can be

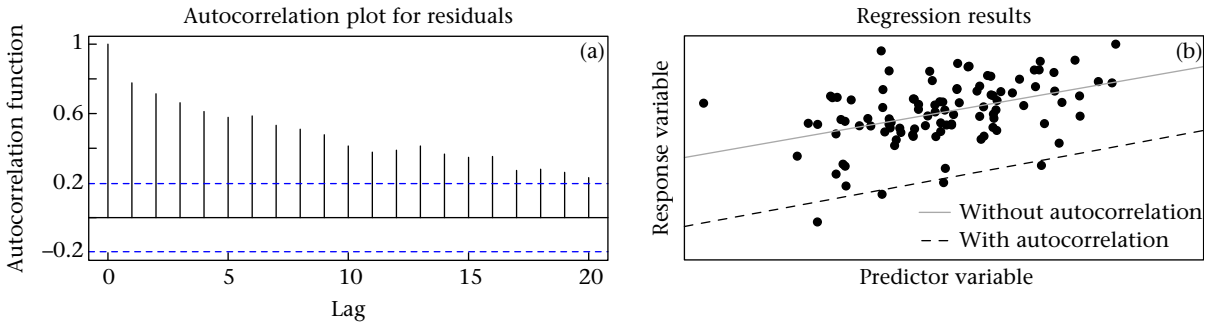

Figure 4. Analysis of temporally autocorrelated data assuming that the probability of two observations originating from the same individual depends on the time delay between the two observations. (a) Autocorrelation plot that informs the observer about the correlation structure in a sequentially obtained data set (how subsequent observations correlate with each other). The current diagnostics show strong serial nonindependence. (b) Behaviour of an autocorrelation function (modelling residuals at time s as a function of the residual of time $s-1$) and how the implementation of an expected temporal autocorrelation structure corrects for the model parameters. The solid grey line is the regression line that could be obtained by a standard ordinary regression analysis. The dashed grey line is the estimate that corresponds to the regression analysis that includes temporal autocorrelation. The corresponding R codes are available in the Supplementary Material.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

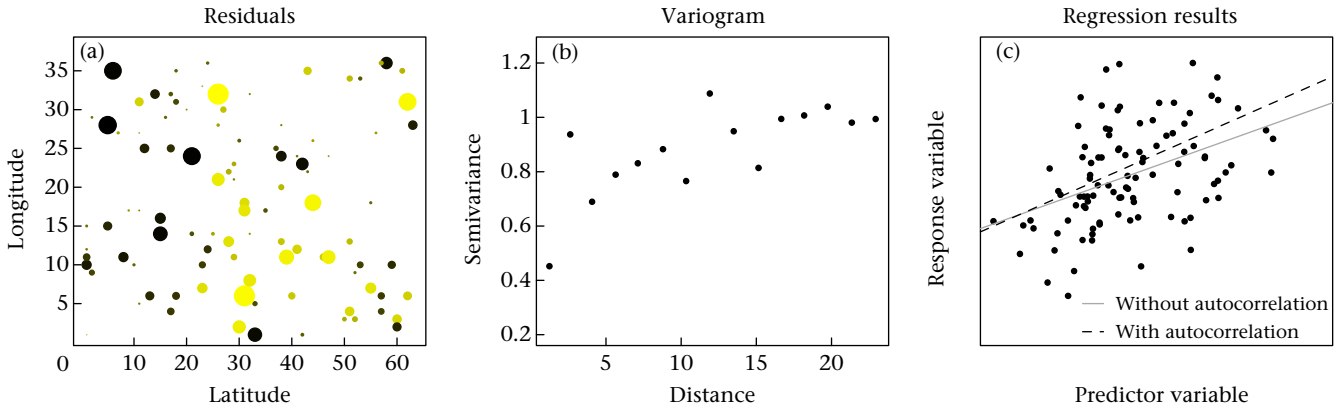

Figure 5. Analysis of spatially autocorrelated data assuming that the probability of two observations originating from the same individual depends on the distance between the localities where the two observations were taken. Following the logic of Fig. 4, (a) and (b) are diagnostic plots that inform the observer about the correlation pattern in a spatially structured data set: (a) model residuals along geographical coordinates, with the attributes of dots representing their absolute values (bigger dots are for larger residuals) and signs (●: positive residuals; ○: negative residuals); (b) a variogram (function describing the degree of spatial dependence) generated for the observed data that indicates a strong spatial nonindependence. (c) Behaviour of an autocorrelation function (modelling residuals by assuming exponential variogram functions) and how the implementation of an expected spatial autocorrelation structure corrects for the model parameters. The solid grey line is the regression line that could be obtained by a standard ordinary regression analysis. The dashed black line is the estimate that corresponds to the regression analysis that includes spatial autocorrelation. The corresponding R codes are available in the Supplementary Material.

obtained from them. In addition, model fit statistics (e.g. based on Akaike's Information Criterion; Burnham & Anderson, 2002) can be used to select the model that best describes the data and can be further used to make judgements about the importance for the control of temporal or spatial arrangements of observations. Figs. 4 and 5 demonstrate the philosophy behind this kind of modelling, while in the Supplementary Material, I give the corresponding R codes that can be used in similar exercises for the autocorrelation approach. The autocorrelation structures can also be implemented in the above randomization procedures; thus, the two approaches that deal with the unknown identity of subjects can be effectively combined.

Note that autocorrelation models may not only be promising when dealing with the unknown identity of subject, but also in a whole range of situations when the similarity of behaviour between two individuals depends on the spatial or temporal distance between them. This can happen, for example, when neighbouring individuals learn or copy behavioural elements from each other (Aplin et al., 2015; Dobkin, 1979; Freeberg, 1998; Sprau & Mundry, 2010). Furthermore, time series would be an obvious choice for studying long-term changes in population numbers. One can also envisage several situations where a conservation work includes an important spatial component and when autocorrelation models could be applied.

COMPLEX BIOLOGICAL QUESTIONS: COMPLEX STATISTICAL ANSWERS

Mixed Modelling: the Specific Hierarchical Structure of Behavioural Data

Studying the adaptive role of behaviour in the light of the rapidly and unpredictably changing environment may be interesting from a conservation aspect, as such adaptation processes are highly relevant under the recently occurring climatic changes. An individual's behaviour can change rapidly from moment to moment, and this plasticity offers a mean by which the animal can quickly react to an emerging stress factor in the environment (Dingemanse, Kazem, Réale, & Wright, 2010). Given that quick changes in the environment occur mostly in an unpredictable manner, there will be weak directional selection for particular phenotypes. However, in such conditions, individuals that maintain their ability to flexibly display a certain range of phenotypes will be favoured, and considerable within-individual variance in behaviour will be preserved. Despite the importance of such plasticity, some degree of behavioural consistency is maintained because individuals cannot vary their behaviours unlimitedly (Wilson, 1998). This drives some remaining between-individual differences in mean behaviour within the population, which can also be advantageous if environmental conditions change from year to year in a stochastic manner (Dingemanse, Both, Drent, & Tinbergen, 2004). Altogether, within- and between-individual variances in behaviour contribute to the observed distribution of behavioural phenotypes in the population. Hence, decomposing such variance components and linking them to fitness allows an understanding of adaptation processes that, as a buffer mechanism, precede life-history, morphological adaptations and genetic alterations. Establishing the adaptive role of within- and between-individual variance of behavioural traits might be of particular importance in species with conservation concern.

The most commonly applied approach to the study of behaviour (both from the evolutionary and conservation perspective) focuses on between-individual differences. Under this focus, if few within-individual repeats are available, these measurements are averaged at the individual level (Garamszegi, Markó, & Herczeg, 2012). The biological relevance of such individual-specific estimates can be assessed beforehand by calculating repeatability (the proportion of between-individual variance relative to the sum of the between-individual and the within-individual variance). In practice, if such repeatabilities are statistically differentiable from zero (i.e. they are significant), individual-specific means are calculated and used in the subsequent statistical analyses to investigate their relationship with environmental factors or fitness. However, numerically, the repeatability of behaviour is modest at the best (mostly falls between 0.3 and 0.4; Bell et al., 2009), indicating that a huge amount of variation is thrown away when within-individual patterns are disregarded by focusing on individual-specific mean estimates of behaviours.

Statistical methods based on generalized linear mixed models (GLMM) offer an elegant way to work in parallel with phenomena at the within- and between-individual levels and to explore rigorously the hierarchical structure of data (Dingemanse &

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

Dochtermann, 2013; Gałecki and Burzykowski, 2013; Snijders & Bosker, 1999; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). For example, one can explore how individual behaviour changes along an ecological gradient (reaction norm), which allows making inferences for behavioural plasticity (Dingemanse et al., 2010). Furthermore, in the same GLMM one can also examine how the characteristics of the reaction norms (i.e. mean and slope) differ among individuals, which can be used to draw conclusions about behavioural consistency. As these within-individual and between-individual characteristics can be crucial to understand how animal populations acclimate to their ecological environments, I suspect that GLMM can be effectively exploited in studies targeting conservation issues as well. The benefits are especially transparent in endangered species, in which obtaining large sample sizes in terms of the number of individuals is highly impractical. In such cases, investments in gathering repeated measurements from the same individuals could be increased, and the biological information that reside in the within-individual samples can be fruitfully recovered in a mixed model approach while controlling for pseudoreplication. However, there is always a balance to make between sampling many individuals few times and sampling few individuals several times (Dingemanse & Dochtermann, 2013; Garamszegi & Herczeg, 2012; Martin et al., 2011; van de Pol, 2012). Note also that GLMMs can handle hierarchical structures of any kind, and thus this may be an ideal approach when comparing different populations (that are exposed to different levels of pollution, for example) as focal units based on individual-specific measurements within these units. Mixed models have a number of statistical properties (e.g. borrowing of strength, global smoothing of uncertainty and shrinkage to the population mean) that can prove useful when working with conservation-related questions on hierarchical data.

Comparative Approaches: Working with Substitute Species and Dealing with Phylogenetic Constraints

When the species or population to be protected is difficult to study due to logistical/ethical/political reasons, a practical alternative is to perform investigations on a substitute species and extrapolate the findings to the species of concern (Caro, Eadie, & Sih, 2005). This approach relies on the assumption that the scientific knowledge obtained in the substitute species will cast light on the original conservation problem in the target species. Therefore, the effectiveness of this approach is directly linked to how the substitute species is chosen. In general, the substitute species should be very similar biologically to the target species, while in particular, substitutes and targets should share the same ecological, life-history or behavioural traits that make the target sensitive to the current environmental situation. If these circumstances are fulfilled, one can develop reliable predictive models to the conservation target based on biologically informed functions that were defined in the substitute species.

However, for substitutes to be appropriate, the functions that will be used to impute data and make inferences in the target species should not only be defined based on ecological, life-history or behavioural parameters, but should consider phylogenetic constraints. For example, even if the two species appear very similar in several aspects, phylogenetic relatedness can still set up limits for the generalization of the results if the two species are not sister species or either of them represent an evolutionary singularity (Nunn & Zhu, 2014). Therefore, the analysis of data obtained from substitute species needs to be carried out with the appropriate phylogenetic methods that are able to account for effects due to common descent (Garamszegi, 2014; Harvey & Pagel, 1991; Nunn, 2011; Paradis, 2011). Furthermore, interspecific comparative studies that include several closely and distantly related 'substitute' species that are placed in the appropriate phylogenetic context can be used to define the predictive functions. Note that phylogenetic comparative methods are not only relevant when substitute species are used, but they can also be applied when different populations of the same species are compared with respect to a conservation-related issue (Stone, Nee, & Felsenstein, 2011), or when entire animal communities in which the endangered/threatened species is involved are being examined (Pearse, Purvis, Cavender-Bares, & Helmus, 2014).

Bayesian Approaches: Implementing Prior Beliefs and Imputing Missing Data

Bayesian statistics is gaining popularity in various biological fields, and conservation biologists also seem to follow this tendency (Brooks, Freeman, Greenwood, King, & Mazzetta, 2008; Marin, Diez, & Insua, 2003; Wade, 2000). Bayesian inference provides alternative means to analyse data from traditional hypothesis-testing techniques (Congdon, 2006; Gelman, Carlin, Stern, & Rubin, 1995; McCarthy, 2007), and also offers some techniques that can be exploited when researchers are faced with the constraints of studying species with concerns for conservation.

In a Bayesian framework, the existing knowledge available before a study is performed (current belief that can be placed on models, hypotheses or parameters) is summarized in a quantitative form, which is used as a prior probability distribution. Another component of a Bayesian inference is a likelihood function that expresses how one expects the data to look given that the model/hypothesis/parameter is true. This prior probability is updated in light of the new data and the likelihood function through a Markov chain Monte Carlo sampling process, and a posterior probability distribution is generated. This posterior density provides a direct measure of the probability of a parameter or hypothesis given the data (and not the probability of data given the null hypothesis, as in the NHT context), and will serve as the scientific results (thus, central tendency and data spread can be used to describe the parameter estimate and the uncertainty around it, respectively).

One benefit of this type of inference is the possibility of the efficient exploitation of prior information, which can be quite useful when sample size is small. Priors define an initial probability distribution by setting out the boundaries of the parameter space where the Markov chain is allowed to evaluate the parameters of interest. Therefore, the careful and biologically motivated choice of priors can increase the precision and accuracy of the posterior estimate at a lower sample size. Furthermore, the use of informative priors can avoid extreme/nonsensical effects. For example, if a previous meta-analysis integrating information from almost 100 experiments documented that predator removal successfully increased hatching success of bird populations by 77%, on average (Smith, Pullin, Stewart, & Sutherland, 2010), this information can be incorporated into an analysis of new field data. Suppose that an ongoing study aims at investigating this relationship in a small population of an endangered species and finds a nonsignificant difference when comparing removal areas against control areas based on the traditional NHT-based analysis of the available small sample. The researcher may tend to conclude that predation pressure is not a concern for the species' conservation programme. However, by applying a Bayesian approach, the mean effect size and the associated confidence interval from the meta-analysis could be used to condition an informative prior distribution, which could lead to a posterior distribution on the data that has different suggestive value. It is likely that the posterior density for the effect size for the between-group differences will fall in the positive range, indicating that predation has some impact on the breeding

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

performance of the studied bird. Therefore, the NHT-based and Bayesian-based inference of the same data might provide completely different motivations for the design of protection programmes (see a similar example with illustrations in Garamszegi et al., 2009).

Another advantage of the Bayesian method is that it can be prosperously utilized when dealing with different kinds of missing data by imputation methods (Kong, Liu, & Wong, 1994; Little & Rubin, 2002; Nakagawa & Freckleton, 2008). The constraints on data collection may lead to cases when observation is available for certain variables but not for others. Such cases are typically casewise deleted from multivariate analyses; thus, missing data can drastically reduce the sample size that is available for models involving more than one variable. However, estimating the unobserved data and making predictions are at the heart of the Bayesian process. Such a feature can be exploited to impute missing data based on a defined function in parallel to the characterization of the posterior distribution of parameters. Note that data imputation cannot only be applied in the within-species context, but also in a between-species context when inferences are being made based on a substitute species in a phylogenetic framework (see Comparative Approaches: Working with Substitute Species and Dealing with Phylogenetic Constraints).

CONCLUSIONS

Here I have provided an overview on various statistical methods that can be useful for conservation biologists and animal behaviourists when analysing behavioural data that are often loaded with various constraints. The common theme appearing in this discussion is that some aspects of these constraints should be regarded as sources of noise that generate uncertainty around the estimated parameters (e.g. small sample size, unknown identity of subjects). These uncertainties are inherent components of the results and should be appreciated at each level of a conservation effort: during scientific reporting and interpretation and during the practical phase when protection programmes are being designed. Statistically, classical NHT-based inferences can often give rise to misleading conclusions when confronted with such uncertainty. On the other hand, simulation and randomization techniques (including the Bayesian approach) allow efficient solutions to such problems. The imputation of missing data, the use of prior information, fitting measurement error and/or hierarchical models are means by which uncertainties can be efficiently handled. Some constraints emerging in studies with a conservation angle can cause bias (e.g. use of surrogate variables, measurement errors, within-individual variance, phylogenetic structure), but approaches are also available (e.g. autocorrelation models, mixed models and phylogenetic comparative approaches) that can correct for such bias in the parameter estimate of interest. To enhance the implementation of the discussed methodology into practice, I provide an exemplary demonstration for most of the statistical situations in the R statistical environment as Supplementary Material. My hope is that not only will conservation biologists benefit from this discussion and resource, but that anyone who works on issues in association with animal behaviour will also find this material useful.

Acknowledgments

I thank E. Fernández-Juricic and B. A. Schulte for inviting me to contribute to the Special Issue on Conservation Behaviour. I am also grateful to one anonymous referee and Matthew Symonds for their constructive comments on an earlier version of the manuscript. During this study, the author was supported by funds from the Spanish government within the frame of the Plan Nacional Programme (no. CGL2012- 38262 and no. CGL2012-40026-C02-01) and from the National Research, Development and Innovation Office in Hungary (NKFIH, no. K-115970).

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.anbehav.2015.11.009.

References

Adolph, S. C., & Hardin, J. S. (2007). Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Functional Ecology, 21*, 178–184.
Adolph, S. C., & Pickering, T. (2008). Estimating maximum performance: effects of intraindividual variation. *Journal of Experimental Biology, 211*, 1336–1343.
Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature, 518*, 538–541.
Atkinson, A. C. (1985). *Plots, transformations and regression: An introduction to graphical methods of diagnostic regression analysis*. New York, NY: Oxford University Press.
Bauer, B., Palme, R., Machatschke, I. H., Dittami, J., & Huber, S. (2008). Non-invasive measurement of adrenocortical and gonadal activity in male and female guinea pigs (*Cavia aperea f. porcellus*). *General and Comparative Endocrinology, 156*, 482–489.
Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: a meta-analysis. *Animal Behaviour, 77*, 771–783.
Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75*(5), 785–804.
Bolker, B. (2007). *Ecological models and data in R*. Princeton, NJ: Princeton University Press.
Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: J. Wiley.
Bradshaw, C. J. A., & Brook, B. W. (2010). The conservation biologist's toolbox: principles for the design and analysis of conservation studies. In N. S. Sodhi, & P. R. Ehrlich (Eds.), *Conservation biology for all* (pp. 313–339). Oxford, U.K.: Oxford University Press.
Brooks, S. P., Freeman, S. N., Greenwood, J. J. D., King, R., & Mazzetta, C. (2008). Quantifying conservation concern: Bayesian statistics, birds and the red lists. *Biological Conservation, 141*, 1436–1441.
Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. New York, NY: Chapman & Hall.
Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer-Verlag.
Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.
Caro, T., Eadie, J., & Sih, A. (2005). Use of substitute species in conservation biology. *Conservation Biology, 19*, 1821–1826.
Chesher, A. (1991). The effect of measurement error. *Biometrika, 78*, 451–462.
Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: L. Erlbaum.
Cohen, J. (1994). The earth is round ($p< .05$). *American Psychologist, 49*, 997–1003.
Congdon, P. (2006). *Bayesian statistical modelling*. Chichester, U.K.: J. Wiley.
Dingemanse, N. J., Both, C., Drent, P. J., & Tinbergen, J. M. (2004). Fitness consequences of avian personalities in a fluctuating environment. *Proceedings of the Royal Society B: Biological Sciences, 271*, 847–852.
Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology, 82*, 39–54.
Dingemanse, N. J., Dochtermann, N. A., & Nakagawa, S. (2012). Defining behavioural syndromes and the role of 'syndrome deviation' in understanding their evolution. *Behavioral Ecology and Sociobiology, 66*, 1543–1548.
Dingemanse, N. J., Kazem, A. J. N., Réale, D., & Wright, J. (2010). Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution, 25*, 81–89.
Dobkin, D. S. (1979). Functional and evolutionary relationhips of vocal copying phenomena in birds. *Zeitschrift für Tierpsychologie, 50*, 348–363.
Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap. Monographs on Statistics and Applied Probability 57*. New York, NY: Chapman & Hall.
Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.
Freckleton, R. P. (2011). Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology, 65*, 91–101.
Freeberg, T. M. (1998). The cultural transmission of courtship patterns in cowbirds, *Molothrus ater*. *Animal Behaviour, 56*, 1063–1073.
Fuller, W. A. (1987). *Measurement error models*. New York, NY: J. Wiley.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. New York, NY: Springer-Verlag.
Garamszegi, L. Z. (2006). Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behavioral Ecology, 17*, 682–687.
Garamszegi, L. Z. (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice*. Berlin, Germany: Springer-Verlag.
Garamszegi, L. Z., Calhim, S., Dochtermann, N., Hegyi, G., Hurd, P. L., Jørgensen, C., et al. (2009). Changing philosophies and tools for statistical inferences in behavioural ecology. *Behavioral Ecology, 20*, 1363–1375.
Garamszegi, L. Z., & Herczeg, G. (2012). Behavioural syndromes, syndrome deviation and the within- and between-individual components of phenotypic correlations: when reality does not meet statistics. *Behavioral Ecology and Sociobiology, 66*, 1651–1658.
Garamszegi, L. Z., Markó, G., & Herczeg, G. (2012). A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. *Evolutionary Ecology, 26*, 1213–1235.
Gelman, A. (2015). Working through some issues. *Significance, 12*, 33–35.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London, U.K.: Chapman & Hall.
Gorsuch, R. L., & Lehmann, C. S. (2010). Correlation coefficients: mean bias and confidence interval distortions. *Journal of Methods and Measurement in the Social Sciences, 1*, 52–65.
Goymann, W., Möstl, E., & Gwinner, E. (2002). Non-invasive methods to measure androgen metabolites in excrements of European stonechats, *Saxicola torquata rubicola*. *General and Comparative Endocrinology, 129*, 80–87.
Hansen, T. F., & Bartoszek, K. (2012). Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology, 61*, 413–425.
Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford, U.K.: Oxford University Press.
Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs, 54*, 187–211.
Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics*. New York, NY: J. Wiley.
Kersey, D. C., & Dehnhard, M. (2014). The use of noninvasive and minimally invasive methods in endocrinology for threatened mammalian species conservation. *General and Comparative Endocrinology, 203*, 296–306.
Kong, A., Liu, J., & Wong, W. H. (1994). Sequential imputation and Bayesian missing data problems. *Journal of the American Statistical Association, 89*, 278–288.
Lane, J. (2006). Can non-invasive glucocorticoid measures be used as reliable indicators of stress in animals? *Animal Welfare, 15*, 331–342.
Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association, 73*, 253–263.
Lenz, T. L., Wells, K., Pfeiffer, M., & Sommer, S. (2009). Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the long-tailed giant rat (*Leopoldamys sabanus*). *BMC Evolutionary Biology, 9*, 269. http://dx.doi.org/10.1186/1471-2148-9-269.
Lessells, C. M., & Boag, P. T. (1987). Unrepeatable repeatabilities: a common mistake. *Auk, 104*, 116–121.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: J. Wiley.
Lobato, E., Merino, S., Moreno, J., Morales, J., Tomas, G., La Puente, J. M. D., et al. (2008). Corticosterone metabolites in blue tit and pied flycatcher droppings. *Hormones and Behaviour, 53*, 295–300.
Lukas, D., Bradley, B. J., Nsubuga, A. M., Doran-Sheehy, D., Robbins, M. M., & Vigilant, L. (2004). Major histocompatibility complex and microsatellite variation in two populations of wild gorillas. *Molecular Ecology, 13*, 3389–3402.
Manisha, & Singh, R. K. (2001). An estimation of population mean in the presence of measurement errors. *Journal of the Indian Society of Agricultural Statistics, 54*, 13–18.
Manly, B. F. J. (1991). *Randomization, bootstrap and Monte Carlo methods in biology*. New York, NY: Chapman & Hall.
Marin, J. M., Diez, R. M., & Insua, D. R. (2003). Bayesian methods in plant conservation biology. *Biological Conservation, 113*, 379–387.
Martin, J. G. A., Nussey, D. H., Wilson, A. J., & Réale, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods in Ecology and Evolution, 2*, 362–374.
Martinez-Abrain, A. (2014). Is the '$n = 30$ rule of thumb' of ecological field studies reliable? A call for greater attention to the variability in our data. *Animal Biodiversity and Conservation, 37*, 95–100.
McCarthy, M. A. (2007). *Bayesian methods for ecology*. Cambridge, U.K.: Cambridge University Press.
Mench, J. A. (2000). Refinement in behavioural research. In M. Balls, A.-M. van Zeller, & M. Halder (Eds.), *Progress in reduction, refinement and replacement of animal experimentation* (pp. 1213–1221). Amsterdam, The Netherlands: Elsevier.
Møller, A. P., & Jennions, M. D. (2001). Testing and adjusting for publication bias. *Trends in Ecology & Evolution, 16*, 580–586.
Mundry, R. (2014). Statistical issues and assumptions of phylogenetic generalised least squares. In L. Z. Garamszegi (Ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (pp. 131–153). Berlin, Germany: Springer-Verlag.
Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology, 15*, 1044–1045.
Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 82*, 591–605.
Nakagawa, S., & Freckleton, R. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution, 23*, 592–596.
Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*, 1253–1274.
Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews, 85*, 935–956.
Narayan, E. J. (2013). Non-invasive reproductive and stress endocrinology in amphibian conservation physiology. *Conservation Physiology, 1*(1). http://dx.doi.org/10.1093/conphys/cot011. cot011.
Nunn, C. L. (2011). *The comparative approach in evolutionary anthropology and biology*. Chicago, IL: University of Chicago Press.
Nunn, C. L., & Zhu, L. (2014). Phylogenetic prediction to identify 'evolutionary singularities'. In L. Z. Garamszegi (Ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (pp. 481–514). Berlin, Germany: Springer-Verlag.
Paradis, E. (2011). *Analysis of phylogenetics and evolution with R*. Berlin, Germany: Springer.
Pearse, W. D., Purvis, A., Cavender-Bares, J., & Helmus, M. R. (2014). Metrics and models of community phylogenetics. In L. Z. Garamszegi (Ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice* (pp. 451–464). Berlin, Germany: Springer-Verlag.
Pereira, R. J. G., Duarte, J. M. B., & Negrao, J. A. (2005). Seasonal changes in fecal testosterone concentrations and their relationship to the reproductive behaviour, antler cycle and grouping patterns in free-ranging male pampas deer (*Ozotoceros bezoarticus bezoarticus*). *Theriogenology, 63*, 2113–2125.
Piggott, M. P. (2004). Effect of sample age and season of collection on the reliability of microsatellite genotyping of faecal DNA. *Wildlife Research, 31*, 485–493.
van de Pol, M. (2012). Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. *Methods in Ecology and Evolution, 3*, 268–280.
van de Pol, M. V., & Wright, J. (2009). A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour, 77*, 753–758.
Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, U.K.: Cambridge University Press.
R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science, 46*, 175–208.
Rosenberg, M. S. (2005). The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution, 59*, 464–468.
Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney *U* test. *Behavioral Ecology, 17*, 688–690.
Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences*. New York, NY: McGraw-Hill.
Smith, R. K., Pullin, A. S., Stewart, G. B., & Sutherland, W. J. (2010). Effectiveness of predator removal for enhancing bird populations. *Conservation Biology, 24*, 820–829.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London, U.K.: Sage.
Sokal, R. R., & Rohlf, F. J. (1995). *Biometry*. New York, NY: W. H. Freeman.
Soper, H. E., Young, A. W., Cave, B. M., Lee, A., & Pearson, K. (1917). On the distribution of the correlation coefficient in small samples. Appendix II to the papers of 'Student' and R. A. Fisher. *Biometrika, 11*, 328–413.
Sprau, P., & Mundry, R. (2010). Song type sharing in common nightingales, *Luscinia megarhynchos*, and its implications for cultural evolution. *Animal Behaviour, 80*, 427–434.
Stephens, P. A., Buskirk, S. W., & del Rio, C. M. (2007). Inference in ecology and evolution. *Trends in Ecology & Evolution, 22*, 192–197.
Stone, G. N., Nee, S., & Felsenstein, J. (2011). Controlling for non-independence in comparative analysis of patterns across populations within species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 1410–1424.
Taberlet, P., & Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society, 68*, 41–55.
Taborsky, M. (2010). Sample size in the study of behaviour. *Ethology, 116*, 185–202.
Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher, 31*, 25–32.
Touma, C., & Palme, R. (2005). Measuring fecal glucocortoids in mammals and birds: the importance of validation. *Annals of the New York Academy of Science, 1064*, 54–74.
Wade, P. R. (2000). Bayesian methods in conservation biology. *Conservation Biology, 14*, 1308–1316.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009

Williams, J. H. G., Greenhalgh, K. D., & Manning, J. T. (2003). Second to fourth finger ratio and possible precursors of developmental psychopathology in preschool children. *Early Human Development, 72*, 57–65.

Wilson, D. S. (1998). Adaptive individual differences within single populations. *Philosophical Transactions of the Royal Society B: Biological Sciences, 353*, 199–205.

de Winter, J. C. F. (2013). Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research and Evaluation, 18*, 10.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution, 3*, 129–137.

Zar, J. H. (1996). *Biostatistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

Zuur, A., Ieno, E. N., & Smith, G. M. (2007). *Analyzing ecological data*. Berlin, Germany: Springer-Verlag.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Berlin, Germany: Springer-Verlag.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*, 3–14.

Please cite this article in press as: Garamszegi, L. Z., A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons, Animal Behaviour (2015), http://dx.doi.org/10.1016/j.anbehav.2015.11.009