Research Focus

Missing inaction: the dangers of ignoring missing data

Shinichi Nakagawa[1,2] and Robert P. Freckleton[2]

[1] Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand
[2] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

The most common approach to dealing with missing data is to delete cases containing missing observations. However, this approach reduces statistical power and increases estimation bias. A recent study shows how estimates of heritability and selection can be biased when the 'invisible fraction' (missing data due to mortality) is ignored, thus demonstrating the dangers of neglecting missing data in ecology and evolution. We highlight recent advances in the procedures of handling missing data and their relevance and applicability.

The best solution to handle missing data is to have none. – R.A. Fisher

A ubiquitous issue but a neglected topic

Unfortunately, in real-world data sets, missing data are the norm rather than the exception [1–4]. Researchers usually omit cases containing missing data from analyses, concentrating only on sample units for which complete data are available (complete case analysis). At first sight this procedure might seem reasonable, and indeed it might appear that there is no other option available. However, in doing so, researchers often throw away a large part of their data, especially when a data set contains many variables but whole cases are deleted based on only one or two variables not being measured. Even worse, the parameter estimates from such pruned data sets are often incorrect when data are not missing completely at random (MCAR) [5] (Box 1). The illustrations of missing data which are missing at random (MAR) or missing not at random (MNAR) in Box 1 clearly demonstrate potential biases in parameter estimates that can be caused by deleting cases with missing observations when missing data are not MCAR. Given that MCAR is a very strong (and often incorrect) assumption, it is somewhat surprising that the topic of missing data in ecology and evolution has been largely ignored to date. For example, in recent years, *Trends in Ecology and Evolution* has hosted a series of papers discussing major advances in statistical philosophies and reform of statistical practices in ecology and evolution [6–8], but none of these reviews mention the topic of missing data. Possibly a major reason for this is that dealing with missing data is a rather technical issue. However, we believe that recent advances in handling missing data have made it possible to begin to tackle this difficult issue with the aid of techniques that have become well accepted in the statistical literature [2,4].

Corresponding author: Nakagawa, S. (shinichi.nakagawa@otago.ac.nz).

Here we highlight a recent study that clearly demonstrates the importance of missing data in evolutionary studies, along with some related work showing that ignoring missing data can compromise analyses in general. We then discuss some of the techniques that have been developed to deal with missing data which can be employed by researchers in the field of ecology and evolution.

Visualising the invisible fraction

Although there are a multitude of reasons why data sets contain missing observations, there are often biologically significant reasons for why this might be. For instance, missing data might be particularly important if organisms die before expression of a trait (e.g. secondary sexual traits) or while a trait is still being developed (e.g. weight or height). 'Missing' observations due to premature death in relation to the trait of interest are referred to as the 'invisible fraction' in the evolutionary literature [9]. Although the importance of invisible fractions in trait evolution has long been pointed out [9], researchers have ignored invisible fractions in calculating evolutionary parameters (e.g. heritability and selection gradients and differentials).

A recent paper by Hadfield [10] demonstrates the crucial importance of dealing with the invisible fraction in calculating such evolutionary parameters. The main result of the paper is that when a trait is under viability selection (i.e. a trait relates to survival; e.g. lighter chicks have higher mortality than heavier chicks), missing data due to the invisible fraction are MNAR in most cases (because a sample taken at a certain time after hatching will 'miss' light chicks that have already died). Therefore, estimators of evolutionary parameters such as heritability and selection (e.g. of body weight) are biased if estimated without accounting for the invisible fraction. Although calculating heritabilities and selection gradients is commonplace, very few studies have to date considered this problem. For traits under viability selection, missing observations depend on lifespan. If data sets include lifespan (e.g. if lifespan is the x-variable in Box 1, Figure Ia), missing observations in such traits can be treated as MAR, whereas without information on lifespan, the missing observations remain MNAR. Unfortunately, data on lifespan are rarely measured accurately and are also frequently incomplete (i.e. individuals might die between censoring points, or only after the final censoring point). With such incomplete lifespan data, missing observations of a trait under viability selection (i.e. the invisible fraction) are still MNAR.

When data are MNAR, it is necessary to make some assumptions about how the data are missing. With

Box 1. Problems of missing data and their bewildering classification

The mechanisms (distribution patterns) of missing data are traditionally divided into three classes: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [5]. If missing data are not MCAR, then there are potential problems in analysing data as though they were, but the precise outcome depends on the way in which they are missing, specifically whether data are MAR or MNAR. Figure I shows simple examples of this in the case of a bivariate regression.

In Figure I, assuming complete observations in the *x*-variable (red and blue points), missing data (red points) in (a) are MAR, missing data in (b) are MNAR and missing data in (c) are MCAR. This is because in (a), missing data in the *y*-variable (termed missingness) depend on the *x*-variable ($x < 0$) whereas in (b), missing data in the *y*-variable (missingness) depend on the *y*-variable itself ($y < 0$). For example, (a) represents a situation where the lifespan (*x*; possibly log scale) of chicks is correlated with weight at 13 days after hatching (*y*); the weight information of chicks that die before the 13th day is not available, although information on lifespan is available for all individuals. For (b), imagine a slightly different situation where lifespan (*y*; log scale) and hatching weight (*x*) are correlated; lifespan is only measured after a certain point (e.g. 13 days after hatching) whereas the information on hatching weight for all individuals is available. It is important to note, although confusing, that if the *x*-variable (e.g. lifespan) is not among those measured in a situation such as (a), missing data in (a) should be classified as MNAR. Thus, MNAR comes in two forms: (i) missingness depends on the missing value itself or (ii) missingness depends on an unobserved variable (see Refs [1–4,10,18] for more technical and precise definitions for missing mechanisms, along with definitions for 'ignorability' of missing data; MCAR and MAR are referred to as 'ignorable' whereas MNAR is 'non-ignorable').

In addition to obvious biases in the means and variances of *x* and *y* due to missing data in (a) and (b) but not in (c), the key point in Figure I is that the estimates of the slope might be biased if missing data are not MCAR, depending on the nature of the missing data. In (a), the expected slope is unbiased as missingness depends on *x*, with the result that covariance between *x* and *y* is reduced by the same amount as the reduction in standard deviation of *x*. However, in (b), this is not true and the slope is biased as the missingness is determined by *y*. In (c), the slope is unbiased. Moreover, in (c), the expected R^2 is the same as for the original data, whereas in the case of (a) and (b), the R^2 is reduced. Clearly, even in this simple example, the consequences of missing data are not straightforward to predict. The situation will become much more complex if a multivariate data set is considered. Moreover, in reality, a data set might contain variables with missing observations which are MCAR, MAR or MNAR.

Notably, this classification (i.e. MCAR, MAR and MNAR) has been criticised because of the confusing nature of its terminology (e.g. MAR does not mean that missing data are distributed at 'random'). Furthermore, MNAR can be difficult to distinguish from MAR owing to the very fact that we have no information regarding missing values when MNAR (but see Box 2) [4]. Therefore, importantly, the most practical assumption is MAR, which is a basis of recent advances of handling missing data (e.g. Box 2) [1–4,18].

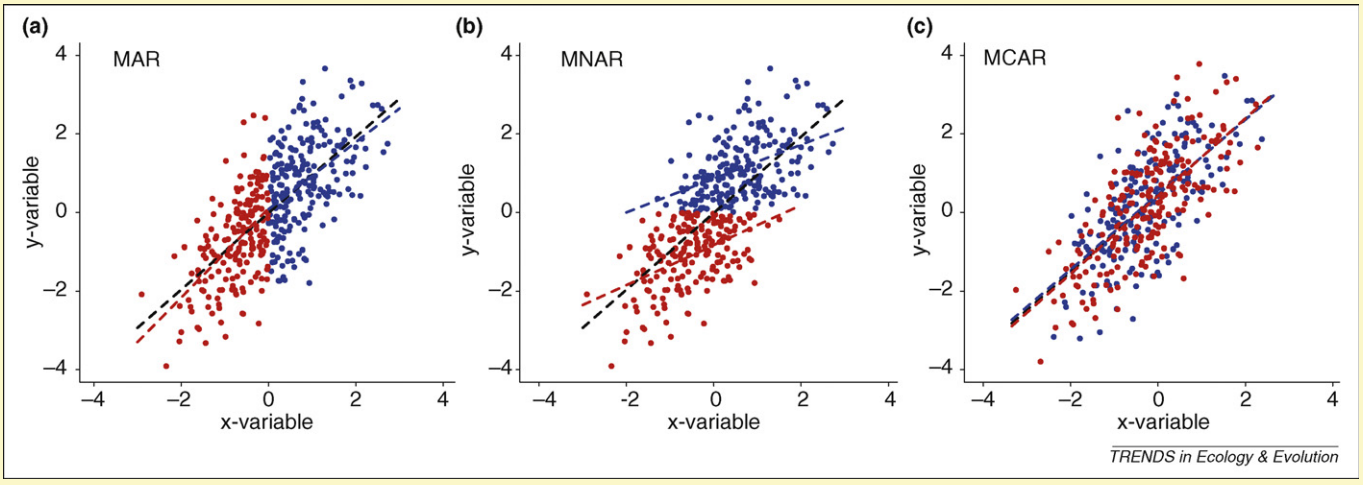

Figure I. Illustrations of the classification for the mechanism of missing data. Blue points are observations whereas red points are missing observations in the *y*-variable; statistics for complete data (blue and red combined) are slope (b) = 1, standard error (se) = 0.05 and R^2 = 0.5. Assuming observations in the *x*-variable are complete, **(a)** represents missing at random (MAR), **(b)** represents missing not at random (MNAR) and **(c)** represents missing completely at random (MCAR). For the observed data (blue points), the estimated slope, se and R^2, are (a) b = 0.86, se = 0.11, R^2 = 0.29, (b) b = 0.432, se = 0.06, R^2 = 0.23 and (c) b = 0.957, se = 0.07, R^2 = 0.49.

censored survival data, it is generally assumed that the model of survival between censoring points is consistent with what is observed at the censoring points. For example, imagine that the population mean body size increases systematically between three censoring points (because small individuals are the first to die). Then, it is reasonable to assume that among those individuals who died in the first interval, the smaller ones died shortly after the first censoring point, and the larger ones died shortly before the second censoring point. However, this assumption might not be justified in certain situations (e.g. involving traits with less predictable expression or development, such as secondary sexual characters).

In the quantitative genetic framework, as Hadfield [10] points out, this assumption on the invisible fraction (e.g. relationship between lifespan and a trait) can be verified and adjusted with pedigree information. This is because some information on trait values of individuals that died prematurely can be obtained from observed trait values of their relatives that exhibit similar trait values (e.g. the lifespan of an individual that died without a weight measurement can be compared with the weights of its relatives to verify the relationship between lifespan and weight). Thus, if missing observations are modelled accurately with a pedigree, the bias in heritability and selection estimates can be reduced.

One of the major conclusions of the paper by Hadfield was surprise regarding the neglect of missing data in evolutionary biology in general, and a call for more attention to this problem [10]. Indeed, missing observations in ecology and evolution might often not be MCAR. For example, in addition to the situation described by Hadfield