



**Universitat  
Pompeu Fabra**  
*Barcelona*

Department  
of Economics and Business

**Economics Working Paper Series**

**Working Paper No. 1558**

**Data reporting and visualization in  
ecology**

**Michael Greenacre**

**February 2017**

# **Data reporting and visualization in ecology**

Michael Greenacre

*Department of Economics and Business, Universitat Pompeu Fabra, and Barcelona*

*Graduate School of Economics, Ramon Trias Fargas 25-27, E-08005 Barcelona, Spain.*

e-mail: [michael.greenacre@upf.edu](mailto:michael.greenacre@upf.edu)

## **Abstract**

The reporting and graphing of ecological data and statistical results often leave a lot to be desired. One reason can be a misunderstanding or confusion of some basic concepts in statistics such as standard deviation, standard error, margin of error, confidence interval, skewness of distribution and correlation. The implications of having small sample sizes are also often glossed over. In several situations, statistics and associated graphical representations are made for comparing groups of samples, where the issues become even more complex. Here, I aim to clarify these basic concepts and ways of reporting and visualizing summaries of variables in ecological research, both for single variables as well as pairs of variables. Specific recommendations about better practice are made, for example describing precision of the mean by the margin of error and bootstrapping to obtain confidence intervals. The role of the logarithmic transformation of positive data is described, as well as its implications in the reporting of results in multiplicative rather than additive form. Comments are also made about ordination plots derived from multivariate analyses, such as principal component analysis and canonical correspondence analysis, with suggested improvements. Some data sets from this Kongsfjord special issue are amongst those used as examples.

**Keywords:** confidence interval, confidence plot, logarithmic transformation, margin of error, ordination, skewness, standard deviation, standard error.

## **Introduction**

Quantitative ecological research involves collecting data, either through observation or experimentation, then analysing the data using statistical methodology, and finally reporting and interpreting the results. In this last stage, either for research articles, reports or presentations, data and results are summarized in tables and figures.

However, the numerical tabulations as well as the visual display of the data are often poor and can be improved. For example, I often find that the selected graphical styles serve to disguise the data rather than to show them! Visual literacy and graphical design are subjects that are not specifically taught in a regular scientific curriculum, and most scientific publications tend to concentrate more on correct textual expression than on the format of the graphics and tables.

Another problem is the choice of statistics that are the basis of the data presentation, both in tabular or graphical form. The best graphic in the world cannot improve unsuitable statistical analysis, so careful attention has to be paid to aspects such as appropriate summary measures, scale of measurement and data transformations.

Having treated these aspects properly, data presentation and visualization can be performed. Articles in certain fields of ecology tend to follow the style of previous articles when it comes to the choice of statistics and graphics, and there is a certain resistance to change. One reason could be the existence of statistical packages with their prepared analytical and graphical options, with few researchers being willing to extend their programming skills to create innovative graphics in environments such as R and Matlab.

The dangers of incorrect use of statistical summary measures and graphical presentations are numerous. Groups can appear to be similar when in fact they are

statistically different. Or, conversely, groups can be judged different when in fact there is no significant difference between them. Variables can seem unrelated when there are indeed interesting relationships amongst them, or their actual inter-relationships can be incorrectly inferred. Comparisons of results from similar studies, where only data and model summaries are reported, not the original data, can suffer from incorrect judgements due to the misspecification or misrepresentation of the data distribution. For example, since positive data are often skew and have some very high values, summarizing them as if their distribution were symmetric can miss these upper tails of the distribution completely, leading to incorrect comparisons of the same variable observed in two comparable studies.

I will consider some examples of univariate, bivariate and multivariate data, some of which are extracted from articles in this special issue of *Polar Biology*, in order to convince readers that there are better ways to summarize and visualize their data. This will involve clarifying some statistical concepts, especially those of variability, which in my experience many researchers find confusing. Thus, I hope that this article leads to a radical change in the way ecological data are considered, and convinces researchers to leave behind the inappropriate, yet apparently accepted, ways of reporting data numerically and visually.

### **Some typical univariate data**

Ecological and biological data are often characterized by unusual distributions, not resembling the idealized bell-shaped and symmetric normal distribution, and sometimes quite small samples. The data in Fig. 1a are a typical example: 15 values of egg counts

of a species in a control group of an experiment (later a set of values for a treatment group will be introduced as a comparison with the control).

Before showing some typical misleading plots of these data, let me point out that I could have alternatively presented the data set in one of the two equivalent forms shown in Fig. 1b. This is a very neat way of showing all the data values as well as a visual impression of their distribution, called a "stem-and-leaf" plot, due to the most innovative of data analysts, John Tukey (Tukey 1977). The "stems" of the plot are 0 (for values 0 to 9), 1 (for values 10 to 19), 2 (for values 20 to 29) and 3 (for values 30 to 39). The "leaves" are the second digit: hence there are 4 0s, 3 1s, 2 2s, a 3 and an 8, then values of 16, 20, 26 and 31. All the data values are given, and in addition the approximate shape of the distribution can be seen, very positively skew, as most count data are, with a lot of low values and a few high values. After a simple sorting of the data in ascending order, this plot can be constructed very easily by simple keyboard entry, especially the horizontal form on the left. But clearly it is only appropriate to certain data sets where the first two or at most three significant digits are accurate enough to report the values. I just mention stem-and-leaf plots in passing to demonstrate that there is always room for innovation in data presentation, and one does not necessarily have to follow the same ways that are inculcated in the literature.

### **Numerical reporting of summary statistics**

Now consider some of the ways these data are typically summarized, first numerically in terms of the mean and three alternative measures of dispersion:

mean	SD	mean	SE	mean	ME
7.40	10.54	7.40	2.72	7.40	5.82

where SD=standard deviation, SE=standard error, ME=margin of error. Each measure of dispersion has its own objective and interpretation. The first is the standard deviation (SD), which is the square root of the variance. The variance in turn is the average of the squared differences of the values from the mean. If one considers all the numerical observations on a number line, a geometric interpretation of the mean is that it is the closest point to all the observations as measured by sum-of-squared distances, and the variance is the value of the measure of closeness that is minimized. Thus, in the simple dot plot representation of the data in Fig. 1c, the mean is the point on the line that minimizes the sum-of-squared distances to all the points<sup>1</sup>. Some points have very large squared distances to the mean, for example the rightmost point 31 has a squared distance of  $(31-7.40)^2 = 557$  from the mean, while the four points piling up at the value 0 each have a squared distance of  $7.40^2 = 55$  from the mean. Thinking about the mean in this way makes it obvious that its value will be highly influenced by outlying data points, which pull the mean towards them. The squared distances are on average<sup>2</sup> equal to 111.1, which is the variance, the minimum value that can be obtained in measuring squared distances of the data to a point on the line. The square root of the variance is the SD, equal to 10.54 here (notice that the SD is the same unit of measurement as raw data,

---

<sup>1</sup> This same least-squares principle is used in a more general form in multivariate ordination methods, to define closest subspaces to a set of points, to be treated later.

<sup>2</sup> Strictly speaking, for theoretical reasons, the sum-of-squared distances (i.e. the sum of the squared deviations from the mean) is divided by  $n - 1$ , not the sample size  $n$ , to obtain the "average". In practice, this only makes noticeable differences in the case of very small sample sizes.

and thus the mean too). However, the SD is a useful single measure of spread of the data only when data are distributed symmetrically about the mean. If the data happen to be distributed according to a normal distribution (we usually say "normally distributed"), then approximately two SDs on both sides of the mean include 95 % of the observations. This way of thinking about the SD as a single all-purpose measure of spread of the data is clearly inappropriate in this case where the data values are highly skewed. Even one standard deviation on the left hand side of the mean extends into negative values, which is impossible for the egg count variable. One can partially solve this problem by separating the terms of the variance calculation into two parts, those for values above the mean and those below the mean. Then the average sum of squares of values below the mean is computed separately from those above the mean, leading to two measures of standard deviation, which in this example would be 6.82 to the left of the mean and 15.07 to the right, clearly very asymmetric and not suitable for applying the 2SD rule below and above the mean (for a single SD estimate). If one is prepared to summarize the dispersion of the data by two numbers, then a straightforward alternative is to report estimates of the 0.025 and 0.975 quantiles (i.e., 2.5 % and 97.5 % percentiles) of the data set, which give an estimate of the limits of 95 % of the observed values: in this example this gives the interval [ 0 , 29.25 ], rounded to [ 0 , 29 ] for these integer-valued data<sup>3</sup>. In medical reporting, benchmark values for a particular parameter in a population are given in the form of a *reference range* (or *reference interval*), usually with 95 % coverage; for example, Total Cholesterol: 120 to 200 mg/dL. In the present context, it is not necessarily a normal group that is being

---

<sup>3</sup> The lower 0.025 quantile is clearly 0, while the upper 0.975 quantile is based on an interpolation between the 14th and 15th ordered values of 26 and 31, and closer to the 31 than the 26. There are at least nine slightly different ways of computing this interpolation, as detailed in the documentation of the R function `quantile`, the default option of which was used to obtain the estimate of 29.25.



described, so I suggest using the term *dispersion range* (or *dispersion interval*), which can be quantified with the percentage such as 95 %. Hence, the estimated 95 % dispersion interval for the egg counts in the control group would be 0 to 29.

To understand the difference between a standard deviation (SD) and a standard error (SE) one has to make a clear distinction between the variability of the original data and the variability of means computed on potentially infinite repetitions of the sampling, each of the same size as the original sample, from the same parent population (this is the aspect that I find many students get confused about). This subject is treated in all introductory statistics books, a good example of which is Quinn and Keough (2002, chapter 2), aimed at biologists. In this example we would like to know how the means would vary when computed for repeated (hypothetical) samples of size 15 from the same original distribution, in other words what other possible values of the mean could have been obtained. Intuitively, means based respectively on multiple samples will be less dispersed, i.e. less variable and more stable, compared to the dispersion of the individual observations. The mean also starts to become approximately normally distributed as the sample size increases (see below). The SE is just the SD of the mean, so its reporting implies that interest lies in describing the behaviour of the mean itself and not in describing the dispersion of the original data. The SE is equal to the SD divided by the square root of the sample size: in this example,  $SE = 10.54/\sqrt{15} = 10.54/3.87 = 2.72$ . Once more, whereas the SD refers to the variability of the original individual data values, the SE refers to the variability of means computed, notionally at least, on multiple samples (in this case, samples each of size 15). The SE is obviously less than the SD and diminishes in value as the sample size increases and the mean becomes estimated more precisely. In practice, to describe dispersion researchers

use the SD or SE with approximately equal frequency in research papers, according to Krzywinski and Altman (2013).

If interest is in the mean of a variable, possibly for comparing with means computed for other groups of samples, then one wishes to know the variability in the mean that is reasonably expected due to sampling, because the mean would be different if you had sampled again. This brings us to the margin of error (ME), which leads to the definition of a *confidence interval* for the mean. Curiously, even though this is the most useful way of expressing knowledge about the mean's variability, it is rarely reported in the ecological literature. The ME is approximately equal to the standard error multiplied by 2, which comes from normal theory (where the exact multiplying factor is 1.96 for large sample sizes) – we often say "two standard errors around the mean". More accurately than 2, the value is obtained from the t-distribution with degrees of freedom one less than the sample size  $n$ , i.e.  $n-1$ . In the present example, where  $n = 15$ , the value corresponding to the t-distribution with 14 degrees of freedom is 2.14, which is the value used to compute the ME in this example:  $ME = 2.14 \times 2.72 = 5.82$ . Especially for "large" ( $>30$ ) samples, the ME can justifiably be reported with the prefix  $\pm$  ("plus or minus"), because it is the value that can be added to and subtracted from the mean to give a 95 % confidence interval. In the present example where  $n = 15$  and the data are quite asymmetrically distributed, the confidence interval would be roughly approximated as  $[7.40 - 5.82, 7.40 + 5.82] = [1.58, 13.22]$  – in the next section the confidence interval will be shown to be slightly asymmetric. A 95 % confidence interval includes the true mean of the underlying distribution with a probability of 0.95. Equivalently, the probability is 0.05 that the interval does not include the true mean, which is the conventional low level of risk that researchers are prepared to take that their estimated confidence interval is "off target". The  $\pm$  prefix is often used,

erroneously in my opinion, for the one SD and one SE forms – they do not give useful intervals, since  $\pm$ SE gives a 68 % confidence interval, which is not intuitive. The  $\pm$ ME form, on the other hand, includes the true mean with a probability of 0.95 (i.e. 95 % confidence interval), which is more familiar and conventionally used to indicate high confidence about the mean's true value.

In summary, in the case of ecological data, which are most often not symmetrically distributed, the SD is a poor dispersion measure of the distribution. To measure precision of the mean, the SE in itself is not useful, but it is stepping-stone to compute the ME, which provides a useful confidence interval for the mean. So my first recommendations are the following:

1. Avoid reporting SD and SE, but if you have to report one of them don't use  $\pm$  to precede them. An option is to put them inside parentheses after the mean.
2. If you have to report SDs (assuming then that your aim is to summarize the behaviour of the individual data, not the mean), check the skewness of your observations and consider the possibility of reporting estimated quantiles that include 95 % of the data distribution, giving what I call a dispersion interval.
3. If interest is in the behaviour of the mean rather than the original individual values, for example in group comparisons, report the margin of error (ME). The ME defines the actual precision of the estimate, namely the interval that most likely includes the true mean, i.e. with high confidence (usually 95 %). The ME can justifiably be preceded with the sign  $\pm$  because it applies to the distribution of the mean, which is usually quite symmetrical, especially in larger samples ( $n > 30$ ).

## Graphical ways of showing the distribution

When it comes to visualizing a set of data graphically, the situation is even more problematic, especially for small samples. Summary statistics of the egg count data are displayed in 11 different ways in Fig. 2.

The dot plot (Fig. 2a, the vertical equivalent of Fig. 1c) shows the complete situation, whereas all the others are summaries. The boxplot (Fig. 2b), also invented by Tukey (1977), is one of the best ways shown here to summarize the distribution of the original individual observations (but not for very small samples, say less than 10). Instead of the mean, Tukey's boxplot shows the median (i.e. the 50th percentile) as the horizontal bar, then the 25th and 75th percentiles as a box, and whiskers extending to the minimum and maximum values. Some versions of the boxplot include a heuristic rule that identifies outliers, as illustrated in this particular case where the highest value of 31 is signalled as an outlier – see Wickham and Stryjewski, 2011). The upper whisker is then extended to the next highest value, which is 26 in this example. The boxplot clearly shows that the distribution of egg counts is very skewed.

Frequently, biologists display the mean of a variable as a bar rising from zero to the mean value. Fig. 2c is a typical example, adding a whisker equal to 1 SD to the mean value. In my opinion, this is one of the worst summaries of the distribution, and Fig. 2d, where a 1 SD whisker is added as well as subtracted from the mean (another typical display used by biologists) shows the reason. There is no hint of the skewness of the distribution and the standard deviation is so high that plus/minus 1 SD extends into negative values in this particular example, which is impossible. One could rather say that this is the best way of disguising the actual data! This problem becomes acute and almost nonsensical for very small samples: for example, for a sample of size 2 (Fig. 3a),

the mean is at the midpoint of the two observations and the standard deviation is approximately 0.7 times the range of the two observations (the standard error would be half the range); and for a sample of size 3 (Fig. 3b), exactly the same bar and whiskers can be obtained for samples with three completely different configurations. Thus, for small samples, the best option is to show the actual data in a dot plot and dispense with any summaries of variability.

Returning to Figure 2, the asymmetric SD plot (Fig. 2e) gives a more realistic impression of the distribution below and above the mean, so this form can be used if one really insists on using the standard deviation measure, but two summary measures of dispersion are involved (i.e., 6.82, 15.07). The quantile plot (Fig. 2f), showing the estimated extent of 95 % of the data's distribution, is also acceptable, but again involves two different values to be reported numerically (i.e., 0, 29.25).

The remaining plots (Figs 2f-j) for SE and ME are trying to summarize the distribution of the mean, not the distribution of the individual data values, and the distribution of the mean will be less skewed than the parent distribution, thanks to the central limit theorem. This fundamental theorem of statistics states that as the sample size increases, the distribution of the mean of the sample tends to become normal, no matter what the original distribution of the individual values is. But the applicability of this theorem depends on the sample size. Conventionally, this theorem can be regarded to "kick in" seriously for samples of 30 or more, although if the data are not too skewed to begin with, means will be approximately normally distributed for smaller sample sizes. In the present example of Fig. 1, the data are very skewed and the sample size is only 15, so the theorem is not wholly operative.

As explained before, there is no clear reason why the SE should be of interest graphically (Figs 2g and h), since it is rather the ME that is interpretable. The double-whisker version of the ME plot (Fig. 2j) is an acceptable option, but not the best, since it assumes symmetry in the distribution of the mean. If I had to choose the best option for displaying the mean's variation, I would choose Fig. 2k, based on "bootstrapping" the mean – this is explained in the next section. Notice that the confidence interval in Fig. 2k is asymmetric, as might be expected in this example of a highly skewed distribution of the original data and a relatively small sample size. The way this can be reported numerically is to state the mean, which is 7.4, and then the confidence interval, either as an interval in parentheses, (3.2, 13.7), or in the range format 3.2 – 13.7, or 3.2 to 13.7. An alternative notation is  $7.4^{+6.3}_{-4.2}$ , i.e. the mean and the different MEs respectively above and below the mean shown as super- and subscripts.

My next recommendations are:

4. Boxplots or quantile plots are a realistic representation of the distribution of the original data.
5. For displaying variability of the mean, always choose margins of error rather than standard deviations or standard errors, since margins of error define confidence intervals (usually at 95 % confidence level).
6. Consider the alternative option of the bootstrap confidence interval, which assumes that the sample is representative of the population and does not rely on any assumption of the distribution – see the next section.

## Skewness and small samples

As sample size increases, the distribution of the mean becomes more symmetric. As stated above, a rule of thumb is that sample sizes greater than 30 can be considered large enough for normality of the mean to be assumed. In this case, the ME can be computed, using the rule of thumb of 2 times the SE (or, more accurately, the critical point of the appropriate t-distribution), and then the distribution of the mean can be summarized numerically or graphically as  $\text{mean} \pm \text{ME}$ . In the present example of sample size 15, however, we should be a bit more cautious, since it may well be that the confidence interval for the mean is not symmetric, given that the data themselves are highly skewed. There are two ways to approach the problem, either transform the data to make the parent distribution more symmetric, even close to normal if we are lucky, or use the modern resampling-based approach of bootstrapping (see, for example, David and Hinkley 1997).

One of the standard transformations of positive data is the log-transform (see later section below), but this is tricky here since there are data zeros, which complicates the situation even more, but there are theoretical treatments of this case, as indicated by Tian (2005). Adding a constant to the value before taking logs is a "cosmetic" way out of this problem, for example  $\log(1+x)$  in this case for integer count data, but the choice of the constant and its effect on the results is usually not considered. Taking a root transformation, usually the square root or fourth root, is another cosmetic solution, often with no justification of the choice of root transformation other than it "acceptably" reduces the skewness of the distribution. Interestingly, the cube root (i.e., third root) is never considered, showing how arbitrary the choice is.

An alternative approach is to use the bootstrap. Bootstrapping has become more and more prevalent in package software, e.g. Stata (Statcorp. 2015), and is easy to execute in R, using the `boot` package, for example. The idea is to take a large number (in our application, 10000) of samples of size 15 of the original sample, with replacement, and to compute the mean on each one, giving an approximate empirical distribution of the mean, from which any percentiles can be estimated. A summary of the distribution of these simulated means (Fig. 4) shows the estimated confidence interval, where whiskers extend from the mean (shown as a dot), to the upper 97.5<sup>th</sup> and lower 2.5<sup>th</sup> percentiles, thus giving an estimated 95 % confidence interval for the mean, which I call a *confidence plot* (these are the same limits used in Fig. 2k). To enhance the interval in a style similar to the boxplot, a box has been added to show the boundaries of the 75<sup>th</sup> and 25<sup>th</sup> percentiles, thus enclosing 50 % of the mean's distribution. Thus there is a 50:50 chance that the true mean lies within the box, and a 95 % chance that it lies in the range of the interval. (Coincidentally and independently, in a recently published book Gierlinski (2015, Figure 1-1) makes similar plots to summarize the distribution of individual values, where "boxes encompass data between the 25th and 75th percentile; whiskers span between the 5th and 95th percentiles", hence a 90 % dispersion interval in that case). In my confidence plot, the box is of secondary importance and can be omitted, especially if it is confused with Tukey's boxplot. Once again, at the risk of repetition, the boxplot, such as the one in Fig. 2b, shows the variability in individual values, whereas the confidence plot in Fig. 4 shows the variability in the mean.

The bootstrap should not be used for very small samples, however, where the sample distribution is often a poor image of the population distribution – generally, for  $n = 10$  and higher it starts to be acceptable, but not for smaller samples. There are



enhancements to the bootstrap that improve its properties – see, for example, Good (2005, Section 3.3.3).

A further recommendation is thus:

7. For sample sizes 30 or higher, the mean  $\pm$  ME gives a good approximation to the confidence interval. For smaller samples, and for data from very skewed or non-normal distributions, a realistic confidence interval for the mean is achieved by bootstrapping the sample to obtain an empirical distribution of the mean, and thus any percentiles. The bootstrapping option for drawing confidence plots is valid for any sample size (except very small ones, less than 10, say), and is distribution-free, so is a good choice in general for all cases.

The format of the confidence plot (Fig. 4) lends itself to comparisons between means of different data sets, which is the subject of the next section.

### **Comparisons of two sets of data**

Suppose that another data set of 15 egg count observations is available for a treatment group, with values shown in stem-and-leaf form side-by-side with the previous control data, as well as a "back-to-back" version for an even easier comparison (Fig. 5a). The treatment group has values generally higher than those in the control group, but is the difference in the group means statistically significant? Using the confidence plot style (Fig. 4) and bootstrapping the mean of the new data set gives Fig. 5b. The confidence intervals overlap, which gives the impression that the means are not significantly

different. However, the overlapping of two 95 % confidence intervals does not necessarily imply non-significance of a difference at the 5 % significance level. No overlap, however, almost surely implies significance, as pointed out by several authors, for example Krzywinski and Altman (2013). The 95 % confidence plot is thus conservative when comparing groups, but being conservative is, in any case, an advantage when several comparisons are being made. The graphical display is not intended to be an inferential instrument, only an indication of difference. In all cases, whether confidence intervals overlap or not, the actual p-value should be computed – in this example an exact permutation test, using the `coin` package in **R**, gives a p-value of 0.06 (and consistently above 0.05 for other randomizations), so the means are not significantly different at the 0.05 significance level.

So the next recommendation is:

8. Use 95 % confidence intervals in plots that are comparing means from different data sets, with the understanding that these are conservative for showing significant differences. The actual hypothesis test that is appropriate for the data, which leads to a p-value, is the accurate way of arriving at a conclusion. If the confidence intervals do not overlap, then it is highly likely that the means will be found to be significantly different.

### **The logarithmic transformation and multiplicative effects**

Figs 4 and 5b showed non-symmetric confidence intervals for the mean, achieved by bootstrap resampling. The errors above and below the mean are different because of the skewness of the data and the small sample size, so to report the confidence interval

numerically one would have to specify two errors, which may be cumbersome. In describing the distribution of the original data, or the distribution of the mean, it is common practice to report a table such as Table 1 (extracted from Table 1 of Piquet et al., 2016, this volume), showing means and SDs of temperature and NO<sub>x</sub> (nitrates & nitrites) for two seasons at five stations. These values are supposed to summarize a central value (the mean) and a deviation from this value (the SD) for the raw data. As discussed earlier, the use of the symbol  $\pm$  really has no meaning here and should be dropped if the SDs are required to be reported, similarly if SEs are reported. For the temperature data, the addition and subtraction of some quantity to describe the distribution has some inherent sense, because temperature is considered a variable on an interval scale, can take negative values and might even be approximately normally distributed. Adding and subtracting a quantity to the NO<sub>x</sub> mean, on the other hand, can lead to paradoxes such as the "Late Spring" values, several of which drop below zero if the SD is subtracted from the mean, e.g.  $1.08 \pm 2.19$  for sample "M", and  $0.80 \pm 1.10$  for sample "KG". A solution to this dilemma is to think of variables that measure positive quantities on a ratio scale, which implies that any differences, and in particular measures of variability, are expressed as multiplicative (or percentage) differences.

It is true that almost every computation involved in regular statistical analysis is one of adding (e.g. summing values to obtain a mean), or subtracting (e.g. computing differences to obtain a variance). Confidence intervals for estimates such as means, group differences and regression coefficients are often expressed by the estimated value plus or minus a single value, the ME. For ratio-scale data, however, one should be thinking of multiplying and dividing rather than adding or subtracting. A good test of how data differences are inherently considered is the following: suppose you have two

samples A and B in a nutrient analysis with these values of inorganic phosphorus and NO<sub>x</sub>:

	Inorganic phosphorus (μM)	NO <sub>x</sub> (μM)
Sample A	0.50	5.00
Sample B	0.52	5.02

For both variables, which use the same measurement scale (μM), sample B contains 0.02 units more than sample A. But are these the same differences, or is the difference in inorganic phosphorus larger, because 0.52 is 4 % more than 0.50 whereas 5.02 is only 0.4 % higher than 5.00? If the latter case is preferable for comparing the samples, then the researcher is – consciously or unconsciously – thinking of the scale as multiplicative and not additive. And the simplest way to pass from a multiplicative scale to an additive one, so that all the usual additive statistical methodology can be applied, is to transform to logarithms. On the log scale the differences are:

Inorganic phosphorus:	NO <sub>x</sub> :
$\log(0.52) - \log(0.50) = 0.03922$	$\log(5.02) - \log(5.00) = 0.00399$

showing that the difference for the first variable is measured about 10 times larger than the second one. The property of the logarithm to convert ratios to interval differences is exemplified by the identity:  $\log(x/y) = \log(x) - \log(y)$ . Moreover, positive data are often skewed to the right, in which case the log-transform pulls in the high values and makes the distribution more symmetric, which is an additional benefit for many statistical model assumptions.

There are, however, some notable repercussions of this transformation and of seemingly innocuous statements found in many "Methods" descriptions, such as "Data were log-transformed prior to analysis" (quoted from an article in this special issue). If one really

needs a single number for reporting the dispersion of a ratio-scale variable, one needs a statistical model for the underlying distribution of the data, and the log-normal distribution provides the most obvious solution, just like the normal distribution is natural for symmetrically distributed interval-scale data. The log-normal distribution is the distribution of a variable  $X$ , whose log-transform  $\log(X)$  is normally distributed. Thus, doing classical statistics like computing means, variances and regressions on log-transformed data, using the normal distribution, implies that one is assuming the log-normal distribution as the underlying distribution of the untransformed data. However, computing the mean and SD on log-transformed data and then simply back-transforming to the mean and SD on the original scale, using the exponential function, does not give the correct estimates of the mean and SD of the log-normally distributed variable. Because the log-transformation is not linear, the mean of the logged data is not the correct estimate of the log of the mean. The acceptable way of transforming back to the original scale involves a bias-correction, which is well-known in the statistical literature, but not so familiar to ecologists. It results in a SD "factor", a value higher than 1, which multiplies and divides the estimated mean in order to give a region of dispersion, rather than a SD that is added and subtracted (for more technical details, see Aitchison and Brown (1957) or Feng et al. (2014) – the method is programmed in the R package `EnvStats` by Millard (2013)). I illustrate this alternative approach with the data set from the study of Piquet et al. (2016, this volume), part of which was given earlier in Table 1. In this table there is a reported mean of 0.80 and SD of 1.10 for the variable NO<sub>x</sub>, for a subsample of the data defined by season "Late spring", and station "KG". The NO<sub>x</sub> data for this subsample, with  $n = 10$  observations, along with a dotplot of the values and a boxplot summary, are given in Fig. 6.

The data are clearly very skewed, with an outlying high value of 3.71, which has caused the SD to be very high. Summarizing these data by their mean and SD, which implicitly assumes a normal – or, at least, a symmetric – distribution, once again does not reflect the true nature of these data, like the egg count example discussed before. The model-based alternative introduced in this section is to assume that the data follow a log-normal distribution. The distributional assumption can be checked visually by doing a quantile plot. For normally distributed data the plotted values should lie approximately on a straight line, and can be more formally verified by a test of normality, e.g. the Shapiro-Wilks test. Fig. 7 shows that for the original untransformed NOx values the pattern of points in the quantile plot is convex, indicating right-skewness, and normality is soundly rejected (Shapiro-Wilks test for normality has p-value  $p = 0.0004$ ), whereas for the log-transformed data the pattern is approximately linear and normality is accepted ( $p = 0.99$ ). Thus, the log-normal distribution is a good assumption here.

Here are some summary statistics of this data set, where the variable NOx is denoted by  $x$ , and the log-transformed variable by  $y = \log(x)$ :

$$\begin{array}{lll} \text{mean of } x: & \bar{x} = 0.796 & \text{mean of } y: & \bar{y} = -0.922 & \exp(\bar{y}) = 0.398 \\ \text{SD of } x: & s_x = 1.106 & \text{SD of } y: & s_y = 1.278 & \exp(s_y) = 3.590 \end{array}$$

The SD  $s_y$  is also larger than the mean  $\bar{y}$ , but logarithms are allowed to be negative, so one might be tempted to compute statistics on the log-scale and then transform back to the original one using the exponential function. This is not correct, however, and all estimates will be biased: for example, the exponential of  $\bar{y}$ ,  $\exp(-0.922) = 0.398$ , is a severe under-estimate of the mean of the log-normally distributed  $x$  in this case. The theory of the log-normal distribution leads to the following estimates of the mean of the

log-normal and a *multiplicative* factor defining its variability (see, for example, Aitchison and Brown 1957, Parkin et al. 1990, Limpert et al. 2001).

First, the mean of  $x$  is estimated by exponentiating  $\bar{y}$  plus a term that depends on the variance  $s_y^2$ :

$$\text{Estimated mean: } \exp(\bar{y} + s_y^2/2) = \exp(-0.922 + 1.278^2/2) = 0.900 \quad (1)$$

i.e. the back-transformed value of  $\exp(\bar{y}) = 0.398$  is multiplied by the factor  $\exp(s_y^2/2) = \exp(1.278^2/2) = 2.263$ .

Second, the standard deviation is this estimated mean multiplied by  $\sqrt{\exp(s_y^2) - 1}$  :

$$\text{Estimated SD: } \exp(\bar{y} + s_y^2/2) \sqrt{\exp(s_y^2) - 1} = 0.900 \times 2.030 = 1.827 \quad (2)$$

Hence, the quantity  $\sqrt{\exp(s_y^2) - 1}$  is equal to the SD divided by the mean, which is the *coefficient of variation* of the log-normal distribution.

The estimated SD in (2) is a multiplicative factor that has to be multiplied, or divided, by the mean to get one standard deviation above or one standard deviation below the mean:  $0.900 / 1.827 = 0.493$  and  $0.900 \times 1.827 = 1.644$ . Two SDs above and below the mean would be calculated as  $0.900 / 1.827^2 = 0.270$  and  $0.900 \times 1.827^2 = 3.004$  (notice that the multiplicative factor is squared, not doubled). These would be the limits of an interval including approximately 95 % of the data (for this particular sample it covers only 6 out of the 10 actual observations, but this is due to the low sample size).

In summary, an alternative model-based way of reporting a mean and the variability of positive observations considered to be on a ratio scale and lognormally distributed, is given by the estimated mean of a log-normal distribution, along with a multiplicative

SD factor that is always greater than 1, in this case 0.900 (1.827) – this would replace Piquet et al.'s mean and SD of 0.80 (1.10).

Computing a 95 % confidence interval for the mean of lognormally distributed data involves a more complicated calculation (see, for example, Land 1972, Parkin et al. 1990), but – similar to the SD – the ME is a factor that multiplies and divides the mean to get the interval limits. In this example, this factor turns out to be 3.343, so rather than "plus-minus", a "multiply-divide" symbol is required, e.g.  $\times\div$ , or the symbol  $\times/$  proposed by Limpert et al. (2001): 0.900  $\times/$  3.343. The confidence interval is (0.269, 3.009).

This "multiply-divide" style is such a radical change in reporting summary statistics of positive-valued right-skewed data in the biological literature, that I do not recommend it (at least, not yet), even though it is a logically consistent model-based approach to the summarising of such ratio-scale data, analogous to the usual normal distribution approach to interval-scale data. The above results depend on the data being lognormally distributed and this is not always the case, and the only benefit of this approach appears to be the ability to summarize dispersion by a single number (the multiplicative SD or ME factors) rather than two numbers (the lower and upper limits of an interval, or the deviations below and above the estimated mean).

This description of the lognormal case is included here so that researchers can realize the non-trivial consequences of a log-transformation, or any other non-linear transformation such as the square root for that matter. The consequences for the estimation of the mean are particularly important. For estimating a confidence interval for the mean, a nonparametric distribution-free alternative such as the bootstrap confidence interval is one of the most suitable. Another important comment about the



use of the log-transformation in the statistical modelling of count data is that it can be avoided, as maintained by O'Hara and Kotze (2010), who prefer using the correct theoretical distribution, which can be Poisson or negative binomial, amongst others, depending on the data's properties.

### **Bivariate data**

Considering two variables simultaneously is also common in ecology, for example the joint analysis of stable isotopes of carbon ( $\delta^{13}\text{C}$ ) and nitrogen ( $\delta^{15}\text{N}$ ) – see Huenerlage, Graeve and Buchholz (2016, this volume, Figure 5), who show a scatterplot of the means of five krill species, and margin of error bars in the directions of the "x" ( $\delta^{13}\text{C}$ ) and "y" ( $\delta^{15}\text{N}$ ) axes. This plot is repeated here (Fig. 8a), but using the distribution-free bootstrap approach described earlier, showing rather symmetric confidence intervals, thanks to the original data not being particularly skewed. Hence, there is no noticeable difference between Fig. 8a and Huenerlage et al.'s Figure 5, which is based on normal theory. The symmetry in this case gives some assurance that the means of such small samples (of sizes 21, 9, 6, 5 and 5, respectively, for the five species abbreviated as Ti, Tr, TL, Mn and Nm) might be close to normal.

As a comparison with the 95 % confidence intervals, Fig. 8b shows the 68 % confidence intervals, based on plus/minus a single SE and hence much shorter. Fig. 8b is again not recommended, since it gives a too optimistic view of the group differences. This can be seen in Table 2, which details the overlapping or not of the plotted intervals in the two cases, as well as the actual p-values obtained by the standard t-test between pairs of species and by the distribution-free permutation test (notice that we avoid the issue of multiple comparisons for the moment, to be discussed in the next section). Looking at

the first two columns, there are in a few cases large differences between the p-values obtained by the t-test, restrictive in its assumptions, and the distribution-free exact permutation test. The conservative 95 % confidence intervals perform well when compared to the p-values, while the 68 % confidence intervals are too optimistic, as expected, showing some apparent differences that are not significant according to the respective tests. Given that the 95 % confidence level is a familiar concept to most researchers, it can again be concluded that it is the preferred approach for visualizing precision in means.

The style of Fig. 8 might give the impression that the confidence region for the mean of each species lies in a square area, but these regions tend to be elliptical due to the additional effect of correlation between the two variables. Following the bootstrapping approach, the resampling of the data for each species leads to an estimation of the distribution of the mean in the plane of the two variables, and 95 % confidence regions can be defined, shown as ellipses in Fig. 9. Technically, this is done by generating a large number, 10000 say, of bivariate means on bootstrapped samples, and then the **ellipse** function in the R package **ellipse** is used to compute the elliptical regions. This pattern suggests that species *TL* and *Mn* have different isotopic signals from all the others, and *Tr* is different from *Nm* or, putting it another way, the only non-significant differences are between *Ti* and *Tr* and between *Ti* and *Nm*. These conclusions are confirmed by multivariate ANOVA (MANOVA) tests performed on each pair of species (Table 3).

In summary, recommendations about the display of grouped bivariate data are:

9. Use 95 % confidence ellipses to show the variability of bivariate group means. As for univariate data, if two confidence ellipses do not overlap, then it is highly likely that the two respective groups have means that are significantly different.
10. Nevertheless, the appropriate statistical test, using MANOVA or the distribution-free permutation test, should be conducted to obtain a p-value to quantify the significance of the difference.

### **Multiple testing**

The issue of multiple testing, ignored up to now, can be illustrated using the p-values in Table 3. The conventional significance level of 0.05 (or 5 %), equivalent to a confidence interval of 95 %, means that a risk of 1 in 20 is taken that a result found to be significant is actually just random variation (called the "type I error"). If we do 10 tests, as in Table 3, then this risk is increasing on the individual results to a probability of 0.40 of finding one of the results to be significant by chance. This problem is well-known and often referred to as increasing the *false positive rate*. Hence, some adjustment needs to be made to the significance level, reducing it so that the 5 % risk of the type I error is not increased but applies to the whole family of tests. In statistical jargon, "the false positive rate is controlled so that the family-wise probability of a type I error is 0.05". There are many ways of coping with this problem, the most well-known being the Bonferroni correction – see, for example, Moran (2003). This is a simple division of 0.05 by the number of tests being made, i.e.  $0.05/10 = 0.005$  in the case of Table 3, to obtain the significance level one should use. The Bonferroni correction is

too extreme, since it assumes that all the tests are totally independent of one another, which is clearly not true. Performing all the tests could be avoided by pre-specifying a smaller set of comparisons that are planned beforehand according to some research hypotheses. When all comparisons are of interest in an exploratory analysis, an alternative approach, which is popular in the biostatistical literature and easy to implement, is the Benjamini-Hochberg step-up procedure (Benjamini and Hochberg 1995). The  $m$ , say, p-values of all the tests are ordered from lowest to highest, and in parallel a list is made of the arithmetic series  $0.05 \times \{1, 2, \dots, m\} / m$ , i.e. in our example starting with  $0.05/10$ , and then moving up in steps of  $0.05/10$  until  $0.05$  (assuming that  $0.05$  is the desired overall significance level):

$p$ :	<b>0.0001</b>	<b>0.0006</b>	<b>0.0009</b>	<b>0.001</b>	<b>0.001</b>	<b>0.002</b>	<b>0.003</b>	<b>0.029</b>		0.072	0.099
series:	<b>0.005</b>	<b>0.010</b>	<b>0.015</b>	<b>0.020</b>	<b>0.025</b>	<b>0.030</b>	<b>0.035</b>	<b>0.040</b>		0.045	0.050

As long as the p-value stays below the corresponding value of the series, the p-value can be regarded as significant, which includes all the p-values up to  $0.029$ , to the left of the vertical bar. To the right of the bar the p-values are greater than the corresponding series values. Notice that the series starts at the Bonferroni corrected level and moves in equal steps up to the  $0.05$  level, so it is only the smallest p-value that needs to be below the Bonferroni level, not all of them. If we had used the stricter Bonferroni adjustment for all tests, the difference between  $Tr$  and  $Mn$  would have been declared non-significant, even though  $p = 0.029$ .

Recommendations for controlling the false positive rate when many tests are conducted on the same data are thus:

11. Think carefully, preferably before the data are collected, about which hypothesis tests you want to perform.
12. Eventually gather together all the p-values of this set of  $m$  tests being performed and order them from smallest to largest.
13. Using the Benjamin-Hochberg approach, compare this ordered list to an arithmetic series starting from  $\alpha/m$ , incrementing in steps of  $\alpha/m$  until  $\alpha$  (the overall significance level, usually 0.05). The p-values that are below the corresponding values of the series indicate significant tests.

### **Multivariate data**

It is still possible to visualize data on three variables (i.e. trivariate data) simultaneously by using several three-dimensional orientations of the points or even animations to give a three-dimensional effect. Some examples of dynamic graphics are given by Greenacre and Hastie (2010), and the original animations can be seen at <http://www.econ.upf.edu/~michael/dynamic> (last accessed 1 July 2016). An example of the trivariate means of six groups is shown in Fig. 10, where the equivalent of ellipses in three dimensions, called ellipsoids, describe 95 % confidence volumes for the three-dimensional means. Two different views of these confidence volumes are shown: first in Fig. 10a, where the groups A5 and A6 clearly separate from the others in the direction of variable 1, whereas the others are generally overlapping but with A1 and A4 visibly separated. The different viewpoint in Fig. 10b, where the display has been rotated 90° around the axis of variable 1, shows that the ellipsoids of A2 and A4

are separated, and A1 and A3 overlapping only slightly. Hence, the three-dimensional configuration has to be inspected from different angles to judge whether the ellipsoids really do overlap or not in their three-dimensional positions. Again, the acid test is to compute p-values using MANOVA (reported in Table 4) or a permutation test. Table 4 confirms the visual inspection, showing that A2 and A4 are significantly different as well as A1 and A3.

In this section, however, I will be more concerned with multivariate data that exist in a high-dimensional space, usually much higher than three, so it impossible to see the data exactly. In order to visualize such data, a dimension-reducing algorithm is used to produce "flat" two-dimensional planar versions that minimize loss of information, since information is necessarily lost by any such reduction. In practice, this is exactly the same objective as described previously, where the mean was defined as the point that came closest to all the observations in terms of least squares – the only difference is that here the objective is to identify a plane that comes closest to all the points. This approach is typical of methods such as principal component analysis (PCA), multidimensional scaling (MDS), correspondence analysis (CA), redundancy analysis (RDA) and canonical correspondence analysis (CCA). The general idea in these methods is that the data points have positions in a high-dimensional space, which is impossible to view, but that their positions can be approximated in a two-dimensional space, called an *ordination*. This technology is enhanced by adding points that represent the original variables to the scatter of points in the ordination, called a *biplot*, as well as adding confidence ellipses if points are grouped. Most of the issues for visualizing bivariate data, discussed in the previous section, apply to these two-dimensional ordination plots, whose axes are no longer actual variables, but combinations of several variables that account for the most variation in the data.

As an example, the data of Voronkov et al. (2016, this volume) were re-analysed using CCA, presented in Fig. 11 (see that paper for a description of the samples, which are classified into seven groups, the species variables and the environmental variables used to constrain the solution). There are minor differences here compared to their Figure 7, mostly in scaling factors along the two ordination axes, but these are of no consequence for the general interpretation of the result. As is common for CCA ordinations of this type, there are three possible sets of points in the resultant *triplot*: the samples, the species and the constraining variables. Several enhancements have been added compared to the output of standard software.

Firstly, not all the species are shown in Fig. 11 (notice that in Voronkov et al.'s Figure 7 no species are shown at all, possibly because of the proliferation of points). The coordinates of the species in Fig. 11 are chosen to be the so-called *contribution coordinates*, which show the small subset of species that contribute more than average to one of the ordination dimensions (Greenacre 2012). These highly contributing species have positions in this display far from the centre, whereas all lower-contributing species, with positions close to the centre, have simply been eliminated from the ordination since they are not worth interpreting. But there are three other species that are highly correlated with the ordination, even though they have not contributed highly to it, and these are added with a smaller font size. Filtering out these highly correlated species is performed slightly differently from distinguishing the highly contributing species, which contribute more than average to the CCA solution. The percentages of inertia on the ordination axes (see Fig. 11) are 50.6 % and 16.4 % respectively, which are overall measures of how much the axes explain the data and are the averages over the individual percentages for the species (or for the samples). The squared correlations of individual species with the axes are measures of how much the axes explain each

species. Hence, those species that are explained more than 50.6 % by axis 1 or more than 16.4 % by axis 2 are judged to be correlated more than average.

Secondly, the 95 % confidence ellipses of the means of the seven groups are shown in the ordination. These are actually the projections of the multivariate hyper-ellipsoids onto the two-dimensional ordination, so separation of the ellipses again indicates significant difference, but again a formal test should be conducted. Overlap of the ellipses, however, does not imply that they might in fact be separated in the "full space" of the data – they are just overlapping in the projected ordination space that is the support of the visualization. Only groups 4 and 5 are overlapping in the ordination, and permutation tests on all pairs of groups show that indeed these two groups are not significantly different, whereas all the others are. Notice that several group ellipses do separate horizontally as well as vertically, suggesting that these axes are significant. To confirm this, the permutation test in the `vegan` package (Oksanen et al. 2014) in R shows both these axes to be highly significant ( $p < 0.001$ ).

The ordination thus shows the directions of the environmental variables used to constrain the ordination space, the positions of the samples (numbered points), their group means and confidence regions (larger numbers and ellipses), and the species that are highly contributing to the separation of the samples, with a total of 67 % of the constrained inertia explained. This is a very concise summary of the features of a large data set, showing just the essential features necessary for the ecological interpretation.



When reporting results of ordinations, the following recommendations are relevant:

14. Try to include only those points and symbols that are necessary for the interpretation of the ordination. In this respect, using contribution coordinates and showing only highly contributing species is one way to cut down the number of points showing the biological variables.
15. When sample points are grouped, use confidence regions (as for bivariate data) to show the variability in the group means, and perform multivariate tests of differences between group means, thereby controlling the false positive rate.

### **Discussion and concluding remarks**

In his seminal book on statistical graphics, Tufte (2001) states two basic principles: "1. Above all else, show the data; 2. Avoid chartjunk by maximizing the data to ink ratio". The aim of summarizing observed data into some numbers (e.g., mean or median and some measure(s) of dispersion) or visualizing data in some graphical form, be it univariate, bivariate or multivariate, is to communicate the features in the data to the reader or observer. As I have tried to demonstrate, several tabular and graphical styles in common use can be misleading and should be avoided – the researcher's choice of data summaries and visual presentation should aim to reveal the data's features rather than disguise them. This topic is not often treated in the biological literature beyond the mere description of certain well-known graphical representations found in standard software packages. An exception is the book by Quinn and McKeough (2002), who include a specific chapter on exploratory data analysis and graphics, but with little critical comment on graphics found in present-day literature. The recent paper by Weissgerber

et al. (2015), pointed out by a referee, is closely aligned to my ideas in the present paper, where they say, for example, that "as scientists, we urgently need to change our practices for presenting continuous data in small sample size studies". In their investigation of 703 research articles, they found that "most papers presented continuous data in bar and line graphs", which is "problematic as many different data distributions can lead to the same bar or line graph" (cf. my Fig. 3). These authors go so far as to recommend that journals change their policies and peer review practices so as to "discourage the use of bar graphs" and "provide specific guidance about what types of figures are preferred". They further conclude, coinciding with my experience, that "scientists need better training in how to select the appropriate type of figure for their data".

Concerning the practice of reporting standard deviations or standard errors alongside mean estimates of a variable, I have made a strong case for replacing this with estimated quantiles, when intervals are required to show the dispersion of the data values themselves, or confidence intervals to express precision of the estimated mean. A referee makes the following defence of the standard deviation: "...even with asymmetric distributions it can still be useful, as a larger SD means more spread". The problem with this statement is that spread can be measured in many different ways, so why use the SD rather than, say, the inter-quartile range or the mean (or median) absolute deviation from the mean, which also go up when the dispersion of the values increases. The issue is: what is the meaning of the SD in the context of the particular data at hand? The SD has relevance only when the data are symmetrically distributed, and has specific quantifiable and interpretable meaning for normally distributed data. Ecological data are typically not symmetrically distributed, so this rules out any single figure summary of dispersion about the mean. A lesson can be learnt from the reports provided by a

medical laboratory, where the patient's values on cholesterol, sugar, etc., are reported, alongside a reference range, consisting of two values, a lower and upper one, computed on data from a healthy population – this reference range would never be reported by a mean and an SD! As for confidence intervals, I have also made a strong case for reporting bootstrap estimates. While bootstrapping has some drawbacks, especially for small samples, it is nevertheless a good all-purpose distribution-free alternative that can be used for continuous data in general and an improvement over much of common practice at present.

I have made several specific recommendations in this article to improve data presentation in ecology – these are my particular and personal suggestions (out of many more potential ones) inspired by many years of working with ecologists and considering the statistical aspects of this special Kongsfjorden issue. I do not expect all researchers to follow my advice, since it is difficult to change established ways, such as reporting large tables of means and standard deviations and showing bar plots of means with a single "whisker" showing a standard deviation or standard error. Yet I hope that this article serves, at least, to alert readers to the problems surrounding the reporting and graphing of numerical data, to show them alternative ways of proceeding and to lead to an eventual improvement of statistical reporting in the ecological literature.

To assist readers to improve their data presentation skills, I have set up a webpage at the address [www.econ.upf.edu/~michael/graphicsR](http://www.econ.upf.edu/~michael/graphicsR), with commented R scripts that reproduce most of the graphics and statistics in this article.

## **Acknowledgments**

The author would like to express his sincere thanks to Haakon Hop for his encouragement and constant constructive feedback related to this article, also to Markus Molis for our many discussions on this topic, as well as Walter Zucchini for additional comments. Thanks are also due to Kim Huenerlage, Andrey Voronkov and Henk Bolhuis for their co-operation in being able to use some of their data from this special Kongsfjorden issue of Polar Biology.

## References

- Adams DC, Gurevitch J, Rosenberg MJ (1997) Resampling tests for meta-analysis of ecological data. *Ecology* 78:1277–1283
- Aitchison J, Brown JAC (1957) *The Lognormal Distribution*. Cambridge University Press, Cambridge UK
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 57:289–300
- Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge UK
- Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM (2014) Log-transformation and its implications for data analysis. *Shanghai Arch Psych* 26: 105–109
- Gierlinski M (2015) *Understanding Statistical Error: A Primer for Biologists*. Wiley, New York
- Good P (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Third edition. Springer, New York
- Greenacre M (2012) Contribution biplots. *J Comput Graph Stat* 22:107–122
- Greenacre M, Hastie T (2010) Dynamic visualization of statistical learning in the context of high-dimensional textual data. *J Web Semant* 8:163–168
- Huenerlage K, Graeve M, Buchholz F (2016) Lipid composition and trophic relationships of krill species in a high Arctic fjord. *Polar Biol*, doi:10.1007/s00300-014-1607-6

- Krzywinski M, Altman N (2013) Error bars. *Nat Methods* 10:921–922
- Land CE (1972) An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* 14:145–158
- Limpert E, Stahel WE, Abbt M (2001) The log-normal distribution across the sciences. *BioScience* 51:341–352
- Millard SP (2014) *EnvStats: An R Package for Environmental Statistics*. Springer, New York.
- Moran MD (2003) Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos* 100:403–405
- O'Hara R, Kotze J (2010) Do not log-transform count data. *Methods Ecol Evol* 1:118–122
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015) *vegan: Community Ecology Package*. URL <http://CRAN.R-project.org/package=vegan>
- Parkin TB, Chester ST, Robinson JA (1990) Calculating confidence intervals for the mean of a lognormally distributed variable. *Soil Sci Soc Am J* 54:321–326
- Piquet A, Maat DS, Confurius-Guns V, Sintès E, Herndl GJ, van de Poll W, Wiencke C, Buma AGJ, Bolhuis H (2016) Springtime dynamics, productivity and activity of prokaryotes in two Arctic fjords. *Polar Biol*
- Quinn, GP, Keough MJ (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK

StataCorp (2015) Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Tian L (2005) Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Stat Med* 24:3223–3232

Tufte ER (2001) *The Visual Display of Quantitative Information*. 2nd Edition. Graphics Press, Cheshire, Connecticut.

Tukey JW (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Voronkov A, Hop H, Gulliksen B (2016) Zoobenthic communities on hard-bottom habitats in Kongsfjorden, Svalbard. *Polar Biol*

Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond bar and line graphs: Time for a new data presentation. *PLoS Biol* 13(4): e1002128. doi: 10.1371/journal/pbio.1002128

Wickham H, Stryjewski L (2011) 40 years of boxplots. URL: <http://vita.had.co.nz/papers/boxplots.pdf> (last accessed 1 July 2016)

Table 1: Extract from Piquet et al. (2016, this volume, Table 1), showing means and standard deviations of two variables for various subsamples of the data set.

Station	Temperature	NO <sub>x</sub>
Early spring		
O	1.35 ( $\pm 0.38$ )	8.25 ( $\pm 2.48$ )
M	1.14 ( $\pm 0.26$ )	7.10 ( $\pm 2.34$ )
G	0.05 ( $\pm 0.57$ )	7.33 ( $\pm 1.57$ )
KM	0.10 ( $\pm 0.40$ )	8.71 ( $\pm 2.06$ )
KG	-0.30 ( $\pm 0.64$ )	9.17 ( $\pm 1.76$ )
Late spring		
O	2.91 ( $\pm 1.38$ )	0.62 ( $\pm 0.68$ )
M	2.74 ( $\pm 1.49$ )	1.08 ( $\pm 2.19$ )
G	3.07 ( $\pm 1.60$ )	1.23 ( $\pm 0.74$ )
KM	2.89 ( $\pm 1.26$ )	0.36 ( $\pm 0.36$ )
KG	3.14 ( $\pm 0.86$ )	0.80 ( $\pm 1.10$ )



Table 2: Significance of pairwise species comparisons (critical level of 0.05), using classical t-test as well as distribution-free exact permutation test, and whether the confidence intervals are non-overlapping.

Comparison	p-values		non-overlap	
	t-test	exact	in 95% CI	in 68% CI
$\delta^{13}\text{C}$				
<i>Ti</i> vs. <i>Tr</i>	0.696	0.623	no	no
<i>Ti</i> vs. <i>TL</i>	<b><u>0.020</u></b>	<b><u>&lt;0.001</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Ti</i> vs. <i>Mn</i>	<b><u>&lt;0.001</u></b>	<b><u>0.001</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Ti</i> vs. <i>Nm</i>	0.098	0.077	no	<b><u>yes</u></b>
<i>Tr</i> vs. <i>TL</i>	<b><u>0.020</u></b>	<b><u>0.026</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Tr</i> vs. <i>Mn</i>	0.053	0.102	no	<b><u>yes</u></b>
<i>Tr</i> vs. <i>Nm</i>	0.381	0.453	no	no
<i>TL</i> vs. <i>Mn</i>	<b><u>0.002</u></b>	<b><u>0.002</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>TL</i> vs. <i>Nm</i>	<b><u>0.004</u></b>	<b><u>0.006</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Mn</i> vs. <i>Nm</i>	0.163	0.191	no	<b><u>yes</u></b>
$\delta^{13}\text{N}$				
<i>Ti</i> vs. <i>Tr</i>	0.064	0.036	no	<b><u>yes</u></b>
<i>Ti</i> vs. <i>TL</i>	<b><u>0.004</u></b>	<b><u>&lt;0.001</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Ti</i> vs. <i>Mn</i>	<b><u>0.004</u></b>	0.065	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Ti</i> vs. <i>Nm</i>	<b><u>0.018</u></b>	0.091	no	<b><u>yes</u></b>
<i>Tr</i> vs. <i>TL</i>	0.078	0.084	no	<b><u>yes</u></b>
<i>Tr</i> vs. <i>Mn</i>	0.870	0.914	no	no
<i>Tr</i> vs. <i>Nm</i>	<b><u>0.004</u></b>	<b><u>0.012</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>TL</i> vs. <i>Mn</i>	<b><u>0.041</u></b>	0.056	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>TL</i> vs. <i>Nm</i>	<b><u>&lt;0.001</u></b>	<b><u>0.004</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>
<i>Mn</i> vs. <i>Nm</i>	<b><u>0.004</u></b>	<b><u>0.008</u></b>	<b><u>yes</u></b>	<b><u>yes</u></b>

Table 3: p-values from MANOVAs performed on the pairs of species, with significant p-values ( $p < 0.05$ ) shown in bold and underlined.

	<i>Tr</i>	<i>TL</i>	<i>Mn</i>	<i>Nm</i>
<i>Ti</i>	0.072	<b><u>0.0001</u></b>	<b><u>0.001</u></b>	0.099
<i>Tr</i>		<b><u>0.001</u></b>	<b><u>0.029</u></b>	<b><u>0.0009</u></b>
<i>TL</i>			<b><u>0.002</u></b>	<b><u>0.003</u></b>
<i>Mn</i>				<b><u>0.0006</u></b>

Table 4: p-values from MANOVAs performed on the pairs of group means in Fig. 12, with significant p-values ( $p < 0.05$ ) shown in bold and underlined.

	A2	A3	A4	A5	A6
A1	0.235	<b><u>0.003</u></b>	<b><u>&lt;0.001</u></b>	<b><u>&lt;0.001</u></b>	<b><u>&lt;0.001</u></b>
A2		<b><u>0.164</u></b>	<b><u>0.003</u></b>	<b><u>&lt;0.001</u></b>	<b><u>&lt;0.001</u></b>
A3			0.243	<b><u>&lt;0.001</u></b>	<b><u>&lt;0.001</u></b>
A4				<b><u>&lt;0.001</u></b>	<b><u>&lt;0.001</u></b>
A5					<b><u>&lt;0.001</u></b>

## Figure captions

Fig. 1 (a) Sample of 15 egg counts in an experiment on *Calanus glacialis* (b) Horizontal and vertical stem-and-leaf plots of this small data set (c) Dot-plot representation of the data

Fig. 2 Different representations of the count data: (a-f) attempt to show the distribution of individual values; (g-k) attempt to show the distribution of the mean; from (c) onwards, with the exception of (f), a box is drawn from the zero to the mean value. (a) the actual raw data as a dot plot; (b) a boxplot showing median, quartiles and extreme values; (c) and (d) one standard deviation (SD) above the mean, first as a whisker above the mean, then as a whisker above and below the mean; (e) an attempt to reflect the asymmetry of the distribution, showing the SD computed for values below the mean and a separate computation for values above the mean; (f) the mean and whiskers extending to the estimated 2.5 % and 97.5 % quantiles (so that 95 % of the estimated distribution lies between these limits; (g) and (h) one standard error (SE) above the mean and then above and below; (i) and (j) the margin of error (ME) for 95 % confidence level above the mean and then the same margin of error above and below the mean, i.e. the 95 % confidence interval according to normal theory; (k) the 95 % confidence interval based on distribution-free bootstrap sampling. Two ticks (✓✓) indicate acceptable ways of visualizing the variability of the data or the mean, one tick (✓) is borderline and a cross (×) is unacceptable. Horizontal bars drawn at the end of the whiskers enclose intervals that have clear substantive meaning

Fig. 3 (a) Sample of size 2 and mean  $\pm$  SD plot. (b) Three samples of size 3 in completely different configurations, which give exactly the same mean  $\pm$  SD plots, and

which are identical in turn to the plot in (a). The horizontal dashed line indicates the mean common to all the configurations

Fig. 4 95 % confidence plot for mean egg count, based on 10000 bootstrap samples. This is the same as Fig. 2i, but visualized in a different style reminiscent of a boxplot. The mean is indicated by a dot (as opposed to a line for the median in a boxplot). The box encloses a 50 % confidence interval for the mean, and the whiskers a 95 % confidence interval

Fig. 5 (a) Stem-and-leaf plots of two sets of egg-count data, for control and treatment groups, as well as a back-to-back plot (b) 95 % confidence plots for the means (C=control, T=Treatment)

Fig. 6 NO<sub>x</sub> data ( $n = 10$ ), corresponding dot plot display (with points jittered where there are overlaps), and boxplot.

Fig. 7 Normal quantile plots for the original NO<sub>x</sub> values and their log-transformed ones.

Fig. 8 Separate confidence intervals for variables  $\delta^{13}\text{C}$  and  $\delta^{13}\text{N}$  at two levels of confidence: (a) 95 % and (b) 68 % (one standard error), obtained by bootstrapping

Fig. 9 Elliptical 95 % confidence regions for the bivariate means of the five species, based on bootstrapping

Fig. 10 Perspective views of a three-variable example of the ellipsoidal 95 % confidence regions of six group means, from two different angles. In (a) the ellipsoids of groups A2 and A4 appear to be overlapping, whereas in the different view of (b), with variable 3 shown vertically, they are seen to be separated

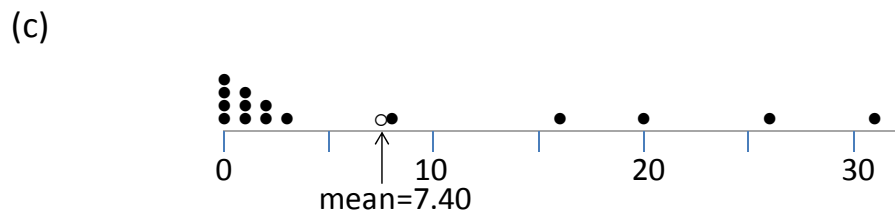
Fig. 11 Two-dimensional CCA ordination of the data of Voronkov et al. (2016, this volume), showing (i) the samples, in seven numbered groups (there is only one sample from group 2), (ii) the 95% confidence ellipses for the group means (except for group 2), shaded with the group number in boldface, (iii) the 13 species that contribute more than the average to one of the two dimensions, (iv) the three additional species (in smaller font, lower left), that contribute less than average but nevertheless are more than average highly correlated with the ordination, and (v) the seven constraining variables (in grey). For abbreviations see Voronkov et al. (2016, this volume)

Fig. 1

(a) 1 0 1 2 2 8 16 31 0 0 20 0 26 1 3

(b)

0	00001112238	8
1	6	3
2	06	2
3	1	2
		1
		1
		1
		0
		0
		0
		0
		6
		0 6 0 1
		<u>0 1 2 3</u>



(a) 1 0 1 2 2 8 16 31 0 0 20 0 26 1 3

(b)

0	00001112238	8
1	6	3
2	06	2
3	1	2
		1
		1
		1
		0
		0
		0
		0
		6
		0 6 0 1
		<u>0 1 2 3</u>

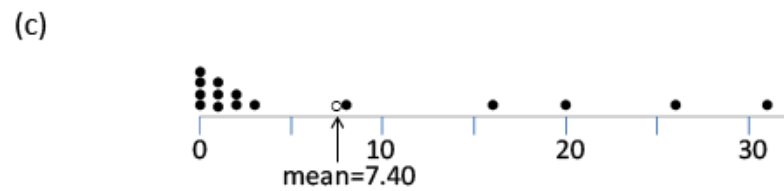


Fig. 2 (should be rotated)

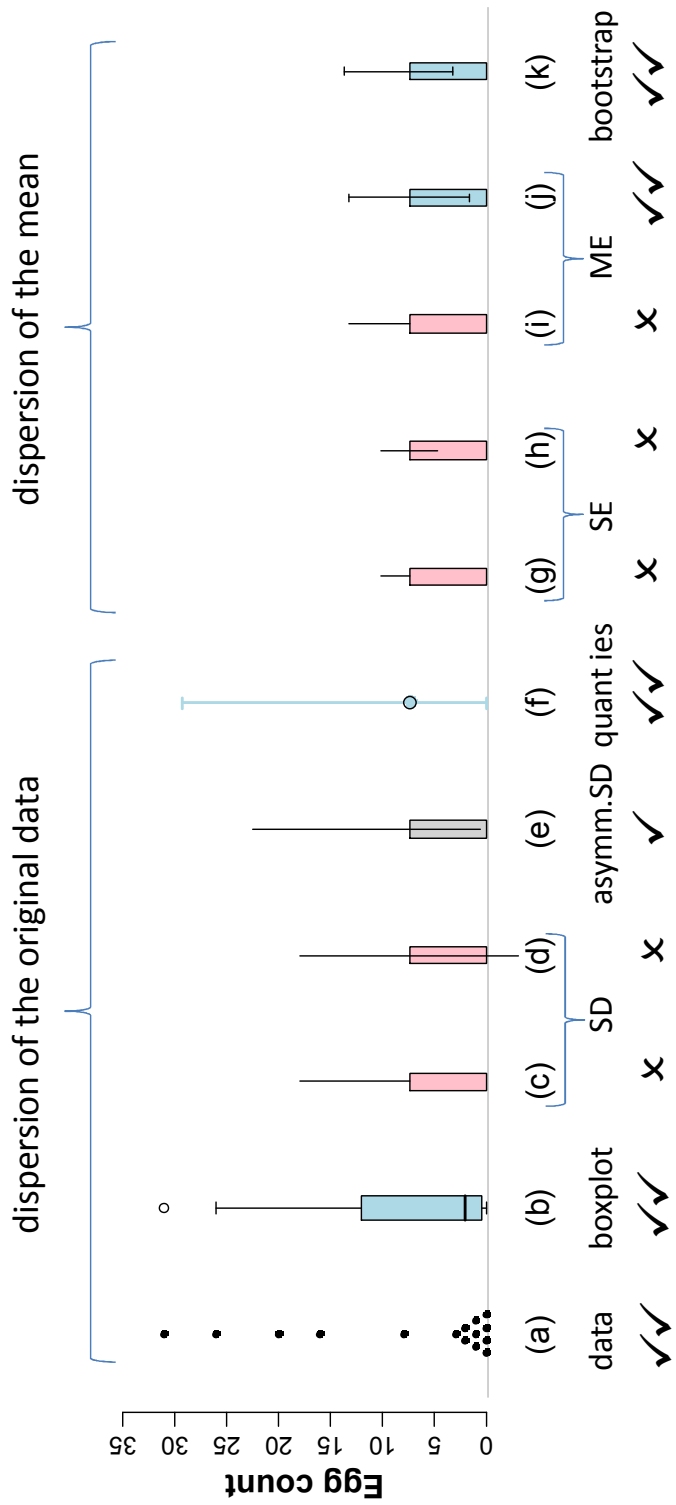




Fig. 3

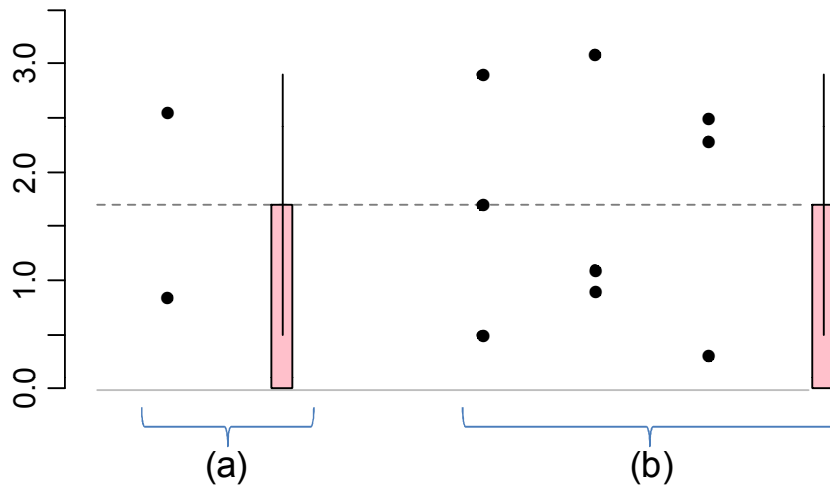


Fig. 4

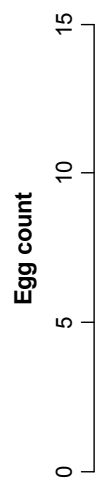


Fig. 5

Control data  
0|00001112238  
1|6  
2|06  
3|1

Treatment data  
0|0011388  
1|28  
2|16  
3|05  
4|8  
5|0

Control Treatment  
83221110000|0|0011388  
6|1|28  
60|2|16  
1|3|05  
|4|8  
|5|0

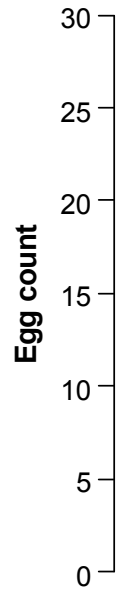


Fig. 6

0.27	1.16	3.71	0.18	1.13
0.59	0.27	0.14	0.47	0.04

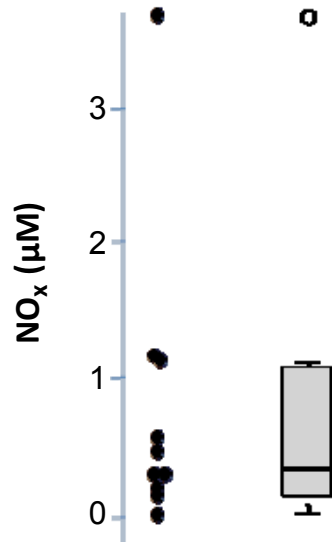


Fig. 7

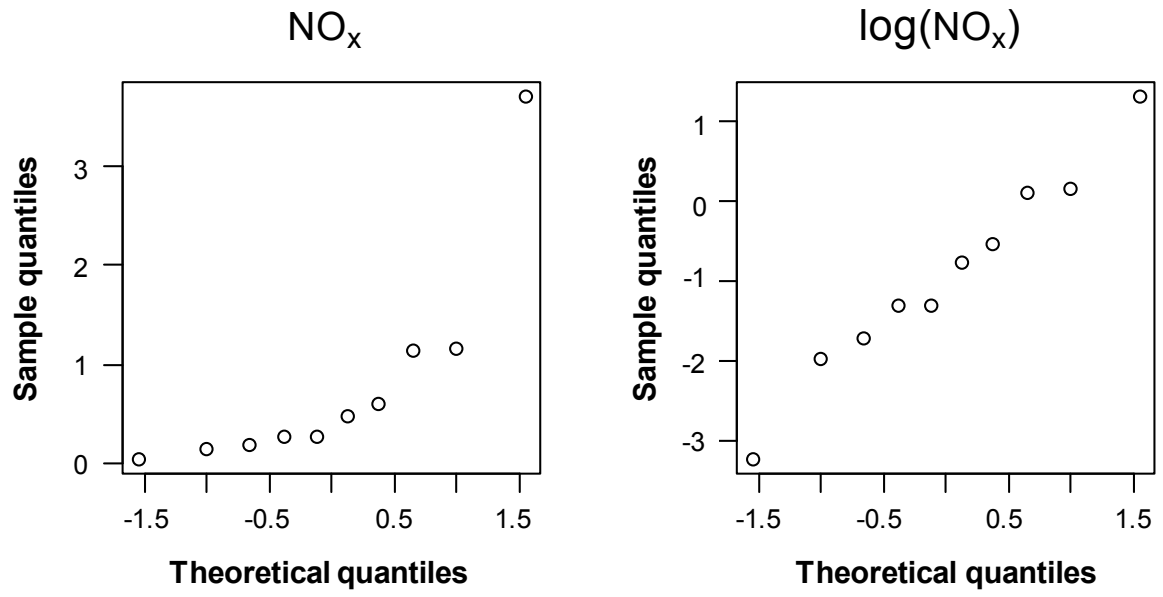


Fig. 8

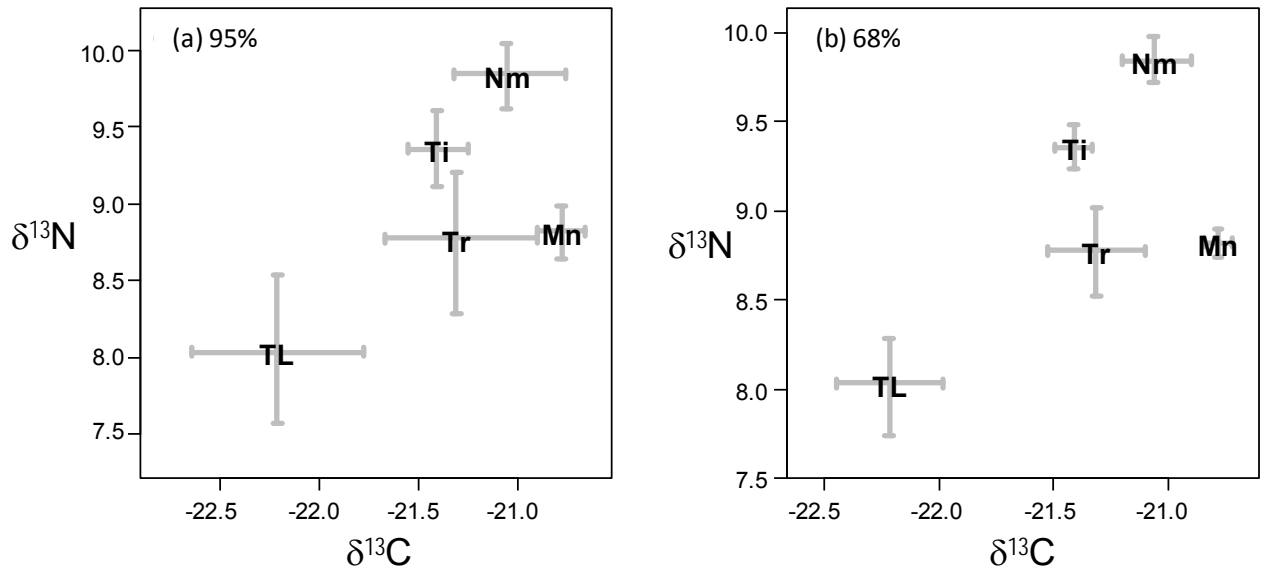


Fig. 9

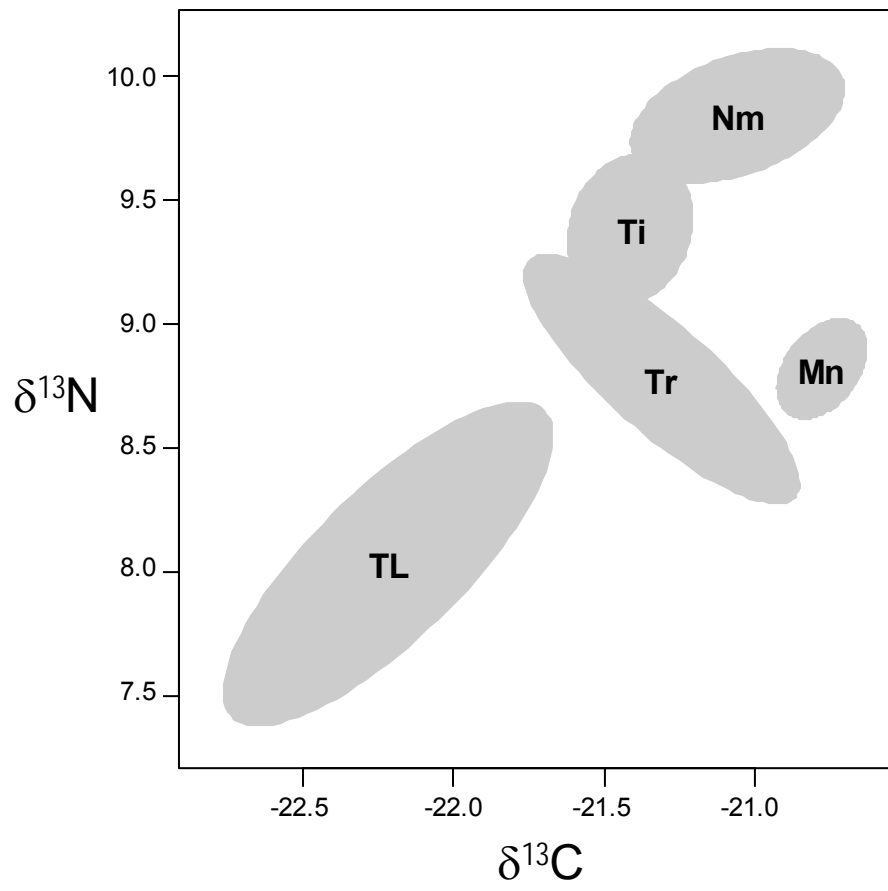


Fig. 10

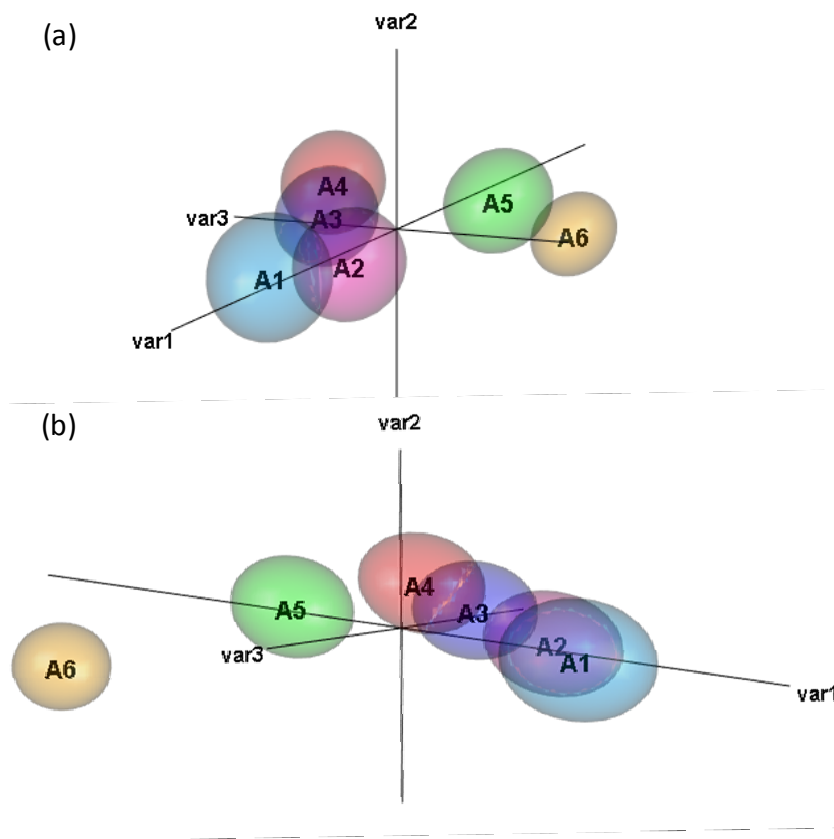




Fig. 11

