Statistics in soil science☆

R Webster, Rothamsted Research, Harpenden, United Kingdom

Key points

- Soil varies from place to place; statistics is a technology for describing that variation quantitatively and separating signal from noise.
- Replicate sampling is required to estimate the variation in experimental material and geographic regions.
- Randomization is needed to estimate mean values and their variances, and to place confidence limits on means.
- Data can be summarized by means, medians, variances, standard deviations and skewness coefficients.
- Distributions of data are best illustrated by histograms and Q–Q plots.
- Analysis of variance apportions the variance in a set of data to sources in the designs of experiments and surveys; any actual analysis must fit the design by which the data have been obtained.
- Relations among two or more measured variables are expressed by Pearson correlation coefficients.
- A variable may be predicted from one or more others on which it depends by regression.

Introduction—Why statistics?

Soil varies from place to place at all scales from the global to the infinitessimal. Much of the variation is natural, arising from variation in its parent material and processes such as erosion and deposition on the land surface. Soil also varies from time to time as it responds to weather, plant growth, and processes in the rhizosphere engendered by that growth. Farmers have added to the variation by their enclosing, reclaiming, clearing, and fertilizing the land, though within fields they have removed some by cultivation and drainage. Further sources of variation are mineral extraction and subsequent reclamation, dumping of waste, and pollution of many kinds. Data on soil embody variation from those many sources, and the aim of statistics at its simplest is to express quantitatively that variation, to separate contributions from different sources and to describe relations.

Investigators design experiments and surveys in such a way that they can estimate the variation from particular sources such as imposed treatments or strata in a population and the differences between them. Variation in data also arises from the way observations are made; from the people who make the observations, from the imprecision of instruments, from imperfect

☆*Change History:* July 2022. R Webster updated the chapter.

laboratory technique, and from sampling fluctuation. One may like to think that the contributions from the laboratory are negligible, though `ring' tests have often revealed them not to be so. In general, however, sampling fluctuation in the field, arising from the spatial variation there, is much the larger.

Any one measurement of a soil property is influenced by contributions from at least some of these sources. It cannot be taken as an absolutely accurate representation of the truth therefore; rather it must be seen in relation to the likely error.

Statistics are needed in soil science to estimate and express this error and to apportion it to the different sources. In this way sense (signal) can be separated from meaningless or uninteresting variation (noise) in comparisons between classes of soil, in expressing relations, in assessing the effects of treatments in experiments, and in prediction. The statistical repertoire is huge, and this article describes the basics and the fairly elementary techniques that soil scientists most often need.

The techniques can be divided into two groups, namely, description and analysis. They also have two fairly distinct fields of application: survey and experiment. In the first investigators observe and record the soil as it is on samples, and descriptive techniques tend to dominate. In the second they deliberately control some of the variation so that they can assess the effects of changes in one or a few factors that are of specific interest by analysis.

There are numerous text books on the statistics summarized in this chapter, and many more tutorial articles in scientific journals and on the World-Wide Web. I recommend strongly three: Snedecor and Cochran (1989), Sokal and Rohlf (2012) and Welham et al. (2015).

Population, units and samples

The soil is regarded for statistical purposes as a *population* comprising elements or *units*. The units are the bodies of soil on which measurements are made. They are more or less arbitrary and determined largely by convenience and practicality. They may be individual cores of soil, pits or pedons, each may be the volume of soil occupied by the roots of a single plant or that deformed under the wheel of a tractor, they may be pots in a greenhouse experiment, lysimeters, plots in a field experiment, or whole fields or farms. The variation among them depends very much on the size of the units; the larger they are the more variation they encompass within them and the less there is between them to be revealed in data. The size, shape, and orientation of the units, known as the "support" in survey, must be defined and adhered to throughout an investigation. The population is then all such units in a specified region or falling within some other definition for the purposes of the investigation. It is an operational definition, often known as the "target population." In a more restricted sense the population and the units comprising it may be the values of a particular soil property in the defined supports.

Populations in surveys are typically very large, in many instances infinite, or hypothetical, and in some they are poorly defined. Measurement is feasible only on small subsets, i.e., on samples, and these subsets must properly represent their populations for the measurements to apply to the larger populations. The units in an experiment, in contrast, are typically a few dozen, though ideally the experimental material should be representative of some larger population.

Replication and randomization

Estimating the mean in a population and its associated error from measurements on a sample requires both replication and randomization. The need for replication is evident; a single value can contain no information on variation. Randomization is needed to avoid bias. At its simplest it means choosing units such that all have the same chance of being selected.

In survey selection can take the form of simple random sampling. To apply it requires (1) that each unit can be identified uniquely, and (2) a rule for the selection. Simple random sampling, as in Fig. 1A, tends to be inefficient in that it takes no account of anything known about the population beforehand, such as its spatial dependence, a soil map of the region, the relations between soil and vegetation or physiography, or the farming history. To take advantage of such knowledge the population is first stratified either into small grid cells, as in Fig. 1B, or by type of soil, vegetation, physiography, or farming as above, as in Fig. 1C. The soil is then sampled at random within each stratum independently; this is stratified random sampling, and it enables the variation due to the strata to be distinguished from variation within them.

Brus (2021) has drawn attention to a growing notion that random sampling is to be avoided where there is spatial correlation. As Brus points out, that notion is false. See also de Gruijter et al. (2006). Random sampling guarantees that estimates of means and variances are unbiased, regardless of any spatial correlation, and estimating them needs to be distinguished from mapping for which other sampling designs are more efficient. Do not be misled by publications that state otherwise; stick to the authoritative texts mentioned above and to Cochran's (1977) standard text on sampling.

In experiments treatments are allocated to the units at random. The units may be arranged completely randomly, as in Fig. 2A. In the field and greenhouse they are usually grouped into blocks such that each block contains one unit of each treatment, as in Fig. 2B; long-range variation then appears in the variation among blocks. There are many more elaborate designs.

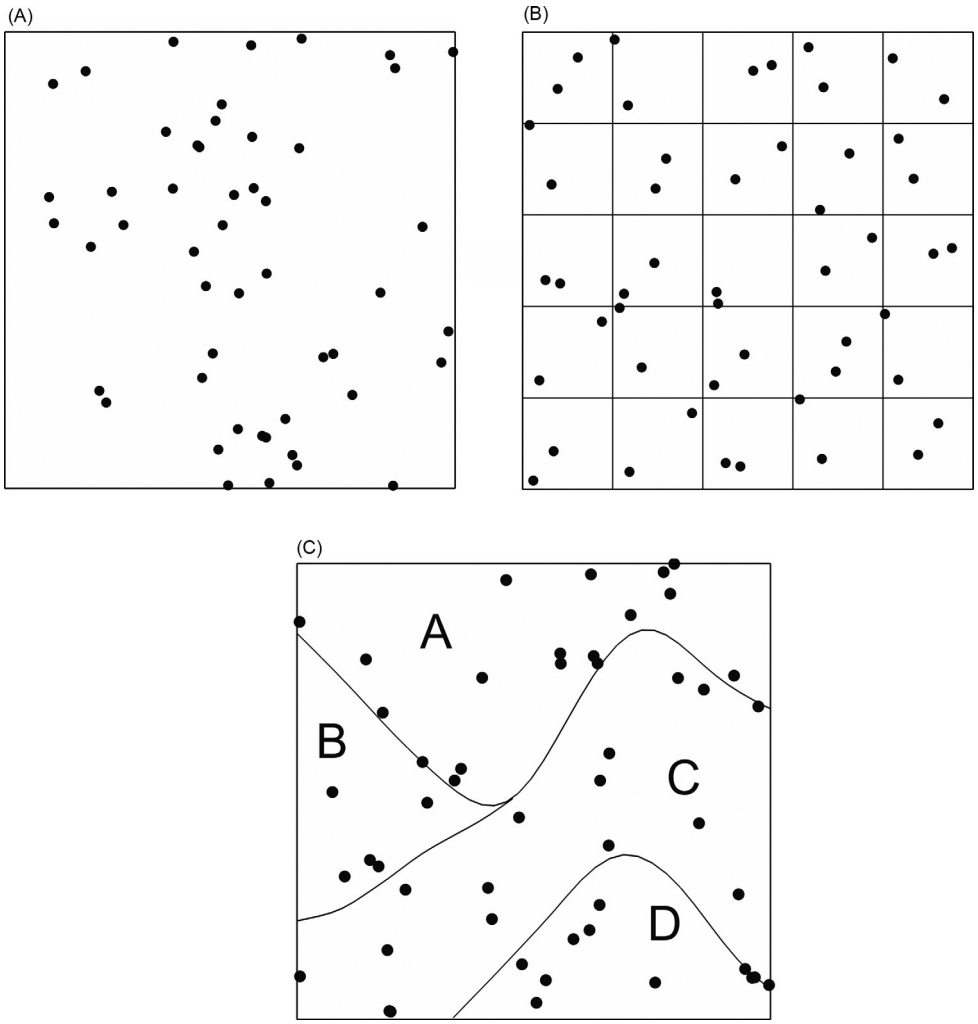

Fig. 1 Probabilistic sampling of a region; (A) simple random sampling; (B) stratified random sampling with the region divided into 25 square cells (strata) and two points per cell; (C) stratified random sampling with four mapped classes of soil as strata.

(A)

W	0	F	W	D
F	D	W	F	0
W	0	W	0	D
D	F	F	D	0

(B)

W	D	F	W	F
F	0	D	F	0
0	W	W	0	D
D	F	0	D	W

Fig. 2 Layout of a randomized field experiment with four treatments, 0 (control), D (dung), W (industrial waste), and F (NPK fertilizer), and fivefold replication; (A) completely randomized; (B) with the replicates arranged in five blocks.

Descriptive statistics

The mean

In almost all investigations mean values are of prime interest. Provided sampling has been properly randomized the mean of the data, denoted $z_1, z_2, \ldots, z_N$,

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{1}$$

estimates without bias the mean, μ, in the population from which the sample was drawn. How well it does so depends on the variation it embodies.

Characteristics of variation

Histogram

The variation in a set of measurements, if there are sufficient of them, can be displayed in a histogram. The scale of measurement is divided into segments of equal width, or "bins," the values falling in each bin are counted, and bars of height proportional to the counts are drawn. Fig. 3 is an example; it summarizes graphically the way in which the frequency is distributed over the range of the data.

Variance

Variation is best expressed quantitatively as variance. For a set of N data it is the average squared difference between the observations and their mean:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N} (z_i - \bar{z})^2 \tag{2}$$

Its square root, S, is the standard deviation, which is often preferred because it is in the same units as the measurements and is more intelligible therefore.

Ideally this variance should estimate the variance of the larger population, of which the N observations are a sample. This variance of a population is by definition

$$\sigma^2 = \mathrm{E}\left[(z - \mu)^2\right] \tag{3}$$

where μ is mean of z in the population and E denotes expectation. Eq. (2) gives a biased result, however; S^2 is a biased estimator. The reason is that $\bar{z}$ in the equation is itself more or less in error as the estimate of μ. To remove the bias N must be replaced by $N - 1$ in the denominator:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (z_i - \bar{z})^2 \tag{4}$$

The result, s^2, is now unbiased, provided the sampling was unbiased in the first place.

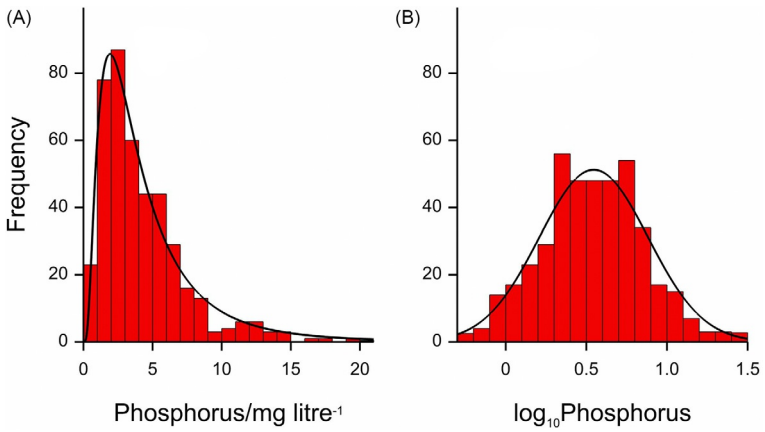

Fig. 3 Histograms of 434 data on available phosphorus (A) in mg L^{-1}, and (B) transformed to common logarithms. The curve in (B) is that of the fitted normal distribution, and that in (A) shows the lognormal distribution.

Estimation variance, standard error and confidence

Both S and s measure the dispersion in the observations; neither expresses the reliability of the estimate of μ. The reliance one can place in an estimate of the mean can be expressed in terms of the expected squared deviation of it from the true mean, i.e., $\mathrm{E}\left[(\bar{z}-\mu)^2\right]$. Provided that the sampling points are drawn independently of one another, as for example in simple random sampling, its estimate is derived from s^2 by

$$s^2(\bar{z}) = s^2/N \tag{5}$$

This is the estimation variance, and its square root is the standard error, which is the standard deviation of means of samples of size N. The larger is the sample the smaller is this error, other things' being equal, and the more confident we can be in the estimate. So, distinguish between the standard deviation, which describes the variation in a sample, and the standard error, which expresses the confidence associated with a mean. Standard errors should accompany means in tables compiled from replicated measurements, and they may be shown by error bars on graphs.

Eq. (5) gives the estimation variance for a simple random sample of size N. If the population has been divided into strata then s^2, the population variance on the right hand side of the equation, can be replaced by ${s_\mathrm{W}}^2$, the variance within strata. The latter is generally less, often much less, than s^2, and so stratified sampling is more precise than simple random sampling by the factor $s^2_{\mathrm{random}}(\bar{z})/s^2_{\mathrm{stratified}}(\bar{z})$. It also means that one can achieve the same precision with a smaller sample, and in this sense stratified sampling is more efficient. This efficiency can be expressed as $N_{\mathrm{random}}/N_{\mathrm{stratified}}$. The within-stratum variance, ${s_\mathrm{W}}^2$, can be estimated by the analysis of variance, see below.

The measures of error above represent uncertainty, which in the current context can be translated into confidence, that is the confidence we can have that a true mean, μ, deviates no more than some value from our estimate. Limits of confidence of our choosing can be obtained from the standard errors for known kinds of distribution. If the data are drawn from a normal (Gaussian) distribution and there are many of them then an interval of $2s/\sqrt{N}$ centered on $\bar{z}$ spans a symmetric confidence interval of approximately 68%. We can readily calculate wider confidence intervals by multiplying by factors, η, such as

Confidence/%	80	90	95	99
η	1.28	1.64	1.96	2.58

The factor η is the value of the standard normal deviate for the desired confidence. Evidently, the more certain, or more confident, we want to be that μ lies within some limits the wider that confidence interval must be.

When data are few ($N < 30$) the above factors need to be replaced by Student's t for the number of degrees of freedom, f. The lower and upper symmetric confidence limits about a mean $\bar{z}$ are

$$\bar{z} - t_f s/\sqrt{N} \text{ and } \bar{z} + t_f s/\sqrt{N} \tag{6}$$

Values of t are available in tables and in most statistical computer packages. For $30 < N \leq 60$ t converges to η, and for $N > 60$ the values in the above table can be used.

There are formulae for calculating the confidence limits for other theoretical distributions. In many instances, however, it is easier to transform data to approximate a normal distribution and subsequently analyze the transformed values.

Coefficient of variation

The coefficient of variation (CV) is the standard deviation divided by the mean, i.e., $s/\bar{z}$. It is often multiplied by 100 and quoted as a percentage. Its merit is that it expresses variation as pure numbers independent of the scales of measurement. It enables investigators and those reading their reports to appreciate quickly the degree of variation present and to compare one region with another and one experiment with another. It should not be used to compare variation in different variables, especially ones having different dimensions.

The coefficient is sensible only for variables measured on scales with an absolute zero. Otherwise the arbitrary choice of the zero affects it. Examples for which it should not be used are soil temperature in degrees Celsius (arbitrary zero at 273 K), color hue (which is approximately circular), and soil acidity on the pH scale (arbitrary zero equivalent to $-\log_{10}[\mathrm{H}^+]$ with H^+ expressed in mol L^{-1}).

For some soil properties physics sets limits on the utility of the CV. For example, the minimum bulk density of the soil is determined by the physical structures that keep particles apart. Particles must touch one another, otherwise the soil collapses. For mineral soils on dry land a working minimum bulk density is around 1 g cm^{-3}. At the other end of the scale the bulk density cannot exceed the average density of the mineral particles, approximately 2.7 g cm^{-3}. So the CV of bulk density is fairly tightly constrained.

The sensible use of the coefficient of variation for comparing two variables y and z relies on the assumption that they are the same apart from some multiplying factor, b, thus:

$$y = bz \tag{7}$$

Then the mean of y is $\bar{y} = b\bar{z}$, its variance ${s_y}^2 = b^2{s_z}^2$, and its standard deviation is $s_y = bs_z$. From there simple arithmetic shows that their CVs are the same. This principle offers a means of comparing variation with logarithms of the observations. Eq. (7) becomes

$$\log y = \log b + \log z \tag{8}$$

The logarithm log$b$ is a constant, and so the variances, $s_{\log y}{}^2$ and $s_{\log z}{}^2$, are equal, as are their standard deviations. The resulting measure of variation is therefore independent of the original scale of measurement.

The measure can be used to compare variation in two groups of observations. Consider again soil acidity. To compare the variation in acidity of a class A with that in another, class B, we treat the hydrogen ion concentration as the original variable, transform it to pH, and compute the variances of pH. Whichever has the larger variance is the more variable, regardless of the mean. Further, we can make a formal significance test by computing $F = s_{\log y}{}^2/s_{\log z}{}^2$ and compare the result with F for $N_y - 1$ and $N_z - 1$ degrees of freedom, see below.

Additivity of variances

Variances are additive; those from two or more independent sources in an investigation sum to the total in the data. Their square roots, the corresponding standard deviations, are not. To obtain an average variation on the original scale of measurement from several sets of data compute the arithmetic mean of their variances, weighted as appropriate by the numbers of degrees of freedom, and then take the square root of it to give an "average" or pooled standard deviation. This divided by the mean will give an "average" CV. More generally, the additive nature of variances confers great flexibility in analysis, enabling investigators to distinguish variation from two or more sources and estimate their components according to the design as by the analysis of variance.

Statistical significance

Significance in a statistical context means distinguishing a signal or the effects of some imposed treatment or detecting differences between strata against a background of "noise." It is matter of separating the variance due to the signal, treatments, or strata from that from other sources that are of no interest. The question being addressed is as follows; given the magnitudes of the several components of variance, is the signal so strong or are the differences observed so large that they are highly unlikely to have occurred by chance. If the answer is "yes," then the result is said to be significant.

A significance test is prefaced by a hypothesis. This is usually that there is no real difference between populations or treatments and that any differences among the means of observations are due to sampling fluctuation. That is the "null hypothesis," often designated H_0, and the test is designed to reject it; *not to confirm it*. The alternative that there is a difference is denoted either H_1 or H_A.

To judge, for example, whether two means differ one computes from the sampling error the probability, P, of obtaining the observed difference if the true means were identical, assuming that one knows the form of the distribution. If P is small (conventionally $<0{:}05$) the null hypothesis is rejected, and the difference is judged significant. If P is large then the null hypothesis is likely to be correct, but we have no measure of the probability that the two means are indeed identical; instead we take the view that we have too little evidence to conclude that the difference observed applies to the population from which the sample was drawn.

One can still draw mistaken conclusions as the result of significance tests. Mistakes can be of two kinds, denoted Type I and Type II. The first occurs when we reject the null hypothesis on the basis of our sample evidence, i.e., declare a difference significant, when the populations do not differ. The second is when we accept the null hypothesis, i.e., state that we have insufficient evidence for a difference, when the populations do differ.

One can diminish the likelihood of drawing wrong conclusions by increasing the sensitivity of the test, and that depends on the precision with which the means have been estimated, i.e., on their estimation variances or on the estimation variance of their difference. The latter is given by

$$s^2_{\text{diff}} = \frac{s^2_{\text{W}}}{n_1} + \frac{s^2_{\text{W}}}{n_2} \tag{9}$$

where n_1 and n_2 are the numbers of observations from which the means in classes 1 and 2 derive, and $s_{\text{W}}{}^2$ is the variance within the classes, assumed to be common. If $n_1 = n_2$ then the variance of the difference is simply twice the estimation variance of the individual means.

Eq. (9) shows two features. One is that the larger is the variance within the populations the larger is the variance between the means and the less likely is one to establish a difference as significant. The second is that larger samples result in smaller variances and hence more sensitive comparisons. If the samples are large enough one can establish that any soil is different from almost any other for whatever property of interest.

The significance test is valuable in preventing false claims on inadequate evidence. Thus, a result might be summarized as follows.

> The mean measured pH of the topsoil was 5.7 compared with 6.7 in the subsoil; but because the samples were small [or because the variances were large] the difference was not statistically significant.

However, when a difference is deemed statistically significant because the null hypothesis is rejected that does not mean that it is important or physically or biologically meaningful. For example, the difference between an observed mean pH of 5.7 in the topsoil

and 5.9 in the subsoil would be of little consequence, whatever the probability of rejecting the null hypothesis. Also, while an investigator might regard a difference as significant if $P \leq 0.05$, a reader may be willing to recognize one if $P < 0.1$ or, more stringently, only if $P \leq 0.01$. The critical value of $P = 0.05$ for significance was originally suggested of R.A. Fisher; it was not a rule to be followed slavishly. Now that one can calculate probabilities so readily in computer packages professional statisticians, led by the American Statistical Association, recommend that we be guided by those values of P (Wasserstein and Lazar, 2016). If they are published instead of statements such as "$P \leq 0.05$" along with the standard errors then readers can reach their own judgments.

Note finally that the null hypothesis is highly implausible when horizons and different types of soil are being compared; they are different.

Transformations

It is often desirable to transform data to their square roots, or logarithms or by other more elaborate functions. One reason for doing so is to obtain a new variate that approximates some known distribution, preferably normal (Gaussian) so that the usual parametric tests of significance can be applied.

The most serious departure from normality usually encountered with soil data is skewness, i.e., asymmetry, as in Fig. 3A. The normal distribution is symmetric, its mean is at its center (its mode), and the mean of the data estimates this central value without ambiguity. The mean of data from a skewed distribution does not estimate the mode, nor does the median (the central value in the data). The meaning of the statistics can be uncertain therefore. A second feature of skewed data is that the variances of subsets depend on their means. If the data are positively skewed (again the usual situation) then the variances increase with increasing mean. This is undesirable for comparisons. Third, estimation is "inefficient" where data are skewed; that is the errors are greater than they need be or, put another way, more data are required to achieve a given precision than would be if the distribution were normal. Transforming data to approximate normality overcomes these disadvantages. We achieve symmetry, and hence remove ambiguity concerning the center. We stabilize the variances. And we make estimation efficient. The second of these is perhaps the most important.

No real data are exactly normal; all deviate more or less from normality. We have therefore to decide whether to transform them. This is best done by judicious exploration of the data aided by graphic display.

Draw a histogram. If it looks symmetrical then superimpose on it a normal curve computed from the mean and variance of the data. The normal curve has the formula

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} \tag{10}$$

where y is the probability density. Scale it to fit the histogram by multiplying by the number of observations and by the width of the bins. If the curve fits well then there is no need to transform the data.

Another popular way of examining data is a Q–Q plot. This is a graph of the cumulative distribution of the data plotted against the quantiles of a standard normal distribution. If the data are approximately normal then the points lie close to a straight line. Any appreciable curve signifies strong departure from normal. Fig. 4 shows the Q–Q plots for the same data as the histograms in Fig. 3.

If the histogram is skewed then compute the skewness coefficient in addition to the mean and variance. This dimensionless quantity can be obtained via the third moment of the data about their mean:

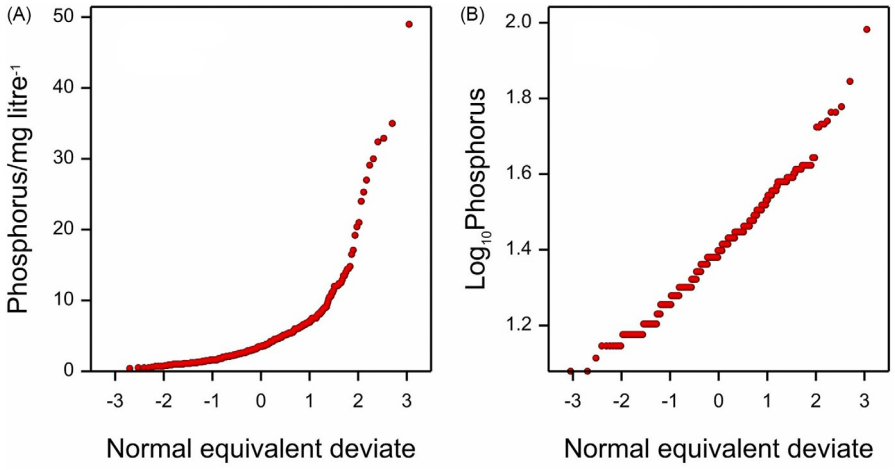

Fig. 4 Q–Q plots of 434 data on available phosphorus (A) in mg L^{-1}, and (B) transformed to common logarithms. The points in (A) lie close to a curve, signifying strong departure from normal, whereas those in (B) lie close to a straight line characteristic of a normal distribution.

$$\gamma_1 = \frac{1}{NS^3}\sum_{i=1}^{N}(z - \bar{z})^3 \tag{11}$$

A symmetric histogram has $\gamma_1 = 0$. Values of γ_1 greater than 0 signify positive skewness, i.e., long upper tails to the distribution and a mean that exceeds the median, which is common. Negative values of γ_1 signify negative skewness, and are unusual.

If γ_1 is positive and less than 0.5 then there should be no need to transform the data. If $0.5 < \gamma_1 \leq 1$ then it might be desirable to convert the data to their square roots; and if $\gamma_1 > 1$ then a transformation to logarithms is likely to give approximate normality.

The following example illustrates the situation. The data, which are summarized in Table 1, are 433 measurements of available phosphorus, P, in the topsoil. Their skewness coefficient is 3.95, i.e., they are strongly positively skewed, and this is apparent in their histogram, Fig. 3A. Transforming to logarithms makes the histogram, Fig. 3B more nearly symmetric, and as the skewness in the logarithms is now only 0.34; the transformation seems satisfactory. Further, the normal curve appears to fit well. Fig. 4, the Q–Q plots of the data, conveys the same message: the curve on which the points lie in Fig. 4A becomes a straight line for the transformed data in Fig. 4B.

There are statistical tests for normality, but they are unhelpful. If data are numerous then small departures from normality lead to supposed significance; if they are few the results are inconclusive.

Fig. 5 shows how the transformation stabilizes the variances. In Fig. 5A the variances of subsets of 44 from the full set of data on phosphorus are plotted against the means. Evidently, the variance increases strongly with increasing mean. Converting the data to their logarithms produces a result in which there is virtually no relation, Fig. 5B.

These simple functions change only the general form of the distribution; they do not change the detail. Normalizing the detail requires a more elaborate normal score transform.

Analysis of variance

The analysis of variance is at once one of the most powerful and elegant techniques in statistics. It and regression analysis are what may be regarded as the "bread-and-butter" of statistics in soil research. Its basis is that variances are additive and that the total variance in a population is the sum of the variances contributed by two or more sources. Working from the design of an

Table 1 Summary statistics of 433 values of available phosphorus measured in a survey of topsoil.

	Scale	
Statistic	*P/mg L^{-1}*	*Log$_{10}$P*
Mean	4.86	0.546
Variance	26.52	0.1142
Standard deviation	5.15	0.338
Skewness	3.95	0.23
X^2_{df18}	368.2	26.7

Values of X^2, with 18 degrees of freedom, are added for the hypothesis that the data or their logarithms are from a normal distribution.

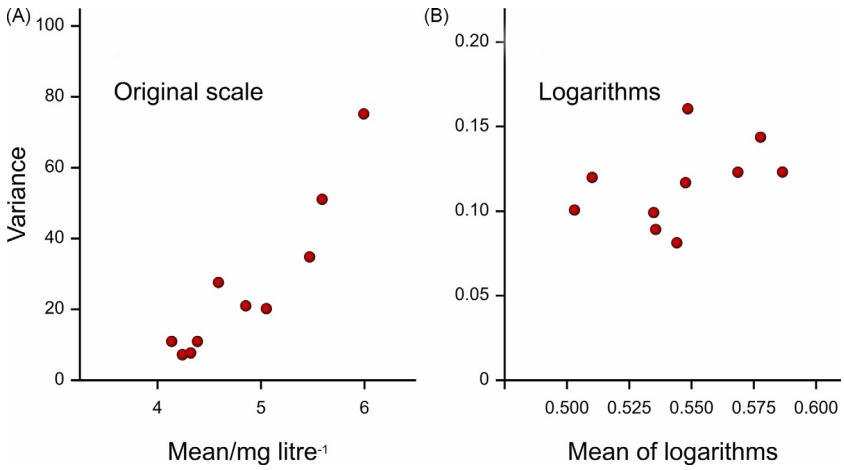

Fig. 5 Graphs of variance against mean for 10 sub-sets of 44 phosphorus data (left) on the original scale in mg L^{-1} and (right) after transformation to common logarithms.

investigation it analyzes the data by separating the contributions from those sources and estimating the variances in them. Designs vary from the simple to highly complex, but all embody the same principle.

Here we consider two of the simplest simple designs, such as appear in Fig. 2. Investigators want to know how manuring improves crop yield in the field. They have several (k) treatments, say, nothing (0), dung (D), industrial waste (W), and a complete artificial (NPK) fertilizer (F). They replicate each m times by assigning the treatments completely at random to plots of equal size. Fig. 2A shows how the experiment might be layed out with $m = 5$ replicates. They apply the treatments, grow the crop, and measure the yield at harvest, designated z.

The total variance in the yields in the experiment, s_T^2, comprises variance from two sources, namely that between the treatments, s_B^2, and that within them, s_W^2, and $s_T^2 = s_B^2 + s_W^2$. The total variance is estimated by Eq. (4). The variance within any one treatment is estimated by the same formula but for only those data in that treatment. Pooling estimates for all treatments gives s_W^2. To complete the analysis we compute also a quantity B:

$$B = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\bar{z}_i - \bar{\bar{z}}\right)^2 \tag{12}$$

where n_i is the number of plots in the ith treatment, z_i is the mean of the ith treatment, and $\bar{\bar{z}}$ is the general mean of the data. The computations are set out in Table 2. Finally, the analysis leads to a test of significance. We compute the ratio $F = B/W$, the distribution of which has been worked out and tabulated for degrees of freedom $k - 1$ in the numerator and $N - k$ in the denominator. If F exceeds the tabulated value at probability $P = 0.05$ (or $P = 0.01$ or $P = 0.001$, according to choice) we judge the treatments to have produced significant differences. As above, however, now that values of P can be calculated so readily we should its value for the ratio F.

In the simple experiment illustrated in Fig. 2A all n_i are equal to 5, so that n_i can be replaced by $n = 5$, and $N = mk = 5 \times 4 = 20$. Things do not always go as planned, however, and if some of the plot yields are lost then the n_i can vary from treatment to treatment, and the more general formulae in Table 2 will take care of that.

The soil might vary across the experiment systematically, so that there is trend, or in an apparently random way at a coarse scale. This variation could swell the residual variation and so mask that due to the treatments. It can be taken into account by blocking. The m replicates are now arranged in m blocks such that in each block every treatment appears once and once only. Fig. 2B shows an example in which five blocks are laid out side by side. The analysis follows the same procedure as in Table 2 except that there is an additional line for the blocks in which the sum of squares is that of deviations from the block means, as in Table 3. The residual sum of squares is diminished by this quantity, and although the residual degrees of freedom are also diminished the residual mean square, i.e., the residual variance is usually less, and the experiment more sensitive therefore.

Experiments should be planned with some hypotheses in view. The treatments should be chosen and arranged in designs to test them. The analysis that follows any experiment must fit its design; no other will do (Webster and Lark, 2018). Further, any comparisons between means should be ones that were set out at the start of the investigation; ideally they should be orthogonal to one another. The common practice of comparing all possible pairs of treatments afterwards is to be deprecated. Despite an investigator's designing an experiment well and analyzing the results in accord with nature may upset the investigator's unforeseen assumptions. Webster and Lark (2019) guide readers on what to do in those circumstances by transformations of the data.

Table 2 Table for the analysis of variance for a completely randomized design with a single classification.

Source	*Degrees of freedom*	*Sum of squares*	*Mean square*	F
Treatments	$k-1$	$\sum_{i=1}^{k} n_i\left(\bar{z}_i - \bar{\bar{z}}\right)^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\bar{z}_i - \bar{\bar{z}}\right)^2 = B$	B/W
Residual	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2$	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2 = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{\bar{z}}\right)^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{\bar{z}}\right)^2 = T$	

Table 3 Table for the analysis of variance for a balanced randomized block design with a single set of treatments.

Source	*Degrees of freedom*	*Sum of squares*	*Mean square*	F
Treatments	$k-1$	$\sum_{i=1}^{k} n_i\left(\bar{z}_i - \bar{\bar{z}}\right)^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\bar{z}_i - \bar{\bar{z}}\right)^2 = B$	B/W
Blocks	$m-1$	$\sum_{l=1}^{m} n_l\left(\bar{z}_l - \bar{\bar{z}}\right)^2$	$\frac{1}{m-1}\sum_{l-1}^{m} n_l\left(\bar{z}_l - \bar{\bar{z}}\right)^2 = M$	
Residual	$(k-1)\times(m-1)$	$T(N-1) - B(k-1) - M(m-1)$	$\frac{T(N-1) - B(k-1) - M(m-1)}{(k-1)(m-1)} = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{\bar{z}}\right)^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \bar{\bar{z}}\right)^2 = T$	

In soil survey different classes of soil replace treatments—see Webster and Lark (2013). In the simplest cases each class is sampled at random, and the n_i are rarely equal, either because it is difficult to obtain equal representation or because we deliberately sample in proportion to area so as to maintain a fairly constant sampling density. Effectively the classes are weighted in proportion to the areas they cover. We can still compute $F = B/W$ and test for significance, but as above this is less interesting than the differences between the means.

The variance between treatments or classes, s_B^2, can be obtained from B. The latter combines variation both from between treatments or classes and within them:

$$B = ns_B^2 + s_W^2 \tag{13}$$

if $n_i = n$ for all i. Rearranging then gives

$$s_B^2 = \left(B - s_W^2\right)/n \tag{14}$$

If the n_i are unequal then n is replaced by n^*, given by

$$n^* = \frac{1}{k-1}\left(N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right) \tag{15}$$

and

$$s_B^2 = \left(B - s_W^2\right)/n* \tag{16}$$

Fixed effects, random effects and intra-class correlation

The value s_B^2 obtained as above estimates $\sum_{i=1}^{k}(\mu_i - \mu)^2/(k-1)$, where μ_i is the expected value of treatment or class i, and μ is the expected value in the whole population. In designed experiments, in which the effects are fixed by the experimenter, it is of little interest. In soil survey, however, where it is often a matter of chance which classes are actually sampled, the differences $\mu_i - \mu$ are subject to random fluctuation. In this event s_B^2 estimates the variance, σ_B^2, among a larger population of means and is termed a component of variance, and it is of considerable interest.

The between-class variance expressed as a proportion of the total variance, $\sigma_B^2 + \sigma_W^2$, is the intraclass correlation, ρ_i:

$$\rho_i = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{17}$$

which is estimated from the analysis variance table by

$$r_i = \frac{s_B^2}{s_B^2 + s_W^2} = \frac{B - W}{B + (n* - 1)W} \tag{18}$$

The intra-class correlation has a theoretical maximum of 1 when every class is uniform. In practice there is always some variation within classes, and so ρ_i never attains 1. The minimum of ρ_i is zero, when $s_W^2 = 0$. The calculated estimate of ρ_i can be negative, because $B < W$, and is usually best explained by sampling fluctuation.

Covariance, correlation and regression

The relations between two variables can be expressed by correlation and regression.

Covariance

The covariance of a pair of variables, y and z, is estimated from data in a way analogous to the estimation of the variance, Eq. (4), by

$$\widehat{c}_{y,z} = \frac{1}{N-1}\sum_{i=1}^{N}\{y_i - \overline{y}\}\{z_i - \overline{z}\} \tag{19}$$

where $\overline{y}$ and $\overline{z}$ are the means of y and z respectively. It is not easy to envisage, especially if y and z have different dimensions. The relation may be standardized by its conversion to correlation, below.

Correlation

The correlation between two variables, strictly the linear correlation, or the product-moment coefficient of linear correlation, is a dimensionless quantity, usually denoted by ρ for the population parameter. Its estimate, r, is obtained from the covariance by

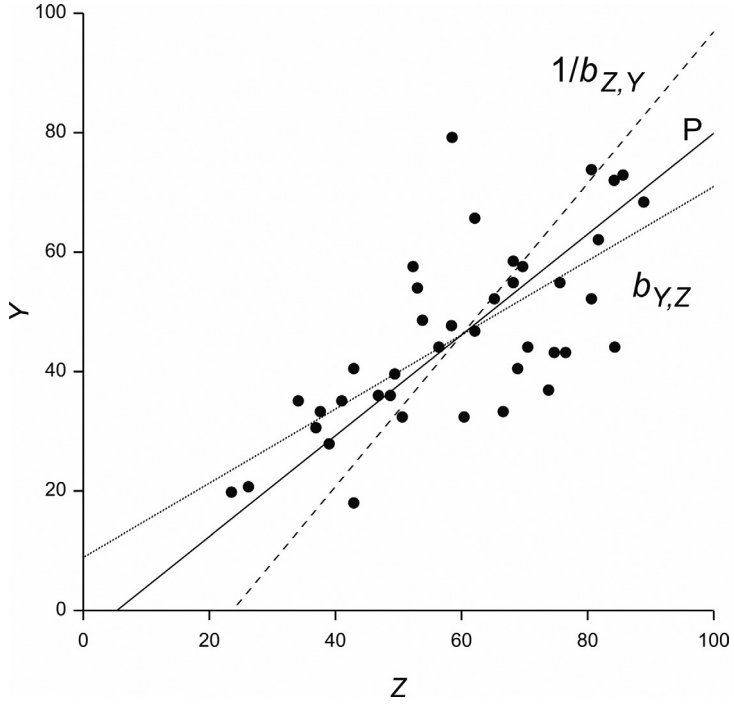

Fig. 6 Scatter graph showing the relations between two variables *Y* and *Z* for which the correlation coefficient, *r*, is 0.699. The symbols are the observed values, the solid line, labeled P, is the principal axis with gradient 0.844, equivalent to an angle of 40.1° to the horizontal. The dashed line (- - -) shows the regression of *Y* on *Z* with gradient $b_{Y,\,Z} = 0.621$, and the dotted line (. . .) shows that of *Z* on *Y*, for which $b_{Y,\,Z} = 0.788$, giving it a gradient in the figure of $1/b_{Z,\,Y} = 1.269$.

$$r_{\gamma,z} = \frac{\widehat{c}_{\gamma,z}}{\sqrt{s_\gamma^2 s_z^2}} \tag{20}$$

$s_\gamma{}^2$ and $s_z{}^2$ are the estimated variances of γ and z. It was proposed by Karl Pearson, and for that reason it is often called the Pearson correlation coefficient.

The coefficient is effectively a standardized version of the covariance. It measures the extent to which the data when plotted as one variable against the other in a scatter graph depart from a straight line; see Fig. 6. It may vary between +1, signifying perfect positive correlation, and −1 for perfect negative correlation. Intermediate values indicate departures from the straight line, as in Fig. 6, for which $r = 0{:}699$. In general positive values of r indicate the tendency of γ and z to increase together, whereas negative values arise when γ increases as z decreases. A value of 0 represents no linear relation.

Notice that the statistic refers specifically to linear correlation. The relation between two variables might be curved; the absolute value of r would then be necessarily less than 1 regardless of any scatter about the curve.

When the data, γ_i; z_i; $i = 1, 2, \ldots, N$, are from a sample then $c_{\gamma,\,z}$ and $r_{\gamma,\,z}$ estimate corresponding population parameters, $\mathrm{cov}_{\gamma,\,z}$ and $\rho_{\gamma,\,z}$. If the data can be assumed to be drawn from a bivariate normal distribution then one can test r for significance. One computes Student's t with $N - 2$ degrees of freedom:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{21}$$

The probability of this value's occurring on the null hypothesis that $\rho = 0$ can then be computed or looked up in a table of t.

Spearman rank correlation

Where the distributions of the underlying variables are far from normal the Pearson coefficient can be replaced by Spearman's rank correlation coefficient, usually denoted r_S. The values of each variable are ranked from smallest to largest and given new values 1, 2, . . ., N. The correlation coefficient is then computed by applying Eqs. (19) and (20). Alternatively, one may take the differences, d_i; $i = 1, 2, \ldots, N$, between the ranks and compute

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d^2}{N\left(N^2 - 1\right)} \tag{22}$$

Many soil variables observed in the field, such as grade of structure and frequency of mottles, are recorded as rankings rather than measured. In these circumstances the correlations between them can be expressed by the Spearman coefficient, whereas the Pearson

coefficient would be inappropriate. Also in these circumstances tied ranks in any large set of data are inevitable, and Eq. (22) must be elaborated. The coefficient can be calculated in various ways, but from Eqs. (19), (20) and (22) we can derive

$$r_S = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{i=1}^{N} T_{i,y} + \sum_{i=1}^{N}(z_i - \bar{z})^2 - \sum_{i=1}^{N} T_{i,z} - \sum_{i=1}^{N} d^2}{2\sqrt{\left\{(y_i - \bar{y})^2 - \sum_{i=1}^{N} T_{i,y}\right\}\left\{(z_i - \bar{z})^2 - \sum_{i=1}^{N} T_{i,z}\right\}}} \tag{23}$$

in which

$$T_i = \frac{t_i(t_i^2 - 1)}{12} \tag{24}$$

where t_i is the number of observations tied at rank i.

For small samples r_S is somewhat less sensitive than the Pearson coefficient in that larger values are necessary to establish statistical significance.

Regression

Regression, which is dealt with comprehensively by Draper and Smith (1981), treats the relation between two variables in a somewhat different way by designating one of them, y, as depending on the other, z, represented by the equation:

$$y = \beta_0 + \beta_1 z \tag{25}$$

The underlying rationale is often physical. For example, we may wish to discover how the soil's strength is changed by additions of gypsum. We could add known amounts of gypsum to the soil, measure the resultant changes in strength, and from the data estimate by how much on average the strength is changed by each increment in gypsum added. We adopt the Gauss linear model for the purpose:

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \tag{26}$$

where y_i is the value of the random dependent variable, Y, here strength, in unit i, z_i is that of the independent variable, gypsum added, and ε_i is random error term that is uniformly and independently distributed with variance ${\sigma_\varepsilon}^2$. The quantities β_0 and β_1 are parameters of the model and are estimated as follows:

$$\hat{\beta}_1 = \frac{\hat{c}_{y,z}}{s_z^2} \tag{27}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z} \tag{28}$$

Eq. (27) gives the rate at which strength changes in response to increments in gypsum. The quantity $\hat{\beta}_0$ in Eq. (28) is the intercept at $y = 0$ and is likely to be of subsidiary interest. Together they may be inserted into Eq. (25) for predicting unknown values of Y if we know those of z:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{1z} \tag{29}$$

The procedure minimizes the sum of the squares of the differences between the measured values y_i; $i = 1, 2, \ldots, N$, and those expected from Eq. (25), $\hat{y}_i$:

$$s_{Y,z}^2 = \frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{30}$$

A somewhat different situation is common in soil survey. In it we sample the soil without knowing in advance what values we shall obtain for the variables of interest. For example, we may measure both cation exchange capacity, CEC, and clay content, and we may regard CEC as depending in a physical sense on clay content. The data on both variables now contain random components, and for this reason we designate them with the capital letters, Y and Z, respectively. We may express the relation as the regression of CEC on clay and estimate the parameters in the same way as for the Gauss linear model. In doing so we assign all the error to CEC and minimize the sum of squares, ${s_{Y,Z}}^2$, as in Eq. (30). The purpose is now prediction, i.e., prediction of CEC, knowing the clay content. We could equally well compute the regression of clay on CEC. The roles of Y and Z are reversed, and we minimize

$$s_{Z,Y}^2 = \frac{1}{N-2}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2 \tag{31}$$

where

$$\widehat{z} = \widehat{\beta'}_0 + \widehat{\beta'}_1 y \tag{32}$$

The primes attached to β_0' and β_1' signify that these quantities refer to the regression of Z on Y and that they are different from those for the regression of Y on Z. In other words, the line defined by Eq. (29) differs from that defined by Eq. (32), as Fig. 5 shows.

The correct line to choose depends on which variable is to be predicted. To predict y from z use Eq. (29); to predict z from y use Eq. (32).

Further matter

Research into the soil and modern developments in instrumentation are leading to ever more advanced statistical understanding and technique. In recent years pedologists have mastered the theory of regionalized variables to model and map spatially correlated soil variables using geostatistics; see the chapter on this topic. Soil biologists are measuring numerous properties of the soil's microora and its genetics and are analyzing the data by various multivariate techniques to identify relations, to group organisms and their functions. Multivariate analysis is also treated as a topic in the encyclopedia. Sensing in the visible and infrared parts of the electromagnetic spectrum, from the air and satellites, from close to the soil surface (proximal sensing) and in the laboratory is producing huge quantities of multivariate data, much of which is redundant. Techniques, including machine-learning, have been developed to compress and smooth the spectral data and extract features that have mineralogical, chemical and biological meaning, and to predict properties that are much more time-consuming and expensive to measure by earlier conventional methods. Machine learning is covered in another chapter. Note, however, that these techniques do not replace or invalidate the sound experimental design, random sampling and the appropriate analysis of the data thereby obtained.

References

Brus DJ (2021) Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science* 72: 686–703.

Cochran WG (1977) *Sampling Techniques*, 3rd edn. New York: John Wiley & Sons.

de Gruijter JJ, Brus DJ, Bierkens MFP, and Knotters M (2006) *Sampling for Natural Resources Monitoring*. Berlin: Springer-Verlag.

Draper NR and Smith H (1981) *Applied Regression Analysis*, 2nd edn. New York: John Wiley & Sons.

Snedecor GW and Cochran WG (1989) *Statistical Methods*, 8th edn. Ames, Iowa: Iowa State University Press.

Sokal RR and Rohlf FJ (2012) *Biometry: the Principles and Practice of Statistical Research*, 4th edn. San Francisco: W.H. Freeman.

Wasserstein RL and Lazar NA (2016) The ASA's statement on p-values: Context, process and purpose. *The American Statistician* 70: 129–133.

Webster R and Lark RM (2013) *Field Sampling for Environmental Science and Management*. London: Routledge.

Webster R and Lark RM (2018) Analysis of variance in soil Research: Let the analysis fit the design. *European Journal of Soil Science* 69: 126–139.

Webster R and Lark RM (2019) Analysis of variance in soil research: Examining the assumptions. *European Journal of Soil Science* 70: 990–1000.

Welham SJ, Gezan SA, Clark SJ, and Mead A (2015) *Statistical Methods in Biology: Design and Analysis of Experiments and Regression*. Boca Raton, Florida: CRC Press, Taylor and Francis Group.