The Journal of Physiology

STATISTICAL PERSPECTIVES

Variation: use it or misuse it – replication and its variants

Gordon B. Drummond[1]
and Sarah L. Vowler[2]
[1]*Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, 51 Little France Crescent, Edinburgh, EH16 4HA, UK*
[2]*Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK*

Email: g.b.drummond@ed.ac.uk

In a previous article (Drummond & Vowler, 2012*a*), we discussed variation between individuals, termed *variance*. When measurements are taken of subjects in separate groups, the variance of the measurements *within* the group can be compared with the variance that exists *between* groups. This allows us to estimate the likelihood that groups were drawn from the same population. Commonly, a treatment has been applied to these groups, and the aim would be to find if this had increased the variance between the groups, which might then indicate that the treatment had caused an effect (Fig. 1*A*).

We have also discussed how to assess correlation, the tendency of two measurements, taken of the same subject, to vary 'in the same direction' (Drummond & Vowler, 2012*b*). This concept can be extended to study the tendency of a particular measurement, made in individuals from a specific (non-random) group, to be *less* variable than measurements made from a more general population (Fig. 1*B*). This is a common feature, particularly in ecology, where the ability to sample at random may be restricted. However, this feature can be introduced easily and inadvertently into laboratory studies by poor study design or lack of attention to detail. The important effect is that samples are no longer truly random. Individuals in the group could have a particular feature that accounts for their presence in that particular group. This 'fixed' (i.e. non-random) factor may affect the results. How this factor could possibly affect the measurements of the group, or even the nature of the factor itself, may not be clear, other than in a prosaic way. For example, we can study only the frogs that we can catch. This fixed factor (catchability) may affect the measurements we are making. It might be that we only catch the ones that can't jump too far, or those that jump towards a flashlight, or those that are awake, or sexually active. . . . There are lots of reasons, but we must recognize that a potential fixed factor is at play, which could affect the measurements. Lab experimental samples are often far from random, and may be very restricted, such as studies of genetic modifications.

Our group of frogs isn't random. These captive frogs are clustered, affected by the fixed factor: the fact that we have been able to catch them. The frogs in the cluster are more related to each other, and less related to the frog population that remain free. This factor may affect how far they jump. We can estimate how related these frogs are in terms of jumping distance. The method is similar to the way we compare the 'related-ness' of pairs of measurements: a correlation. The *intra-class* correlation indicates how closely these frogs are like each other, in respect of the measurement we are making. If they all jump exactly the same distance, then there is a perfect correlation, and intra-class correlation = 1. If their jumps vary as much as the population as a whole (assuming we can estimate it!), there is no correlation within the cluster measurements, and intra-class correlation = 0.

Correlation of features within a group causes problems with many statistical tests. Fixed factors can reduce the within group variation between measurements.

Key points

- Variation between measurements may be reduced if sampling is not random
- Fixed factors can reduce variation
- Replicates are repeated measurements from within an experimental unit
- Replication means that samples can no longer be considered random
- Comparisons of replicated values require correction, usually by reducing the effect of sample size
- Correction requires knowledge of the intra-class correlation

If variation is less than would be expected if sampling were truly at random, then we cannot apply tests that assume that the variance results from random sampling. The measurements now contain *less* variance than would be expected. By taking repeated measures from a sample that is subject to a fixed factor, we may commit *pseudo-replication*. The term was popularized in a long and idiosyncratic article by the ecologist Hurlbert, which is widely cited, but is far from an easy read (Hurlbert, 1984). The article uses colourful vocabulary to discuss a diverse range of concepts, such as rage between statisticians, original sin at Rothamstead, human sacrifice, pseudo-experiments and pseudo-design. One observation in the paper repeats a familiar theme: 'Statistics and experimental design are disciplines with an impoverished vocabulary.' In other words, the specialized, opaque, and non-colloquial use of some of the terms that are used to tackle these important concepts can confuse and bewilder the novice. We shall choose simpler expressions in an attempt to explain the problem, although they may be less exact.

What exactly is a replicate? It is literally 'another version of the same thing', but in this context, it is a repeated measurement from the *same* experimental unit. An experimental unit is the smallest part of an experiment or a study that can be subject to a discrete treatment, or has a recognizable fixed factor present. Other similar words for such an element or unit could be a group, nest, cluster, or cell, although all of these words can be used in other ways. A unit is often a group of individuals, such as the group of frogs that we have captured. Replication consists of measuring

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 9th in a series of articles on best practice in statistical reporting. All the articles can be found at http://jp.physoc.org/cgi/collection/stats_reporting

DOI: 10.1113/jphysiol.2012.234260

more than one individual from within an experimental unit which has been given the same treatment, or has been exposed to the same fixed effects, or has been drawn from the same group. Of course, there is likely to be variation in the values we find in replicate measurements. By taking several samples we can obtain a more representative picture *of that particular unit as a whole*, and this is perfectly legitimate, but we should not assume that the variance (usually gratifyingly small) could be attributed to random sampling.

Individual animals, rather than groups, may of course be 'experimental units'. If we took our captured frogs and randomly divided them into two groups, we could conduct a comparison of two different treatments given to the two randomly selected groups, which both represent the 'captured frog population'. Each frog would be a 'unit' because it had been individually and randomly assigned to a treatment. The variance contributed by each individual would be related to its random origin (from within the group of captured frogs) and the conclusion we drew from our study would be valid, for captured frogs. We might, with reservations, wish to extrapolate the finding to all frogs, but we should do this circumspectly, since the group has not been taken at random.

What is pseudoreplication? This error comes from assuming that all the measurements made of elements within the experimental unit have variance that can be attributed to random factors, when in fact some additional factor is present which could alter (usually reduce) this variance, and then taking the data *out of this unit*, without qualification, and managing them as if they were a random sample. If there is some correlation between the replicate measurements obtained from within an experimental unit, the variance is less than would be expected if the numbers were randomly selected. The practical importance is that the influence of that number of measurements, when used elsewhere, counts for less than we might expect. We have to be more sceptical about their importance, to paraphrase a well-known political saying: 'well they would show that, wouldn't they?'

If we were interested in the features of the entire frog population of California, we would try to randomly sample from all the frogs in the State. This is probably impossible, and not what usually happens in biological experiments anyway. Suppose rather than sampling randomly, we proposed to take a random sample of counties from within the State, and then sample frogs from these counties. One of those randomly chosen counties happens to be Calaveras. Over the years, several champion jumping frogs have escaped after the annual competition. Their progeny are all pretty good jumpers! In the other counties, the frogs we sample are similarly related genetically to others from the same region. Although the jumping abilities of frogs in each county are quite similar, these can vary from place to place, and very few are up to Calaveras standard (Fig. 2). Thus we find that there is an intra-county correlation in jumping abilities in our sampling units (counties).

Using these samples, we can calculate the average jump distance of frogs from each county. However, we are not justified in comparing the counties by taking all the jump distances of these frogs from each sample as a separate measure. We should correct for the correlation between the replicate measures from each county, to take account of the reduced variation. In practice, we reduce the effect of the number of observations in the calculation. This will reduce the likelihood of concluding that there could be a significant difference. To make this correction, a measure of the intra-class correlation is required, and information from previous studies may be needed. However, careful thought or inspection of the data can indicate if there is a need for correction.

Compare two possible experiments. In the first, we wish to see if training affects muscle strength in frogs. We take two samples of 20 frogs at random from Abilene county and train one sample. We measure the muscle strength in the leg muscles of the frogs using a sciatic nerve–gastrocnemius preparation.

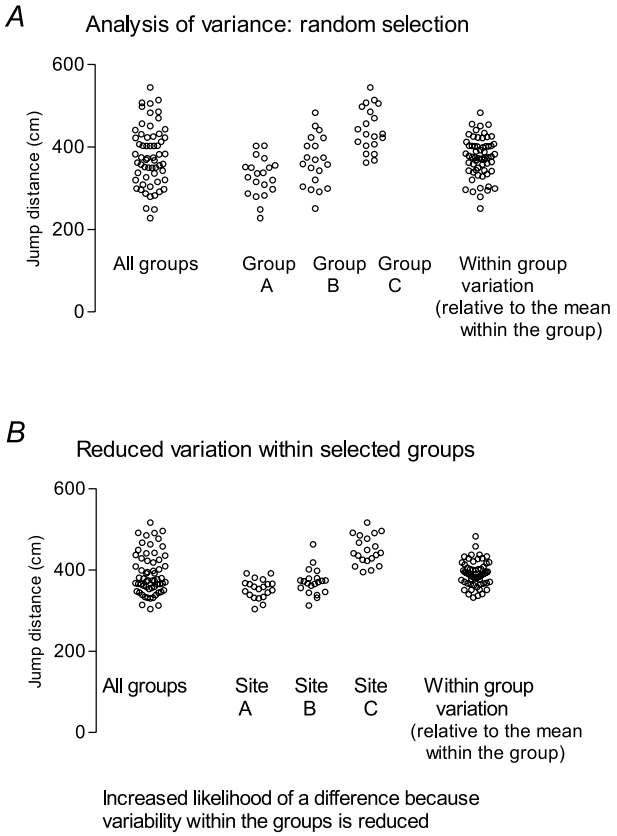

Figure 1
A, analysis of variance is based on comparing variance between the groups, possibly due to a treatment effect, with the variance observed within the groups. It assumes that the groups are drawn at random. *B*, if groups are selected on the basis of another feature, such as the site where they are found, a fixed effect is introduced that may reduce the variance within the groups. Comparison of variance between groups could then yield an apparently more prominent effect.

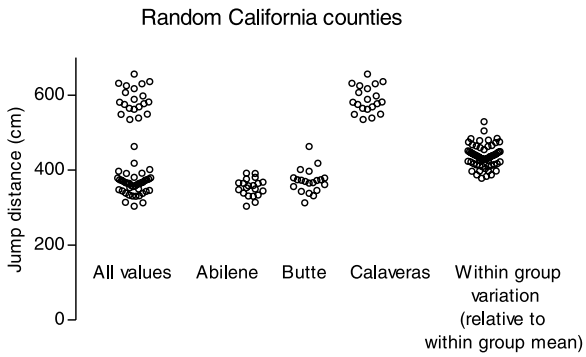

Figure 2
Samples from three counties showing the effect of a fixed factor: a correction should be applied for comparisons between the groups, to reduce the effect of the reduced within group variation.

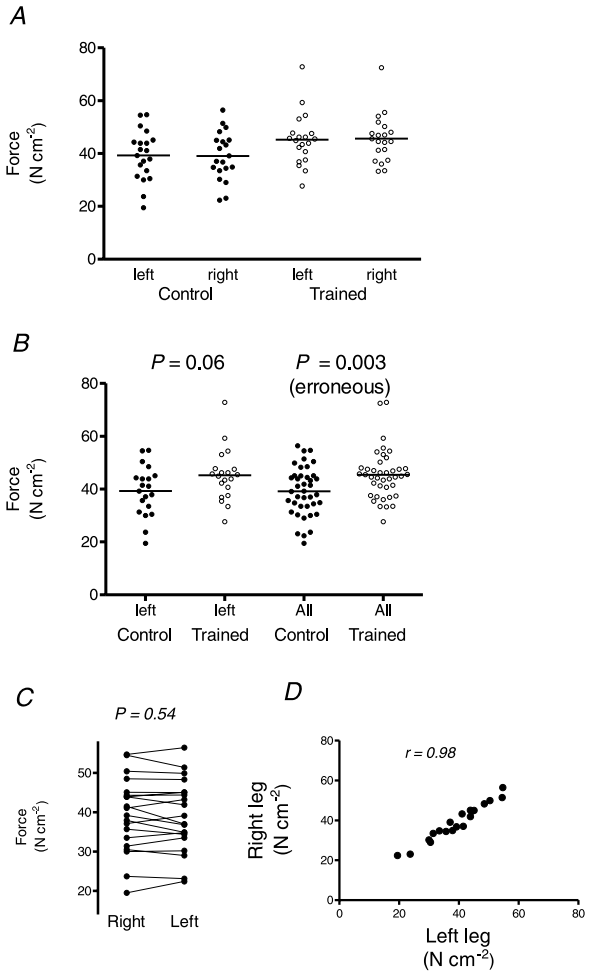

Figure 3
A, muscle force developed in all samples. Two samples of frogs have been taken, one trained. Two sciatic–gastrocnemius preparations have been made from each frog. *B*, comparison of control and trained groups using a single sample from each frog, and comparison using two samples from each frog. The second comparison has NOT been corrected for intra-class correlation (relationship between right and left samples). *C*, comparison of right and left preparations in the control frogs. There is no difference between the paired samples. *D*, relationship between right and left preparations in the control frogs. There is a strong correlation between the tensions developed between right and left leg muscles.

We reason that if we take two preparations from each frog, we could have a greater chance of finding an effect.

Fig. 3*A* shows the results, displaying each sample from each frog, for right and left legs. In Fig. 3*B*, we compare the left leg preparations and the *P* value is 0.06. On the basis of this finding, we might conclude that training shows no evidence of an effect. When we combine the data from right and left leg preparations we get a *P* value of 0.003. However, this comparison is based on the premise that the samples are random samples, which is not the case. Pairs of samples have come from each frog. In each frog, the relationship between the strength of the right and left legs varies slightly and randomly. However, the variation is small: much smaller than the variation between the frogs. This can be seen in Fig. 3*C*. When the right and left leg samples are compared with a paired test, there is no evidence of a difference between the strengths. However, there is a strong correlation between the strength of the right and left legs in each frog (Fig. 3*D*).

In fact, this feature could be usefully exploited if we wanted to compare the effects of applying a drug to the preparation. We would have a statistically efficient comparison if we applied the drug to the preparation from one leg, and used the other leg as a control, in each pair of preparations. This is because there is only a small, random, insignificant difference between these preparations before the treatment is applied.

Pseudoreplication is a widespread and under-recognized problem in biology (Lazic, 2011). In part, this occurs because experimental units (animals, brain slices, cell bodies) can be sparse and expensive, and the impulse to wring results out of the last scrap of data is overwhelming. In addition, there is a desire or perceived obligation to apply tests for statistical significance whenever measurements are reported, even if the logic of significance testing in descriptive studies may be tenuous. Pseudoreplication may be unavoidable. One of the simplest ways to acknowledge and describe the problem is to report test details in full, particularly the sample size, degrees of freedom, and the test statistics.

On the other hand, reduced variation may be a distinct advantage in some experiments, because if the 'noise' is less, then the signal can be detected more easily. Variation can be reduced by suitable study design. In one of our previous examples, we used the

proposal that breeding our frogs reduced the variation in the progeny, and we could then demonstrate an effect that was not shown in the original population. When genetic and familial factors are shared, and if the treatment can be randomized *within* a group with shared fixed factors, the number of individuals needed to demonstrate an effect can be reduced. Similarly, careful design of studies to assess and attribute variation to different factors can increase power, and reduce the number of subjects needed. Randomization can take into account fixed factors such as species, strain, sex, age, and so on, by using a blocked design in which the sample is first divided in terms of the chosen factors, and the sub-groups are then randomized (Festing, 2003). This process recognizes and exploits shared and distinct features from the total data set.

References

Drummond GB & Vowler SL (2012*a*). Analysis of variance: variably complex. *J Physiol* **590**, 1303–1306.

Drummond GB & Vowler SL (2012*b*). Categorized or continuous? Strength of an association – and linear regression. *J Physiol* **590**, 2061–2064.

Festing MFW (2003). Principles: The need for better experimental design. *Trends Pharmacol Sci* **24**, 341–345.

Hurlbert SH (1984). Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* **54**, 187–211.

Lazic SE (2011). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* **11**, 5.