ESC European Society of Cardiology

Cardiovascular Research (2017) **113**, 1703–1704
doi:10.1093/cvr/cvx195

EDITORIAL

Avoiding phantasms

Zana Coulibaly[1], Ye Chen-Izu[1,2,3], and Leighton T. Izu[1]*

[1]Departments of Pharmacology, University of California, Davis, 1 Shields Avenue, Davis, 95616 CA, USA; [2]Biomedical Engineering, University of California, Davis, 1 Shields Avenue, Davis, 95616 CA, USA; and [3]Internal Medicine, University of California, Davis, 1 Shields Avenue, Davis, 95616 CA, USA

Online publish-ahead-of-print 26 September 2017

This editorial refers to 'Hierarchical statistical techniques are necessary to draw reliable conclusions from analysis of isolated cardiomyocyte studies' by M.B. Sikkel *et al.*, pp. 1743–1752.

In the book entitled *The Seven Pillars of Statistical Wisdom*, Stephen Stigler[1] tells the story of William Stanley Jevons's obsession with finding matches between the periodicities of sunspot activity and business cycles. Having found an apparent equality, Jevons constructs a model of the causal relationship between these two cycles. His model was ridiculed by his contemporaries and did not stand the test of time. Jevons's problem was not the implausibility of his model for the relationship between sunspot activity and business cycles. Rather, Jevons erred in building a model based on the *false premise* that the two cycles had the same period. Unless two periodic functions have exactly the same period, they will eventually move further and further out of phase with each other. However, on a short time scale, the cross-correlation between two unrelated cycles of slightly different periods can be very high and it is so tempting, as Jevons fell victim, to construct a causal model. Now we might smugly smile at Jevons's naiveté; the statistical tools developed in the 150 years since Jevons's time ought to save us from Jevons's fate. However, we should not forget the warning of Ecclesiastes 1:9: *What has been will be again, what has been done will be done again; there is nothing new under the sun.*

Sikkel *et al.*[2] show us how *we* might inadvertently fall into the same trap as Jevons—and importantly, how to spot and avoid the trap. The *Methods* section of almost any article in *Cardiovascular Research* will have a subsection on the statistical tests used and what *P* value is deemed the threshold for significance (typically 0.05). These statistical tests, frequently the Student's *t*-test, or some kind of analysis of variance are supposed to reduce the chances of wrongfully rejecting the null hypothesis (usually the uninteresting hypothesis we hope to reject) and committing a Type I error (Perhaps the most famous and important experiment where the null hypothesis was not rejected was Michelson and Morley's measurement of the speed of light along directions parallel and perpendicular to the propagation direction. Not rejecting the null hypothesis–no difference in the speed of light along these directions—set the foundation for Einstein's relativity theory.) Most of us are aware that many parametric tests (such as the *t*-test) assume that the underlying data are normally distributed, and we make some test to check whether our data conform to this assumption. Most statistical tests also assume that the data are independent (apart from the treatment variable). But as Sikkel *et al.*[2] and others[3,4] point out, many researchers ignore this critical assumption, which can result in a falsely small *P* value.

When there is a correlation in the data, the data cannot be treated as being independent statistical analysis. Consider a simple hypothetical example where $m = 10$ measurements of blood glucose concentration are taken from each of $n = 3$ rats (labelled red, blue, and green). The true blood glucose distributions for each rat (based on the number of samples m taken from each rat approaches infinity) are shown in *Figure 1*, and the black curve is the population distribution that is obtained as the number of rats, n, sampled approaches infinity. If the number of samples m from each rat increases, we obtain a better estimate of the mean glucose concentration of each rat, but we get no better estimate of the population mean because n remains equal to 3. In this example, it is clear that we should not treat each measurement as independent and estimate the population mean as the average of these $10 \times 3 = 30$ measurements. More mischief ensues if we treat the measurements as independent and

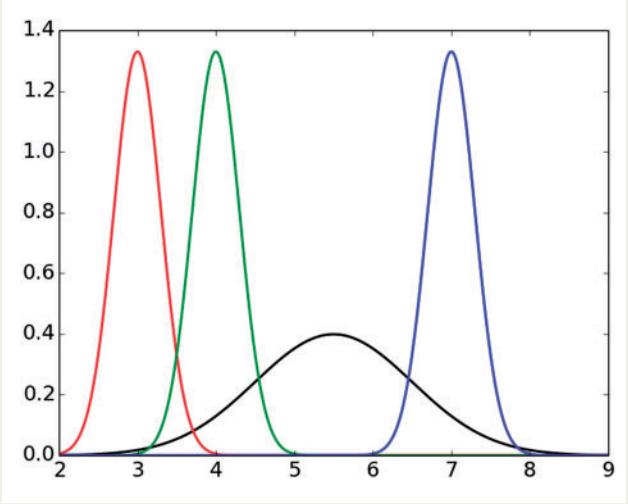

Figure 1 Hypothetical distribution of glucose concentrations in three rats (red, green, and blue curves) and that of the population (black curve).

The opinions expressed in this article are not necessarily those of the Editors of *Cardiovascular Research* or of the European Society of Cardiology.

* Corresponding author. Tel: +1 530 752 0324, fax: +1 530 752 7710, E-mail: ltizu@ucdavis.edu

estimate the error of the mean as $s/\sqrt{30\text{-}1}$, where s is the sample standard deviation (of 30 measurements). If we did the same procedure for a set of three rats that had undergone some treatment and did a *t*-test to compare the means between the two groups, we might find that the *P* value is less than 0.05, and we would happily publish our finding that the treatment had an effect. But this is patent nonsense. Taking more and more measurements from each of the three rats (letting $m \rightarrow \infty$) would make the standard error of the mean appear to approach zero, and *P* could be much less than 0.001, but, in fact, we are no closer to knowing the true population mean value than before. We have committed a Type I error and most likely constructed some mechanistic explanation, like Jevons, to explain the false premise.

Some of us might be guilty of treating each measurement as an independent data point for understandable, but not justifiable, practical reasons. For example, we might measure the L-type Ca^{2+} current in *m* cells from *n* animals where $m > 1$ to give a total of $m \times n$ data points. Animals can be very expensive so *n* can be quite small, and it is natural to get many measurements (large *m*) from each animal. One approach—usually not taken—is to average the *m* measurements for each animal and use the *n* average values for the statistical tests. Because the standard error of the mean scales as $1/\sqrt{n\text{-}1}$, the statistical test might not indicate significance when *n* is small even when the treatment is actually effective. In other words, the statistical test is too conservative, and we have committed a Type II error.

How do we navigate between fooling ourselves by committing a Type I error and missing a real effect by committing a Type II error? Sikkel *et al.* suggest using hierarchical statistical techniques that can detect clustering of data (as in *Figure 1*) and calculate *P* based on the value of the *intraclass correlation coefficient* (*ICC*). The ICC varies between zero for no clustering (where there are really $m \times n$ independent data points) and one for data that have identical values in each cluster (infinitely narrow distribution for each rat in *Figure 1*), but different clusters have different values (where now there are only *n* independent data points). Importantly, Sikkel *et al.* provide an algorithm in the open source R language to calculate ICC and *P*. With this algorithm, we can sleep better knowing that we do not have to throw away all those hard-earned $m \times n$ data points and at the same time not fooling ourselves with artificially small *P* values and expending energy and time making mechanistic models of some phantasm. We wish the readers a good night's sleep.

Funding

This work was supported in part by an NIH Cardiovascular Training Grant (T32 HL0863500) Z.C. and NIH (RO1 HL123526) to Y.C. and L.T.I.

Conflict of interest: none declared.

References

1. Stigler S. *Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard University Press, 2016.
2. Sikkel MB, Francis DP, Howard J, Gordon F, Rowlands C, Peters NS, Lyon AR, Harding SE, MacLeod KT. Hierarchical statistical techniques are necessary to draw reliable conclusions from analysis of isolated cardiomyocyte studies. *Cardiovasc Res* 2017;**113**:1743–1752.
3. Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* 2010;**11**:5.
4. Scariano SM, Davenport JM. The Effects of violations of independence assumptions in the one-way ANOVA. *Am Statis* 1988;**41**:123–129.