Appendix C *Post hoc* power assessment

Estimating power based on an experimentally observed result (i.e. *post hoc* power assessment) has been shown to be illogical and invalid for 0D data (Hoenig and Heisey, 2001). This Appendix aims to explain why in the context of 1D data.

Consider the four datasets depicted in Fig.C.1. Each depicts a true signal to which smooth Gaussian continua were added. Hypothesis testing results (from one-sample t tests) for each dataset are depicted in Fig.C.2. These hypothesis testing results represent all four possible outcomes of an arbitrary experiment. That is, in the case of a null effect one can either correctly fail to reject the null hypothesis (H0) (Dataset A, true negative) or incorrectly reject H0 (Dataset B, false positive). In the case of a true effect one can either correctly reject H0 (Dataset D, true positive) or incorrectly fail to reject H0 (Dataset C, false negative). Note that we constructed these datasets to convey specific points regarding power so we encourage readers to judge their relevance to real Biomechanics experiments.

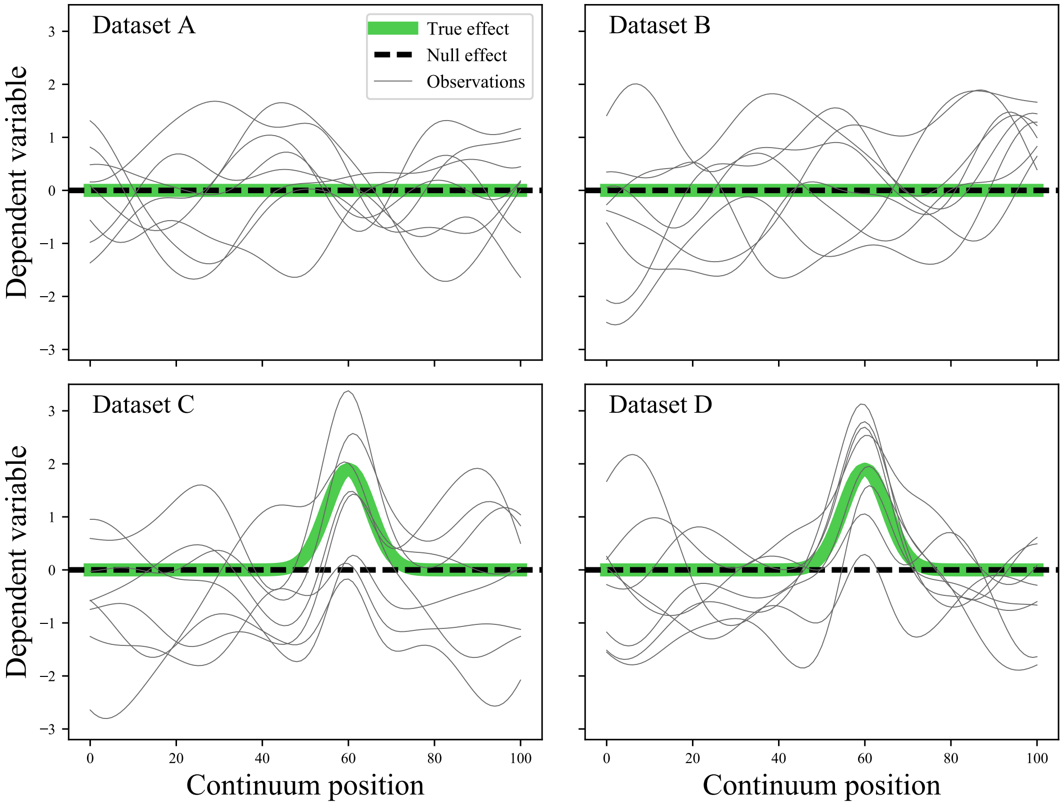

Figure C.1: Simulated datasets.

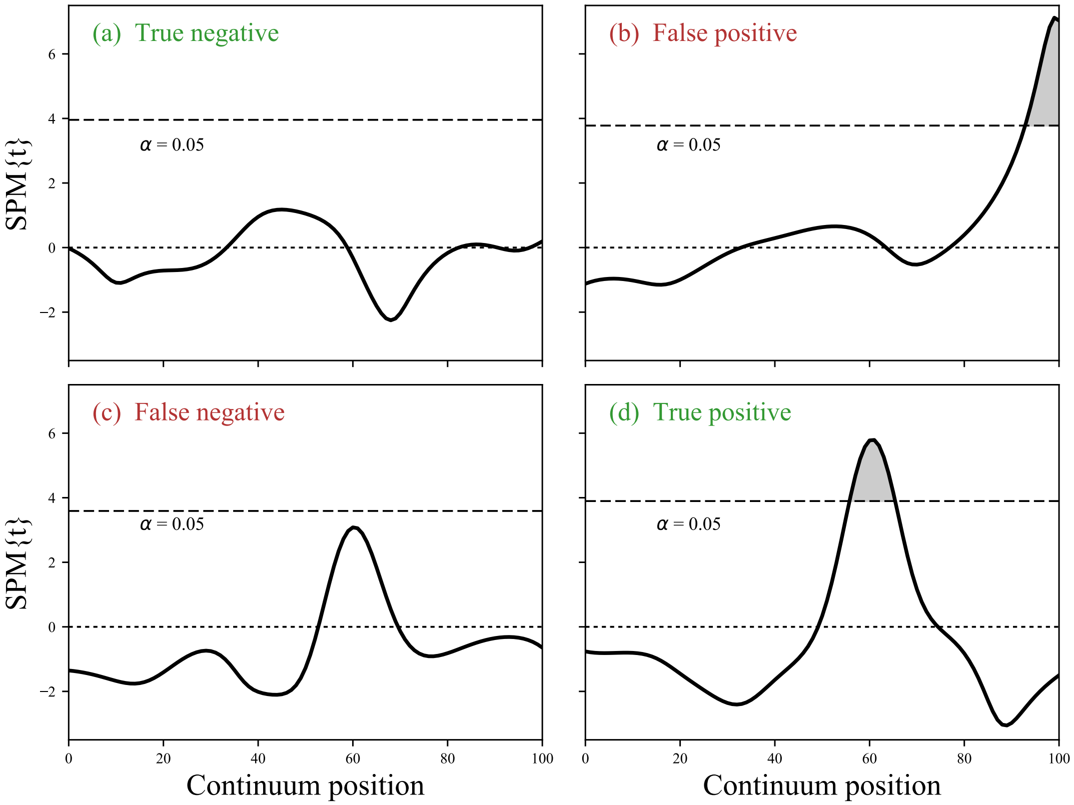

Figure C.2: One-sample t test results. The solid black line depicts the test statistic (t value) continuum. The horizontal line depicts the critical threshold at a Type I error rate of α=0.05.

While Fig.C.2 represents all possible experimental outcomes, power analysis pertains only to those cases where true effects exist (i.e. Datasets C and D). Nevertheless, to understand why *post hoc* power analysis is invalid it is useful to first consider the other cases (Datasets A and B), where there is no true signal, and in particular the concept of Type I error.

Type I error (α) is also termed a 'false positive' and refers to the mistake of inferring an effect when none exists (Dataset B). The Type I error rate is set before an experiment, conventially at α=0.05, and one applies this criterion to the observed data to either reject H0 (Fig.C.2b) or fail to reject H0 (Fig.C.2a). Note that it is possible, albeit scientifically invalid, to adjust α after observing the data. For example, given the results in (Fig.C.2a), one could increase the value of α until the critical threshold decreases enough to reject H0; in Fig.C.2a one would need to use α=0.65 to reject H0. It is scientifically invalid to adjust α in this *post hoc* manner because it is non-objective. In other words, α pertains not to a specific experiment, but instead to the infinite set of identical experiments in which no true effect exists, but in which random 'effects' are observed due to random sampling.

Similarly, Type II error (β) is termed a 'false negative' and refers to the mistake of inferring no effect when one in fact exists (Dataset C). It is related to power as: power = (1 - β), where power is the probability of inferring an effect when one truly exists. Similar to α, β must be set before an experiment because it pertains not to a specific experiment, but instead to the infinite set of identical experiments in which a specific effect exists, and in which a range of 'effects' are observed experimentally due to random sampling. Identical to α, β mustn't be computed based on the results of an experiment because a particular experiment's 'effect' may have been caused by random sampling. That is, one can never know what the true effect is based on an experimentally observed effect, so neither α nor β should be computed based on experimentally observed effects.

In summary, one must specify both α and β (and thus power and effect size) only in an *a priori* manner because random sampling ensures that an experimentally observed effect is generally unrelated to the underlying true effect. Further considerations of *post hoc* power calculations are provided in Hoenig and Heisey (2001).

References

Adler, R. and Hasofer, A. (1976). Level crossings for random fields. *The Annals of Probability*, 4(1):1–12.

Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry.* Springer-Verlag.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Genesis Publishing Pvt Ltd.

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., and Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, 4(3):223–235.

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., and Laurienti, P. J. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, 37(3):721–730.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power. *The American Statistician*, 55(1):19–24.

Hopkins, W. G. and Batterham, A. M. (2016). Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Medicine*, 46(10):1563–1573.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education*, 61(4):317–333.

Knudson, D. (2017). Confidence crisis of results in biomechanics research. *Sports Biomechanics / International Society of Biomechanics in Sports*, page in press.

Mumford, J. and Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39(1):261–268.

Pataky, T. C. (2017). power1d: Numerical power estimates for one-dimensional continua in Python. *Journal of Statistical Software*, 3:e125.

Pataky, T. C., Robinson, M. A., and Vanrenterghem, J. (2013). Vector field statistical analysis of kinematic and force trajectories. *Journal of Biomechanics*, 46(14):2394–2401.