Pataky, TC, Robinson, MA and Vanrenterghem, J

A computational framework for estimating statistical power and planning hypothesis-driven experiments involving one-dimensional biomechanical continua.

A computational framework for estimating statistical power and planning hypothesis-driven experiments involving one-dimensional biomechanical continua

Todd C. Pataky[a], Mark A. Robinson[b], Jos Vanrenterghem[c]

[a] *Institute for Fiber Engineering, Department of Bioengineering, Shinshu University, Japan*
[b] *Research Institute for Sport and Exercise Science, Liverpool John Moores University, UK*
[c] *Department of Rehabilitation Sciences, University of Leuven, Belgium*

Abstract

Statistical power assessment is an important component of hypothesis-driven research but until relatively recently (mid-1990s) no methods were available for assessing power in experiments involving continuum data and in particular those involving one-dimensional (1D) time series. The purpose of this study was to describe how continuum-level power analyses can be used to plan hypothesis-driven biomechanics experiments involving 1D data. In particular, we demonstrate how theory- and pilot-driven 1D effect modeling can be used for sample-size calculations for both single- and multi-subject experiments. For theory-driven power analysis we use the minimum jerk hypothesis and single-subject experiments involving straight-line, planar reaching. For pilot-driven power analysis we use a previously published knee kinematics dataset as a pilot dataset. Results show that powers on the order of 0.8 could be achieved with relatively small sample sizes, five and ten for within-subject minimum jerk analysis and between-subject knee kinematics, respectively. However, the appropriate sample size depends on *a priori* justifications of biomechanical meaning and effect size. The main advantage of the proposed technique is that it encourages *a priori* justification regarding the clinical and/or scientific meaning of particular 1D effects, thereby robustly structuring subsequent experimental inquiry. In short, it shifts focus from a search for significance to a search for non-rejectable hypotheses.

Keywords: biomechanics, hypothesis testing, Type II error control, random field theory, dynamic trajectories, statistical parametric mapping,

Word Counts:
Abstract: 228 (max.250)
Main Text: 1937 (max.2000)

Corresponding Author:
Todd Pataky, Ph.D., Institute for Fiber Engineering, Shinshu University
Tokida 3-15-1, Ueda, Nagano, Japan 386-8567
tpataky@shinshu-u.ac.jp, T.+81-268-21-5609, F.+81-268–21-5318

1. Introduction

Statistical power is the probability that an experiment with a specific sample size will detect a true effect see (Appendix A for terminology) given experimental noise or variance. The main goal of power analysis is to calculate the sample size required to detect a meaningful effect. *A priori* sample size calculation based on appropriate power levels (conventionally > 0.8) represents a strong position for subsequent inferences. When not calculated a variety of inferential errors can result (Hopkins Batterham, 2016; Knudson, 2017).

The secondary goals of *a priori* sample size calculation are to avoid over- and under-powering. Over-powered experiments have too many observations and consequently risk detecting meaningless, possibly small effects. Under-powered experiments conversely have too few observations to convincingly distinguish true effects from noise. In Biomechanics both over- and under-powered experiments are widely prevalent (Knudson, 2017).

Power analysis' critical step is defining a meaningful effect. Biomechanical effects can be derived in a variety of ways: (1) analytically based on mathematical hypotheses (Flash Hogan, 1985; Todorov, 2004), (2) numerically based on optimizations of musculoskeletal models, (3) referentially based on results reported in the literature, and (4) empirically based on pilot results. Nevertheless, very few studies in the Biomechanics literature employ *a priori* sample size calculations (Knudson, 2017).

In our estimation a key problem is data dimensionality. Many experiments measure one-dimensional (1D) processes (e.g. kinematic and force trajectories), yet most power analysis software requires zero-dimensional (0D) data (e.g. body mass, height). This is problematic because it adds complexity to power analysis; not only does one have to justify a particular biomechanical effect, but one also has to justify why a particular 0D metric robustly summarizes one's 1D dataset (Pataky et al., 2013).Nevertheless, in other literatures a variety of open-source tools exist for dealing with 1D and higher-dimensional effects (Appendix B).

The main purpose of this study was to describe sample size calculations for hypothesis-driven Biomechanics experiments involving 1D data. In particular we demonstrate how a numerical method called '**power1d**' (Pataky, 2017) can be used along with theory- and pilot-based 1D effects

to calculate sample sizes in single- and multi-subject experiments.

2. Methods

Analyses were conducted in Python 2.7 using Canopy 1.4 (Enthought Inc., Austin, USA), **spm1d** (Pataky, 2012) and **power1d** (Pataky, 2017).

2.1. Example 1: Theory-driven power analysis

The minimum jerk hypothesis (Flash Hogan, 1985) was tested for planar reaching in a single subject. This hypothesis predicts that, when moving from one point to another in a specific time interval, one follows the kinematic trajectory which minimizes total jerk, where jerk is the third temporal derivative of position in the intended movement direction.

2.1.1. Pilot experiment

One graduate student, naive to the intent of the study, provided informed consent to participate following ethical approval at Shinshu University. The subject sat unconstrained with their chest squarely touching a table. A board (50 cm long, 5 cm wide) was placed on the table at shoulder height and perpendicular to the subject. Two targets were marked on the board 30 cm apart using circular stickers (diameter=4 cm), corresponding to transverse plane elbow angles of approximately 150 deg and 45 deg. The subject was instructed to start with a fully-extended index finger above the proximal target, to move "rapidly but comfortably" to the distal target without touching the board, and to keep their elbow at shoulder height. Ten consecutive trials were performed with 30 s rests between trials.

Two rigid plates containing four reflective markers each were affixed to the subject's forearm and upper arm and two additional reflective markers were attached above the humeral epicondyles. Marker positions were recorded at 100 Hz using four Oqus 5 cameras (Qualisys AB, Göteborg, Sweden). Planar elbow angles were computed using Visual3D v6 (C-Motion, Germantown, MD, USA). Movement start and stop were determined using an angular velocity threshold of 5 deg/s. Data were linearly interpolated to a temporal domain of 0 – 100%. Unfiltered data were analyzed.

2.1.2. Power analysis

First, the pilot data (Fig.1a) were used to compute the temporal smoothness, or 'full-width-at-half-maximum' (FWHM) (Kiebel et al., 1999; Pataky et al., 2013), of the 1D residuals (Fig.1b); here FWHM = 26.1%. Next, noise amplitude was estimated as the average standard deviation (SD) value of 6.2 deg (Fig.1c). The minimum jerk trajectory between the experimentally measured mean starting and ending elbow angles, SD and FWHM formed a **power1d** model for generating random data samples (Fig.1d,e).

Third, we defined kinematics representing an alternative hypothesis (H1) (Fig.1f) which, if true, would cause us to reject the null (minimum jerk) hypothesis (H0). We chose alternative kinematics with a maximum difference of 20 deg at the instant of maximum velocity (time=50%) and whose difference tapered exponentially to zero at the endpoints. H1 meaning is addressed in §4.1.

Last, we generated 10,000 random data samples over a range of sample sizes J and calculated the power for each, where power is defined as the probability that the experiment will reject H0 if H1 is true; the 1D rejection criterion is described extensively elsewhere (Friston et al., 1996; Pataky et al., 2013). Following convention we computed the minimum J required to achieve power > 0.8 at α=0.05.

2.1.3. Hypothesis testing experiments

The subject repeated the experiment but was asked to move at half the previous speed. A one-sample t test was conducted (α=0.05) to compare the mean trajectory to the hypothesized minimum jerk datum. The first experiment was then repeated for a second graduate student, after receiving their informed consent, to judge the relevance of the original power analysis to a novel subject.

2.2. Example 2: Pilot-driven power analysis

A previous dataset (Neptune et al., 1999) was revisited. Ten subjects had lower limb kinematics measurements taken during v-cut and side-shuffle manoeuvres. We focussed on sagittal plane knee kinematics.

Based on observed trajectories (Fig.2a) we constructed two data sample models (Fig.2b-c).

The first, representing H0, used the observed v-cut mean trajectory as the datum for both samples, and noise was modeled as smooth Gaussian with the same average amplitude and FWHM as the original data (Fig.2b). The second, representing H1, added a Gaussian pulse to produce a different datum for the side-shuffle task (Fig.2c), similar to the observed mean (Fig.2a). As above we used **power1d** to calculate the minimum J required to achieve power > 0.8.

3. Results

3.1. Example 1: Theory-driven power analysis

Power analyses (Fig.3) suggested that a sample size of J=5 was needed to reject the minimum jerk hypothesis (H0) with a power of 0.8 (Fig.3c). In the slow reaching experiment H0 was rejected as predicted for J=5 (Fig.4i) and not for smaller sample sizes (Fig.4g,h), where the chronologically first J trials were analyzed in each case. When J=4 the maximum mean difference was approximately 25 deg (Fig.4e) which was slightly larger than the *a priori* 20 deg and thus yielded a near-significant result (Fig.4h).

For the second subject, H0 was not rejected for J=5 (Fig.5g) but was rejected for both J=6 and J=7 (Fig.5h,i). Note that this subject's mean difference from the minimum jerk datum was only 10 deg (Fig.5d–f), or half of H1's 20 deg signal. Nevertheless, Subject 2's variability was also approximately half as large as Subject 1's, implying that the overall effect size was similar. Thus, while H0 was rejected for both J=6 and J=7, both results are over-powered and thus invalid based on our *a priori* criterion of 20 deg.

3.2. Example 2: Pilot-driven power analysis

For the ten-subject knee kinematics dataset a sample size of J=10 yielded power=0.8 (Fig.6). The point-of-interest (POI) and center-of-interest (COI) power continua indicate that maximum power occurred at time=50%, close to the instant of maximum knee flexion, but with smaller-than omnibus powers.

4. Discussion

4.1. Main implications

This paper has shown that traditional power analyses can be conducted for one-dimensional (1D) biomechanical data. Single-subject results suggest that relatively small samples of approximately five may be adequately powerful to detect 1D effects (Fig.3) but also that those sample sizes can greatly over-power single-subject analysis when subject-specific variability is incorrectly modeled (Fig.5). Multi-subject results similarly showed that sample sizes of approximately ten may be suitable to detect population-level effects (Fig.6). However, these dataset- and model-specific conclusions do not necessarily generalize to other experiments. For example, smaller effects than those investigated herein may require much larger sample sizes to obtain similar powers. The only general conclusion we make is that robust power analysis is possible for 1D effects.

In §2.1.2 we chose one model based on the average SD and one specific alternative hypothesis. This would naturally lead one to question why we used those and not others. This question, in fact, emphasizes this paper's main message. Explicit *a priori* justification of the clinical or scientific meaning of alternative effects represents hypothesis-driven research and is scientifically stronger than exploratory research, which attempts to infer meaning in a *post hoc* manner. Exploratory results are of course often necessary as preliminary probes of complex biomechanical systems, but exploratory findings themselves may be scientifically meaningless (Knudson, 2017) because they are derived from a null prediction with no rationale for what effect sizes should cause rejection. *A priori* non-null hypotheses and corresponding power analyses would benefit the literature by shifting focus from a search for significance to a search for experimentally irrefutable hypotheses.

An important strength of the Biomechanics literature is its ability to predict complex 1D trajectories through theory like minimum jerk (Flash Hogan, 1985) and optimal control (Todorov, 2004), or through musculoskeletal model optimization (Pizzolato et al., 2016). Integrating realistic 1D variability with these theoretical and numerical approaches creates explicit *a priori* 1D effects, and coupling those effects with 1D power analyses represents a highly generalizable framework for robust bio- and neuro-mechanical hypothesis testing.

4.2. Power misunderstandings

From the perspective of hypothesis-driven research, a common problem is that power and effect sizes are often assessed in a *post hoc* sense (Knudson, 2017). *Post hoc* power analysis is in fact invalid (Hoenig Heisey, 2001) mainly because both power and α pertain only to *a priori* perspectives regarding the infinite set of all possible experiments (see Appendix Cfor an extended discussion). Just as α cannot logically be computed from a given dataset (i.e. what is the smallest α that would yield significance for the observed effect), neither can power.

A second common misunderstanding is that significance can be obtained simply by increasing sample size. This interpretation is incorrect because it describes over-powering (Hopkins Batterham, 2016), or equivalently employing more subjects or trials than are necessary to elucidate a specific hypothesized effect. *A priori* sample size analysis balances over- and under-powering, conventionally with a target power of 0.8.

4.3. Limitations

This paper does not address the issue of how to derive clinically or biomechanically meaningful effects. We employed relatively large signals (Figs.1f, 2c), but smaller signals may be meaningful. Nevertheless, this paper shows that it is possible to robustly test clearly defined 1D effects, and thereby encourages an *a priori* perspective on meaning, which is scientifically much more robust than attempting to infer meaning from observed signals which, by definition, may be random. Shifting to *a priori* justifications of meaning would transform analyses from exploratory to hypothesis-driven.

The proposed approach is limited to classical hypothesis testing so is largely irrelevant to machine learning studies, including those involving dimensionality reduction techniques like principal components analysis. These techniques can not only yield valuable insights into biomechanical systems but are often essential for engineering applications like neural prostheses control. Nevertheless, since many studies in Biomechanics continue to employ classical hypothesis testing (Knudson, 2017) we expect the proposed framework to remain relevant for some time. In particular, *a priori* power analysis will reduce the prevalence of false positives (Pataky et al., 2016; Knudson, 2017).

4.4. Summary

This paper aligns traditional power analysis procedures with 1D biomechanical data. The proposed framework robustly supports hypothesis-driven research for experiments involving 1D data. Its main scientific advantages are that it encourages *a priori* justifications of meaning, and consequently that it promotes a shift in focus from a search for significance to a search for non-rejectable hypotheses.

Acknowledgments

This work was supported by Wakate A Grant 15H05360 from the Japan Society for the Promotion of Science.

Conflict of Interest

The authors report no conflict of interest, financial or otherwise.

References

Flash, T., Hogan, N., 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience 5*, 1688–1703.

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

Hoenig, J. M., Heisey, D. M., 2001. The abuse of power. *The American Statistician 55*, 19–24.

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

Kiebel, S., Poline, J., Friston, K., Holmes, A., Worsley, K., 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage 10*, 756–766.

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

Neptune, R. R., Wright, I. C., van den Bogert, A. J., 1999. Muscle coordination and function during cutting movements. *Medicine & Science in Sports & Exercise 31*, 294–302.

Pataky, T. C., 2012. One-dimensional statistical parametric mapping in Python. *Computer Methods in Biomechanics and Biomedical Engineering 15*, 295–301.

Pataky, T. C., 2017. power1d: Numerical power estimates for one-dimensional continua in Python. *Journal of Statistical Software 3*, e125. doi:10.7717/peerj-cs.125.

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

Pataky, T. C., Vanrenterghem, J., Robinson, M. A., 2016. The probability of false positives in zero-dimensional analyses of one-dimensional kinematic, force and EMG trajectories. *Journal of Biomechanics 49*, 1468–1476.

Pizzolato, C., Lloyd, D. G., Sartori, M., Ceseracciu, E., Besier, T. F., Fregly, B. J., Reggiani, M., 2016. CEINMS: A toolbox to investigate the influence of different neural control solutions on the prediction of muscle excitation and joint moments during dynamic motor tasks. *Journal of Biomechanics* (pp. 1–35).

Todorov, E., 2004. Optimality principles in sensorimotor control. *Nature Neuroscience 7*, 907–915.

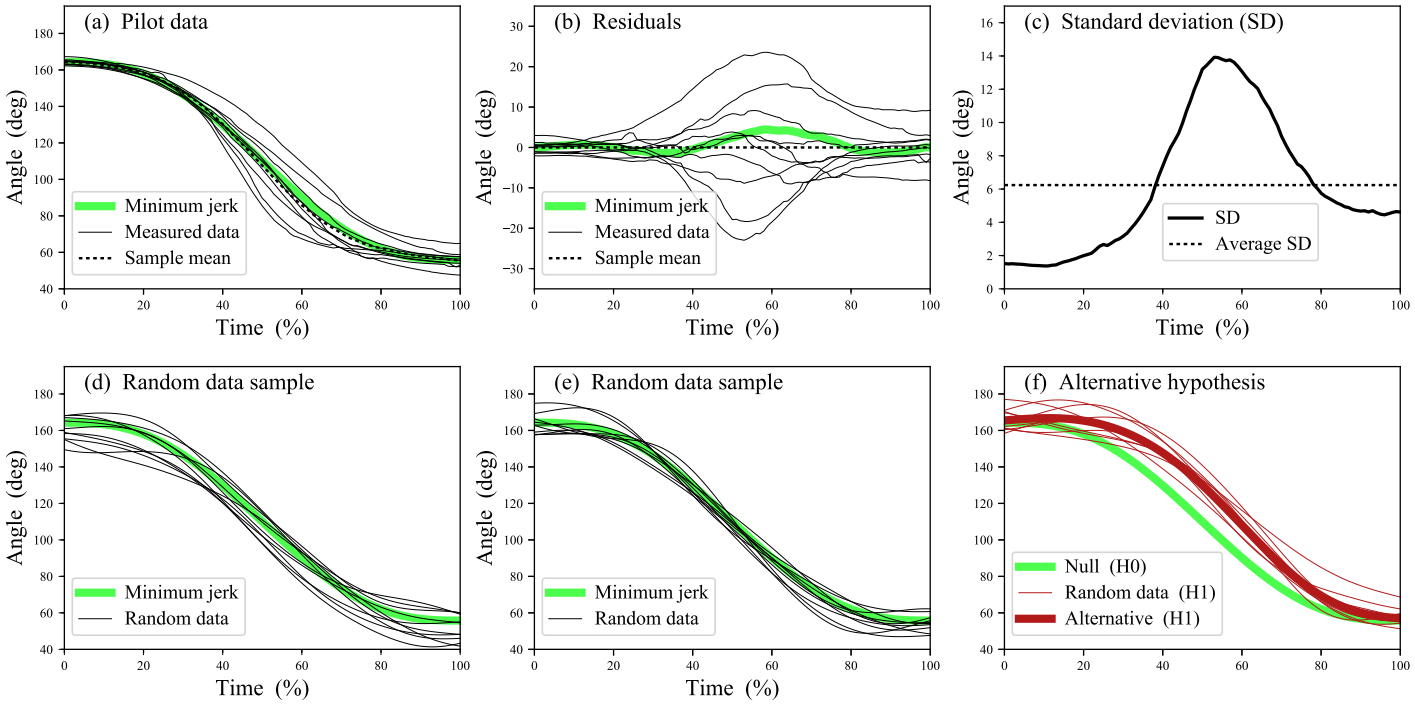

Figure 1: Example 1, components needed for hypothesis-driven power analysis. (a) Pilot data (elbow flexion angle, one subject, ten trials) and minimum jerk trajectory. (b) Residual trajectories (i.e., difference from sample mean). (c) Standard deviation (SD) trajectory with average SD depicted. (d-e) Data samples generated randomly based on residual smoothness (FWHM=26.1) and average SD (6.2). (f) An alternative hypothesis (H1) depicted with null hypothesis (H0) minimum jerk trajectory; the goal of power analysis is to calculate the sample size needed to reject H0 if H1 is true.

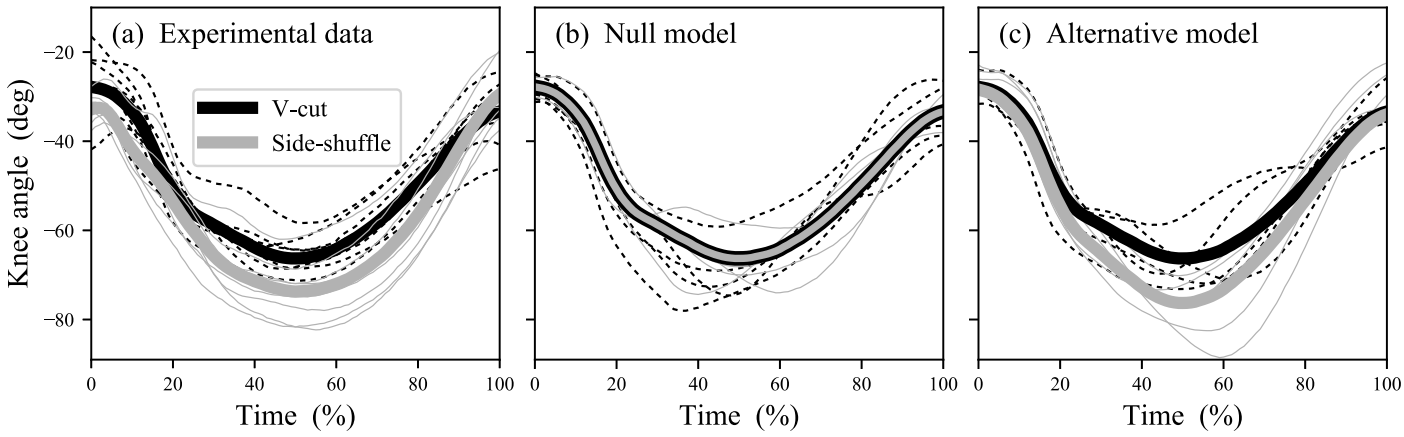

Figure 2: Example 2. (a) Multi-subject 'pilot' experimental data (Neptune et al., 1999); means and individual subjects depicted as thick and thin lines, respectively. (b) Null hypothesis model; individual subject trajectories are randomly generated as smooth Gaussian trajectories which vary about the corresponding hypothesized mean trajectory. (c) Alternative hypothesis model. Power analysis consists of: (i) using the null model to calculate the t threshold that the Gaussian trajectories would reach with a probability of (1 - alpha), then (ii) calculating the probability that the alternative model will cross that threshold.

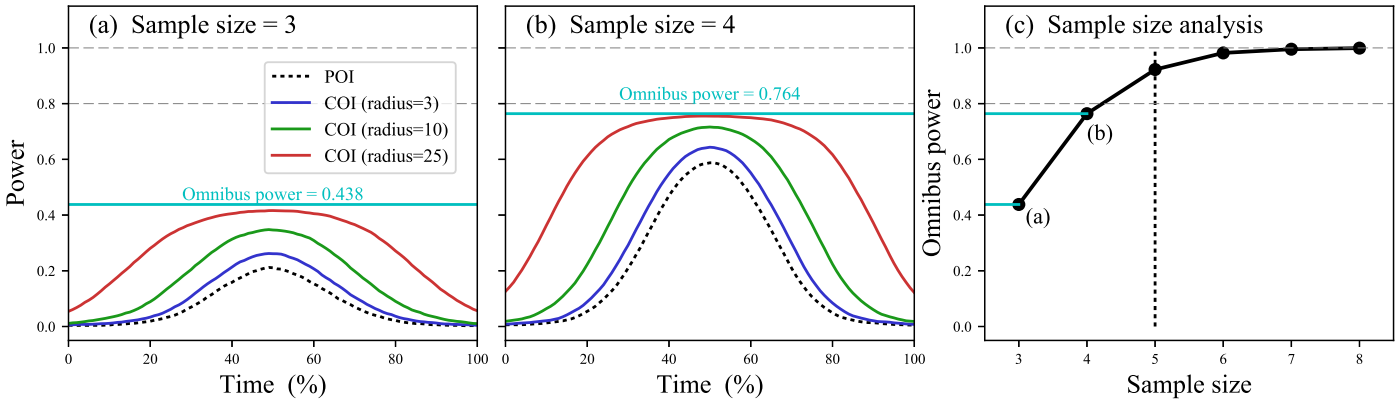

Figure 3: Example 1, power results. The omnibus (whole-continuum) power is depicted along with point-of-interest (POI) and center-of-interest (COI) power continua. Increasing the sample size from three in panel (a) to four in panel (b) results in an omnibus power increase from 0.438 to 0.764, as summarized in the first two values of panel (c). The panel (c) results suggest that a sample size of five is required to reject the null hypothesis with a power of 0.8.

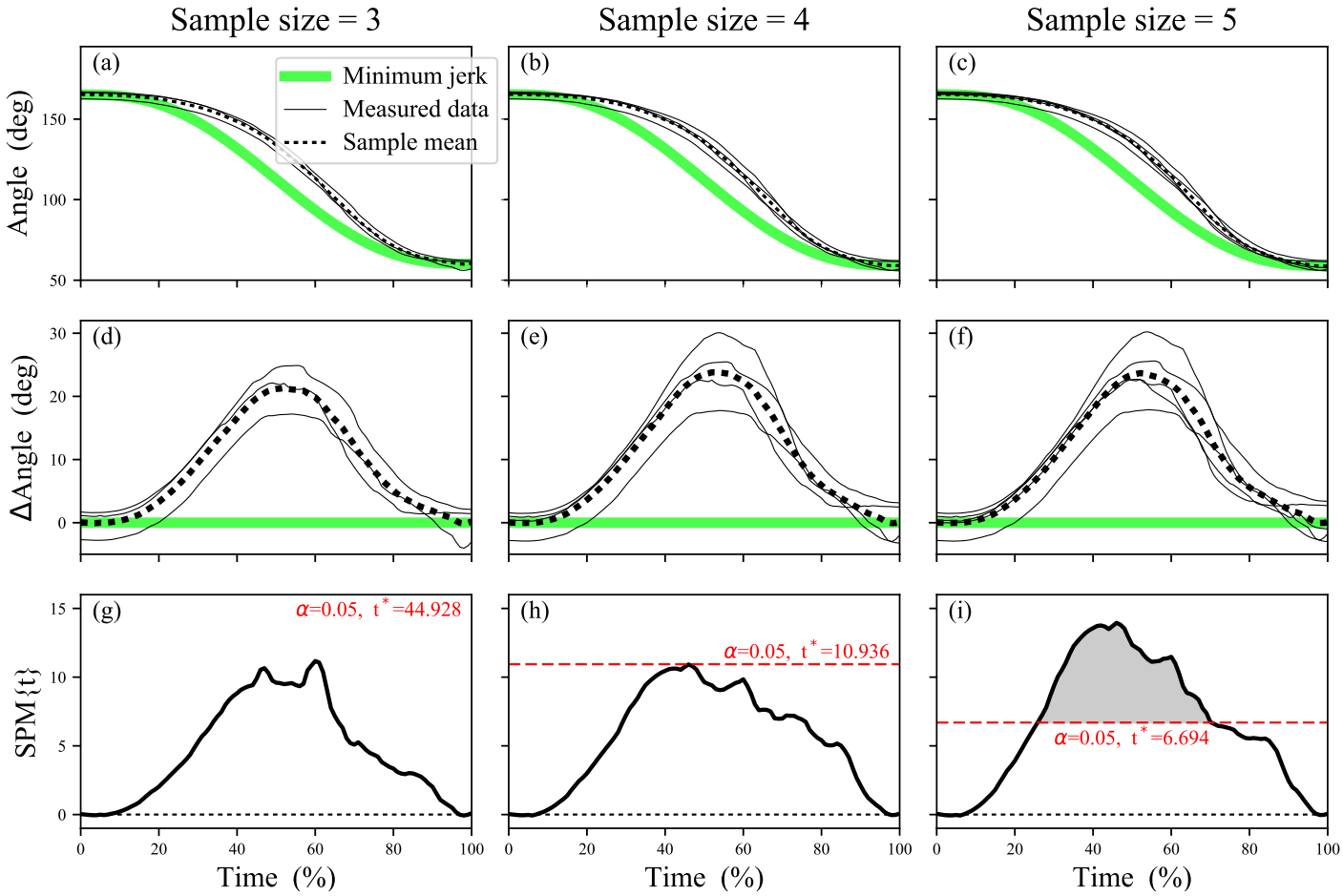

Figure 4: Example 1, main experiment hypothesis testing (Subject 1). Top panels: datasets, means and minimum jerk trajectories for three sample sizes. Middle panels: same data as in top panels but with respect to the minimum jerk trajectory (i.e. the null hypothesis). Bottom panels: hypothesis testing results; SPM{t} = statistical parametric map (t statistic); t* = critical threshold. *A priori* sample size analysis suggested sample size = 5 for power = 0.8 (Fig.3) but here three different sample sizes are presented to illustrate effects on final statistical results.

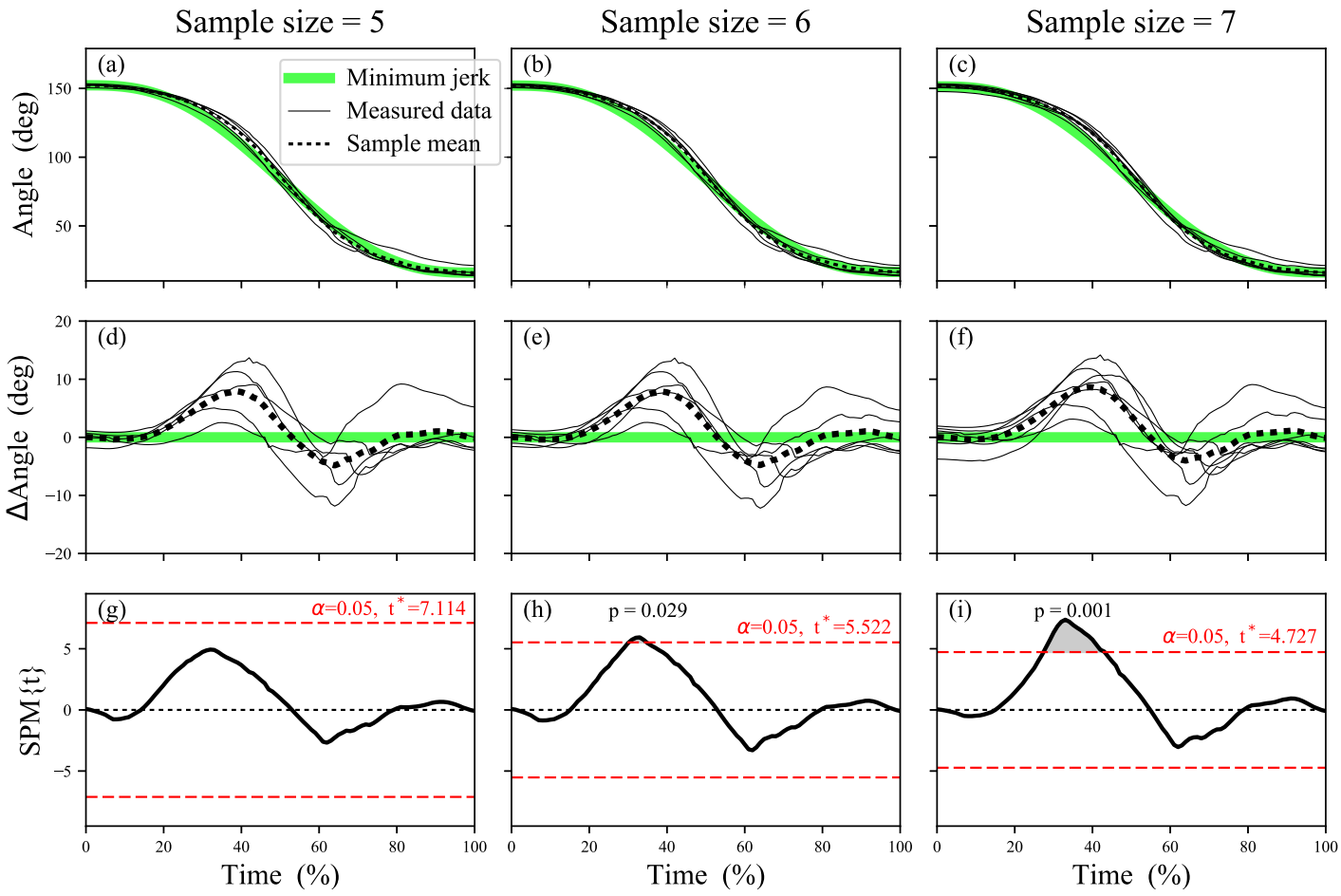

Figure 5: Example 1, second hypothesis testing experiment result (Subject 2). Data presentation follows Fig.4.

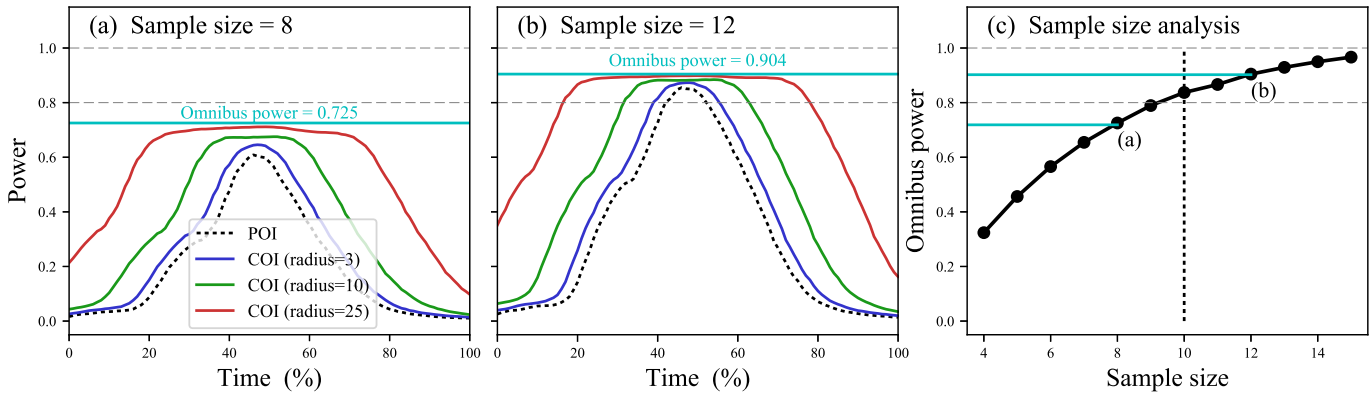

Figure 6: Example 2, power and sample size analysis. Data presented as in Fig.3. Results suggest that a sample size of 10 is required to reject the null (Fig.5b) with a probability of 0.8 when the alternative (Fig.5c) is in fact true.

Appendix A Terminology

Table A.1 and Fig.A.1 below summarize the power-relevant terminology used in the main manuscript. A brief discussion follows.

Table A.1: Power-related terminology.

Category	Term	Description
Data dimensionality	Univariate data	Scalar data
	Multiivariate data	Vector data
Measurement domain dimensionality	0D data	Univariate or multivariate data; the type of data analyzed by most commercial statistical software
	1D data	Univariate or multivariate data which vary continuously over a continuous one-dimensional (1D) domain, usually space or time
Data model components	Signal	Expected mean deviation from a datum (Fig.A.1). Units: same as observation.
	Noise	Residuals about a datum (Fig.A.1). Units: same as observation.
	Effect	Noise-normalized signal (Fig.A.1). Unitless.
Hypothesis testing	Null hypothesis (H0)	A testable statement of equality (e.g. Force change = 0)
	Alternative hypothesis (H1)	A non-testable statement of inequality (e.g. Force change > 0)
Probability	Type I error (α)	The probability of rejecting H0 when it is true
	Type II error (β)	The probability of failing to reject H0 when it is false
	Power (1 - β)	The probability of rejecting H0 given an alternative true effect
1D power types	Omnibus power	The probability of rejecting H0 at any point in a 1D continuum; i.e. in unconstrained 1D hypothesis testing
	Center of interest (COI) power	The probability of rejecting H0 within a given radius about a specific point in a 1D continuum; i.e. in regionally-constrained 1D hypothesis testing.
	Point of interest (POI) power	The probability of rejecting H0 at a specific point in a 1D continuum; i.e. in point-constrained 1D hypothesis testing. Equivalent to 0D power when a single continuum point is selected in an *a priori* manner for hypothesis testing.

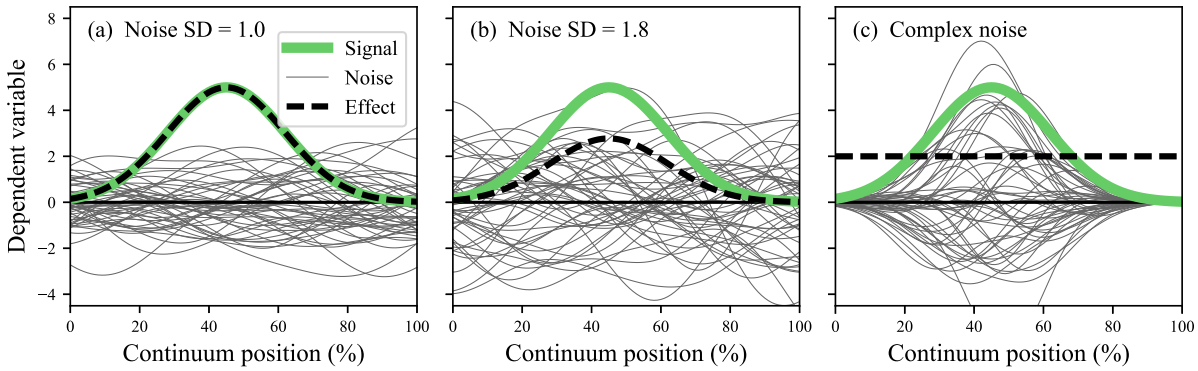

Figure A.1: Power-relevant terminology: "signal", "noise" and "effect". The goal of power analysis is to determine the probability of detecting a signal given the underlying noise. That probability is related most closely to the effect size, or signal-to-noise ratio. (a) An arbitrary signal and smooth Gaussian noise with a constant standard deviation (SD) of 1.0. (b) Same as (a) but with a constant SD of 1.8; since the noise is greater and also constant across the continuum the effect is uniformly compressed. (c) Same as (a) but with irregular, complex noise; in this particular case the noise amplitude scales with the signal so the effect is constant across the 1D continuum.

Discussion

Standard power theory and most commercial software packages deal exclusively with 0D data. Their relations to 1D data can be elusive because it is usually difficult to justifiably reduce a 1D observation to a specific 0D metric in an *a priori* manner (Pataky et al., 2013). Moreover, power-relevant terminology is not readily translatable to 1D data. For 0D data the 'effect' is simply the signal-to-noise ratio, but for 1D data the relation amongst signal, noise and effect can be more complex because both signal and noise magnitudes can change across the 1D domain (Fig.A.1), and also because noise can be signal-dependent, implying that a constant effect size implies neither constant signal nor constant noise. Furthermore, adjacent values in time are often highly correlated due to the viscoelastic nature of biological tissues, implying that 0D approaches, which neglect this temporal correlation, are generally not valid for smooth 1D data analysis.

There are three separate types of power that exist in 1D power analysis: point-of interest (POI), center-of-interest (COI) and omnibus (Figs.3,6, main manuscript) (Pataky, 2017). The POI power continuum shows the probability of rejecting the null hypothesis at each continuum point. The COI continuum shows power increases associated with extending the hypothesis testing scope to adjacent continuum nodes. The omnibus power represents the power of rejecting H0 anywhere in the continuum. By definition, POI power <= COI power <= omnibus power. COI powers are equivalent to POI and omnibus powers when the COI radii are zero and 100%, respectively.

Appendix B Existing power analysis methods

While hypothesis testing emerged in the literature nearly a century ago (Fisher, 1925), statistical power was not formally discussed until the late 1980s (reviewed in Huberty, 1993). Power analysis' relatively recent emergence has led to a number of misunderstandings and misuses (Hoenig and Heisey, 2001), an under-appreciation of its value (Hopkins and Batterham, 2016) and *ad hoc* effect definitions (Knudson, 2017).

Continuum-level approaches to statistical power first emerged in the 1990s (Friston et al., 1996) and currently three main categories of continuum-level power analysis exist (Fig.B.1): inflated variance, non-central random field theory (RFT), and numerical.

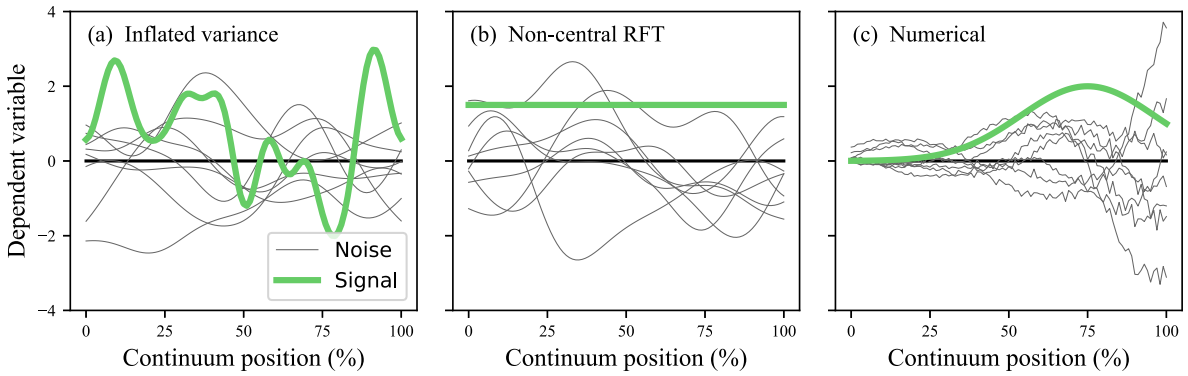

Figure B.1: Overview of existing continuum-level power analysis methods.

The inflated variance method (Friston et al., 1996) is based on (central) random field theory (RFT) (Adler and Hasofer, 1976; Adler and Taylor, 2007). RFT describes the probabilistic behavior of smooth Gaussian continua (see Fig.B.1a,b) and has been shown to accurately model the variance observed in a variety of 1D biomechanical datasets (Pataky et al., 2013). For power analysis this approach defines signal as a second Gaussian process with larger variance and potentially different smoothness (i.e. different frequency content), thereby representing a hypothesis for which the continuum location(s) and precise effect amplitudes are *a priori* unknown. In the context of joint angle trajectories, for example, this inflated variance signal implies that the approximate magnitude of a particular change is known (e.g. 10 deg) but not its location or extent in time. This interpretation of signal is useful for purely exploratory analysis in which precise effect predictions are not made. However, the method itself has

low power because the signal prediction is imprecise Hayasaka et al. (2007). In other words, if there is a true signal which systematically occurs at the same point in time across subjects and/or trials, then the inflated variance method will predict far more subjects and/or trials than are actually needed to detect that signal.

The second approach to continuum-level power analysis is the non-central RFT method (Hayasaka et al., 2007; Mumford and Nichols, 2008). This method models signal as a constant shift (Fig.B.1b), possibly isolated to specific continuum regions. This signal represents, for example, a constant change of 5 deg from a reference joint angle trajectory. This type of signal is perfectly analogous to the classical definition of power, and since the signal is precise the non-central RFT method is more powerful than the inflated variance method (Hayasaka et al., 2007). Neverthless, its main limitation is that it defines signal in a binary sense: a continuum region either possesses constant signal or none. This is not very useful for biomechanics applications in which precise kinematic and dynamic trajectories can be predicted based on theory or musculoskeletal model optimization.

The third approach is a numerical method which iteratively simulates random 1D continua to compute power (Pataky, 2017). This method, while computationally intensive, overcomes the limitations of both aforementioned methods because it affords arbitrary modeling of both signal and noise (Fig.B.1c). This allows an investigator to generate a specific 1D signal of interest, to create 1D noise models to mimic theoretical or experimentally observed variance, and ultimately to compute the sample size required to robustly test arbitrary 1D signal predictions in their arbitrary noise environments. The numerical approach is the most general approach, flexibly implementing either of the two aforementioned approaches or arbitrary signal and noise (Pataky, 2017). Since it is most general the main manuscript uses only the numerical method.

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.

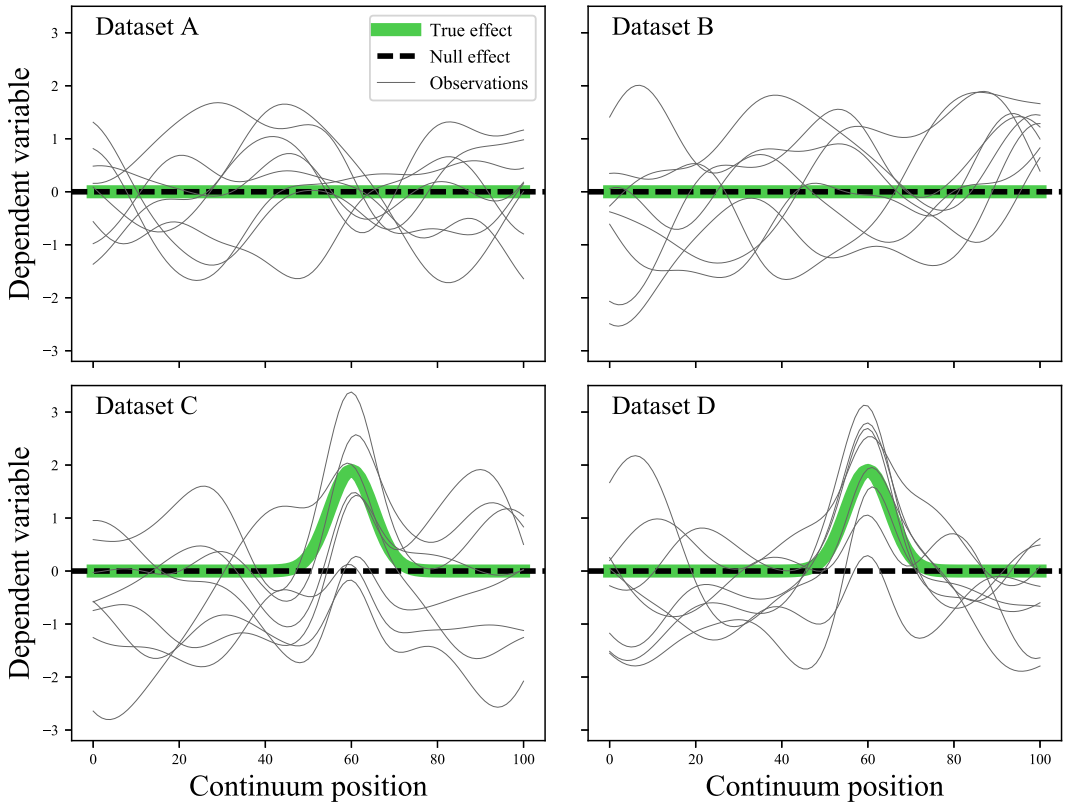

Figure C.1: Simulated datasets.

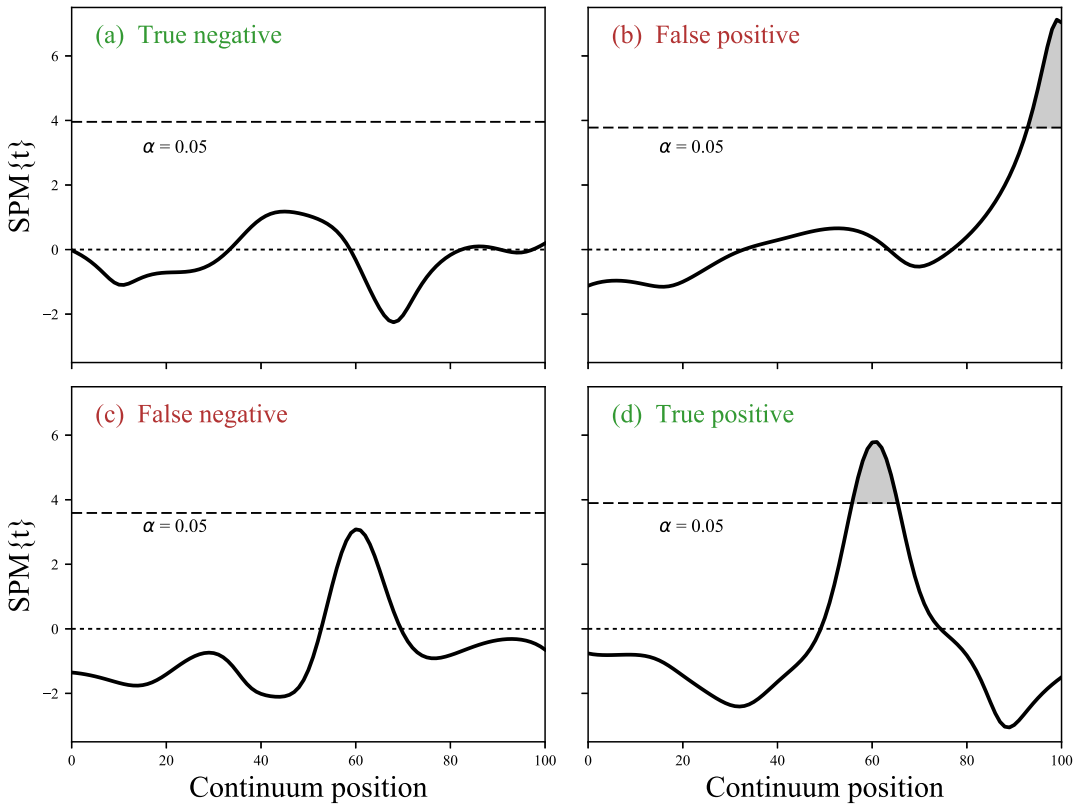

Figure C.2: One-sample t test results. The solid black line depicts the test statistic (t value) continuum. The horizontal line depicts the critical threshold at a Type I error rate of α=0.05.

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.