Data simulation: Theory

Why simulate your experiment

1. Refine your analysis protocol/scripts so that you have a better idea of what you are doing. This protects against nasty surprises when it comes to analysing data for real and may mitigate any need to re-run the experiment in case of catastrophic confound/failure. In particular, simulations can help you discover if your analysis:
    a. Does not tell you what you want to know. This can happen if your test is either insensitive to the effect of interest or returns a significant result due to trivial effects. Both of these can occur when statistical assumptions are violated. For instance, time-series correlations can be significant if both the predictor and response variables are non-stationary.
    b. Is impossible to run because it violates some fundamental statistical assumption (e.g. no perfect multicollinearity in predictors = fixed effects design matrix is not of full rank).
2. Help develop hypotheses and statistical tests for pre-registrations and registered reports. In both of these cases, you will be bound to what you said you were going to do and so need to be 100% sure everything will all work the way you expect.
3. Power analyses for statistical tests when power cannot be computed directly (i.e. any covered by GPower):
    a. Linear mixed effects models (where power depends on the number and convenience of random effects).
    b. Uncontrollable covariance between observations that act as predictors (e.g. in a quasi-experiment).
    c. Generalised linear fixed-/mixed effects models where you have something more complex than an identity link function.
4. Planing the analyses ahead of time saves you stress (because you know what you are doing all along) and is very satisfying (as you can see your simulated predictions come to life in the real world).

Simulating data requires a model… What is a model

Every parametric model has two components...

1. **Parametric equations** that describe how the DV is expected to change in response to the IV.
2. **Probability distributions**: ‘Shapes’ that describe the spread of data points around what is expected given the parametric equation. **Note**: this means that we explicitly acknowledge that our parametric equations will never perfectly reproduce real-world observations. Instead, the distributions allow our models to account for a range of observed responses around the expectation. This spread can be interpreted in many ways but two of the most common interpretations are:
    a. Noise or error in measurements.
    b. Random effects that reflect real-world fluctuations from some group average but are not explicitly accounted for by the model.

Remember: “All models are wrong, but some are useful.” *George Box*. So don’t be too hard on yourself when running a simulation.

Choosing a parametric equation

How do you express the expected relationship between variables of interest when simulating data? In choosing a parameterisation you should consider what you think happens in the real world and the range of values that the IV and DV can take (known as the ***domain*** and ***image*** respectively). Pick the parameters of each equation based on previous research or your own intuition. It is also possible to vary them systematically to simulate a range of possibilities:

1. ***Mean***: $y = \beta_0 + \varepsilon$
    a. Domain: [-Inf,Inf]
    b. Image: [-Inf,Inf]
    c. Example: Modelling the mean difference between two conditions in a repeated measures design.
2. ***Linear***: $y = \beta_0 + \beta_1 x_1 + \varepsilon$
    a. Domain: [-Inf,Inf]
    b. Image: [-Inf,Inf]
    c. Example: Modelling a standard correlation between two continuous predictors (for instance, hight and weight).
3. ***Polynomial***: $y = \beta_0 + \beta_{1,1} x_1 + \beta_{1,2} x_1^2 + \varepsilon$
    a. Domain: [-Inf,Inf]
    b. Image: [-Inf,Inf]
    c. Example: Modelling the inverted-U relationship between arousal and performance… If you are too aroused, your performance will be low :(.
4. ***Exponential***: $y = c + m \cdot exp(\beta_0 + \beta_1 x_1) + \varepsilon \ldots \ log(y - c) - log(m) \sim \beta_0 + \beta_1 x_1$
    a. Domain: [-Inf,Inf]
    b. Image: [c,Inf]
    c. Example: Modelling memory performance over time… It will never quite hit zero and will certainly not go negative.
5. ***Sigmoid***: $y = 1 / (1 + exp(\beta_0 + \beta_1 x_1)) + \varepsilon$
    a. Domain: [-Inf,Inf]
    b. Image: [0,1] although can be scaled to fit in any range with two extra free parameters.
    c. Example: Modelling the probability of recalling a word from memory. Note: a lot of people would use a linear model for this but it would break down when performance is either high or low.
6. ***Power*** (important for log-log analyses): $y = c + (\beta_0 + \beta_1 x_1)^m + \varepsilon \ldots$
   $log(y - c) \sim m \cdot log(\beta_0 + \beta_1 x_1)$
    a. Domain: [0,Inf]
    b. Image: [c,Inf]
    c. Example: Modelling spectral power in EEG epochs across conditions or time.

NOTE: You may only plan to use a simple ANOVA model to analyse your data. However, we often chose more detailed models to generate simulated data in order to make the simulation more realistic. For instance, if my ANOVA will have five conditions (e.g. very low, low, medium, high, and very high concentrations of a drug), perhaps, I should use a polynomial equation to generate the expected responses to that drug by sampling from 5 different x values (e.g. $x_{vlo} = 0;\ x_{low} = 1;\ x_{med} = 2;\ x_{hig} = 3;\ x_{vhi} = 4;$). Feeding this into the polynomial equation: $y = 0 + 1 \cdot x_1 - 0.15 \cdot x_1^2$ gives the following values of y: $y_{vlo} = 0;\ y_{low} = 0.9;\ y_{med} = 1.6;\ y_{hig} = 2.1;\ y_{vhi} = 2.4;$. This is neater and more systematic than simply plucking 5 means out of thin air.

Choosing a probability distribution

How do you characterize the distribution of random factors around what is expected by your parametric equation? As before, this depends on what you think happens in the real world, and the range of values that the DV can take (i.e. the ***image***, also known as, the ***support*** of the distribution):

1. Normal:
    a. Support: [-Inf,Inf]
    b. Example: Problem-solving ability in the general population
2. Gamma: Positively skewed with long tails:
    a. Support: [0,Inf]
    b. Example: Useful for modelling observations that cannot go below zero: e.g. reaction times, hights of participants, location/placement error during a memory task.
3. Beta distribution:
    a. Support: [0,1]
    b. Example: Useful for modelling observations that cannot go below zero or above 1... For instance, the probability of recalling a word from memory.

NOTE: The above are all examples of commonly used continuous distributions. As such, they may only be useful when you are simulating continuous data. There is a whole other class of distributions for generating discrete data. See here for a comprehensive list.

Estimating power from simulations: The sampling distribution, sample size, and statistical inference:

1. What is the sampling distribution?: It is a distribution (almost always a normal distribution) that encodes your confidence in the range of parameter values that are plausible given your data. The narrower the sampling the distribution, the more confident you can be that your estimated parameters reflect reality.
    a. The standard deviation of the sampling distribution is called the ***standard error***. The variance of the sampling distribution is calculated as the standard error squared.

    b. We use the sampling distribution to make statistical inferences, whether they be Bayesian or Classical. The smaller the standard error, the narrower the sampling distribution and so the more power we have.
2. How is the sampling distribution is affected by sample size? Well, that what simulations are for! The variance of the sampling distribution for a parameter should decrease with the number of data points that have been used to estimate that parameter. In some cases, as the number of data points doubles, the variance of the sampling distribution halves… Things like GPower use this fact to calculate statistical power analytically.
3. Note, the sampling distribution is independent of the standard deviation within a group… While the standard deviation of a group should stay about the same as you collect more samples, the sampling distribution will narrow.
4. Regardless of what your data look like and what probability distribution you have chosen to reflect the randomness in your model, the sampling distribution is almost always normal because of the central limit theorem... As long as you have more than 30 responses contributing to the parameter estimate.
    a. This is because of the central limit theorem, see: http://www.ltcconline.net/greenl/java/Statistics/clt/cltsimulation.html
5. How do the variances/covariances of IVs affect statistical power? Again, simulations can help you understand this. Correlations between predictor variables (e.g. IVs in a multiple regression) changes the width of the sampling distribution. This can dramatically affect your ability to identify differences between the effects of each predictor.
    a. Let's say you are trying to predict a person's life expectancy from multiple predictors (including hight and weight)... If there is a strong positive correlation between a person height and their weight then your confidence in any differential effect of hight vs wight on life expectancy will be high... That is, it will be relatively easy to tell whether hight or weight is the more critical risk factor for lower life expectancy.
    b. In contrast, strong positive correlations between predictors make it difficult to determine whether the sum of hight+weight have a significant effect on life expectancy. Such a test would be used if you didn't care whether hight or weight (or both) contributed to life expectancy and so you wanted to test them together (sort of like a main-effect).This is a really terrible example so I will stop typing now.