Simulating data to improve your research: an Introduction, Sessions 1 and 2.

COURSE AIM

In recent years, there has been an increasing focus on the 'replication crisis', with evidence that much published research is not robust. There are many reasons for this: in this course, the focus will be on the importance of having a deep understanding of the basic statistical methods that are commonly used for null-hypothesis significance testing. Human cognition is not well-suited to thinking about probability, and so if researchers are simply trained to apply statistical tests, they may do so in a way that is very likely to generate non-replicable findings. A good way to gain a deep understanding of the nature and limitations of statistical methods is to simulate experimental data with known characteristics, and then apply standard statistical tests.

In session 1, the focus will be on gaining a deep understanding of what a p-value means, and in particular how easy it is to obtain 'significant' results from null data if there is a flexible approach to data analysis. In session 2, the focus moves to simulating datasets where there is a real effect, so consider how choice of experimental design and analytic approach may influence whether the effect is detected.

No prior knowledge of coding is required, and the initial exercises will use Excel, to illustrate the basic principles that are involved in data simulation. Subsequent exercises will use the R programming language. No prior knowledge of R is required, and indeed this course can act as a gentle introduction to R. However, to get benefit from the course, participants should follow along the coding exercises, and for this they will need to have R, R studio and some related packages and scripts installed: the instructions for doing can be found here.

COURSE CONTENT
At the end of session 1, participants should be able to:

- Simulate distributions of variables with known means and standard deviations, using R.
- Understand when and why it is important to apply corrections for multiple statistical tests.
- Understand the value of simulating data to check out an analysis plan prior to running an experiment

At the end of session 2, participants should be able to:

- Understand why doing research with an insufficient sample size is wasteful and can result in false acceptance of a null hypothesis
- Understand how to use simulation to do a power analysis
- Be aware of how power can be influenced by choice of experimental design and measures

Session 1 (total viewing time 114 minutes)

Simulating random data (null effects)

block	Duration	Content	R Script
1.1	18.43	Introduction: using Excel to illustrate data simulation	-
1.2	9.36	Doing a t-test on simulated data (Excel)	-
1.3	7.12	Explanation of p-hacking	-
1.4	11.00	The Garden of Forking Paths	-
1.5	12.38	Simulating data in R	Simulation_ex1_intro
1.6	17.46	Walking through the script	Simulation_ex1_intro
1.7	7.48	Repeatedly running simulation in a loop	Simulation_ex1_multioutput
1.8	14.49	Simulating correlated variables	Simulation_ex2a_correlations
1.9	13.45	Overview; different kinds of p-hacking. Different simulation packages	Forkingpaths_demo

Session 2 (total viewing time 81 minutes)

Simulating data with an estimated true effect

block	Duration	Content	R Script
2.1	8.30	Impact of N on estimates of mean	sampling_demo
2.2	7.16	How N and effect size affect p-value distribution	Simulation_ex1_multioutput
2.3	13.37	More on effect sizes, and a demonstration using *simstudy* package	simstudy_power_demo
2.4	8.16	Confronting the problem of low power	-
2.5	19.30	Comparing power of between vs within-subjects design using *faux* package	faux_demo_bw
2.6	12.04	Increasing number of observations to enhance power	simulating_items
2.7	10.17	Increasing number of observations to improve test-retest reliability	simulating_reliability

REFERENCE

Bishop, D. V. M. (2019). World View: Rein in the four horsemen of irreproducibility. Nature, 568, 435. doi:10.1038/d41586-019-01307-2