www.energy-evaluation.org
ENERGY EVALUATION EUROPE
2021 Virtual Conference
Wednesday 10 – Thursday 11 and Monday 15 – Tuesday 16 March 2021
Formerly IEPPEC
Accelerating the energy transition for all: Evaluation's role in effective policy making

# Ensuring statistics have power: sample sizes, effect sizes and confidence intervals (and how to use them)

*Ben Anderson (University of Southampton, UK. b.anderson@soton.ac.uk, @dataknut), Tom Rushby (University of Southampton, UK), Abubakr Bahaj (University of Southampton, UK), Patrick James (University of Southampton, UK)*

## EXTENDED ABSTRACT

There is ongoing confusion in empirical energy efficiency and energy demand evaluation studies over the meaning, value and use of *statistical significance* and *statistical power*. This is compounded by confusion over how these concepts should be used both in *designing* studies and in deciding what can be *inferred* from them. As a consequence, sample sizes in most energy efficiency studies may be too low to provide adequate statistical power and so statistically robust conclusions cannot be drawn at conventional thresholds[1]. In this paper we explore this problem via the design of a study focused on winter evening heat pump demand to demonstrate how sample sizes, effect sizes and confidence intervals matter.

## Introduction / background

Given these confusions, it is unsurprising that many studies report effect sizes which are not statistically significant at conventional thresholds[2], choose to use lower statistical significance thresholds or lower both statistical power values *and* statistical significance thresholds[3]. However, decisions should never be based solely on statistical significance thresholds set purely by convention[4]. Inference, and thus decision making, should be based on an assessment of the effect size; the level of uncertainty (confidence intervals) and the risk of a false positives (Type I error) or false negatives (Type II error). Only then can we decide if the effect is large enough, certain enough and has a low enough risk of being a false positive or false negative result to warrant action.

We have observed three consequences of this confusion: Firstly, a large number of energy evaluation studies have been implemented with no real idea of whether they will be able to robustly test their hypotheses under normative statistical conventions. Secondly, studies which *have* been robustly designed risk being dismissed and/or themselves dismissing potentially useful results due to a very narrow application of p-value based statistical significance testing. Finally, a lack of consistency of reporting makes comparing across studies and thus developing a synthesised and summative evidence base for strategic or public policy decision making extremely difficult.

[1] E. R. Frederiks, K. Stenner, E. V. Hobman, and M. Fischle, 'Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers', *Energy Research & Social Science*, vol. 22, pp. 147–164.
[2] A. Srivastava, S. Van Passel, and E. Laes, 'Assessing the success of electricity demand response programs: A meta-analysis', *Energy Research & Social Science*, vol. 40, pp. 110–117.
[3] B. Anderson, T. Rushby, A. Bahaj, and P. James, 'Ensuring statistics have power: Guidance for designing, reporting and acting on electricity demand reduction and behaviour change programs', *Energy Research & Social Science*, vol. 59, p. 101260.
[4] S. Greenland *et al.*, 'Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations', *Eur J Epidemiol*, vol. 31, no. 4, pp. 337–350.

## Approach

As an example, Figure 1 shows the observed mean electricity demand by heat pumps in New Zealand households of different sizes in the evening peak period in winter[5]. The sample is small (40) and as a consequence sub-sample counts for each household size are very low – there are only 3 1-person households in the sample. As we can see it would be impossible to conclude that there were any statistically significant differences between the household groups in Figure 1 under normative statistical thresholds (the 95% confidence intervals overlap).

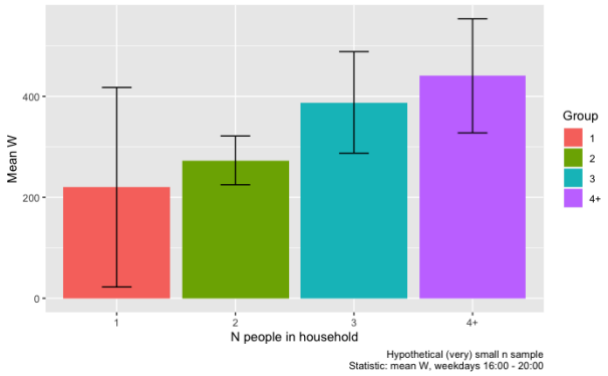


Figure 1: Mean heat pump power demand by household type (Error bars = 95% confidence intervals for the mean – 40 household sample). *Source: authors' calculations*

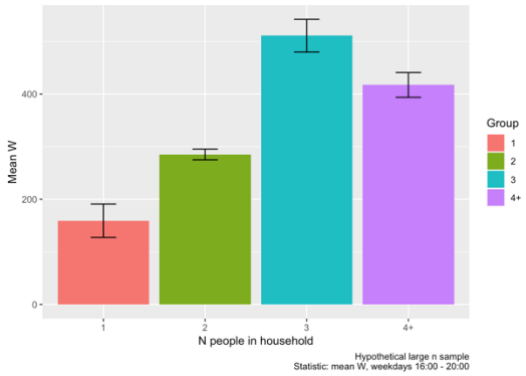


Figure 2: Mean heat pump power demand by household type (simulated 1000 household sample). *Source: authors' calculations*

In contrast Figure 2 simulates the results we would have obtained from a sample of ~ 1000 households. In Figure 2, we see that the 95% confidence intervals do not overlap (statistically significant differences) although we still have considerable uncertainty in the 1-person households. However, we could have reached broadly similar conclusions regarding heat pump energy use and household size from the small sample study reported in Figure 1 if we were comfortable with the increased risk of Type I and Type II errors it implies.

## Conclusion & discussion

Implementing a study from which we may be able to conclude something with some certainty relies on appropriate sample sizing via **statistical power analysis** to reduce the risk of a Type II error / *false negative*. Since there is no post-hoc fix, we need to conduct statistical power analysis *before we start* to make sure the study has a chance of detecting the effects it foresees. This is hardly news, but it certainly seems to be in energy studies.

Inference and subsequent decisions must then pay attention to all of: **difference or effect size** - is it 2% or 22% (i.e. is the result *important* or *useful*, "What is the estimated *bang for buck*?"); **statistical confidence intervals** - (i.e. is there *uncertainty* or *variation* in response, "How *uncertain* is the estimated bang?"); **statistical p values** - (i.e. what is the risk of a Type I error / *false positive*, "What is the risk the bang observed isn't real?"). This means that we *always* need to report all these elements because together they enable the assessment of the substantive significance of the results.

Overall, energy evaluation studies must therefore implement appropriate sample design based on statistical power analysis **and** must report nuanced analysis based on effect sizes, confidence intervals (and associated p values) and statistical power. Project managers can use this guidance to understand what can count as evidence, for what purpose and in what context. They can then more effectively manage study resources and develop a robust, contextually meaningful and *defensible* strategy for making decisions based on the results. Finally, commercial or public policy decision makers can use this guidance to help them make evidence-based and defensible commercial strategy or policy intervention decisions. In particular, it will help them to avoid focusing on results which are 'statistically significant' but small in magnitude and therefore of little *practical significance.*

[5] Data source: B. Anderson et al., 'New Zealand GREEN Grid household electricity demand study 2014-2018', Sep. 2018.