Statistical Power, Statistical Significance, Study Design and Decision Making: A Worked Example

Sizing Demand Response Trials in New Zealand

Ben Anderson and Tom Rushby (Contact: b.anderson@soton.ac.uk, `@dataknut`)

Last run at: 2018-09-18 23:42:51

Contents

1 About

1.1 Report circulation:

- Public

1.2 License

1.3 Citation

If you wish to use any of the material from this report please cite as:

- Ben Anderson and Tom Rushby. (2018) Statistical Power, Statistical Significance, Study Design and Decision Making: A Worked Example (Sizing Demand Response Trials in New Zealand), Centre for Sustainability, University of Otago: Dunedin, New Zealand.

1.4 History

Code history is generally tracked via our git.soton repo:

- Report history

1.5 Data:

This paper uses circuit level extracts for 'Heat Pumps', 'Lighting' and 'Hot Water' for the NZ GREEN Grid Household Electricity Demand Data (https://dx.doi.org/10.5255/UKDA-SN-853334 (Anderson et al. 2018)). These have been extracted using the code found in

1.6 Support

This work was supported by:

- The University of Otago;
- The University of Southampton;
- The New Zealand Ministry of Business, Innovation and Employment (MBIE) through the NZ GREEN Grid project;
- SPATIALEC - a Marie Skłodowska-Curie Global Fellowship based at the University of Otago's Centre for Sustainability (2017-2019) & the University of Southampton's Sustainable Energy Research Group (2019-2020).

We do not 'support' the code but if you notice a problem please check the issues on our repo and if it doesn't already exist, please open a new one.

2 Introduction

In our experiennce of designing and running empirical studies, whether experimental or naturalistic, there is ongoing confusion over the meaning and role of two key statistical terms:

- statistical power
- statistical significance

We have found this to be the case both in academic research where the objective is to establish 'the most likely explanation' under academic conventions and in applied research where the objective is to 'make a robust decision' based on the balance of evidence and probability.

In this brief paper we respond to these confusions using a worked example: the design of a hypothetical household electricity demand response trial in New Zealand which seeks to shift the use of Heat Pumps out of the evening winter peak demand period. We use this example to explain and demonstrate the role of statistical signficance in testing for differences and of both statistical signficance and statistical power in sample design and decision making.

3 Error, power, significance and decision making

Two types of error are of concern in both purely academic and applied research studies:

- Type I: a false positive - an effect is inferred when in fact there is none. From a commercial or policy perspective this could lead to the implementation of a costly intervention which would be unlikely to have the effect expected;
- Type II: a false negative - an effect is not inferred when in fact there is one. From a commercial or policy perspective this could lead to inaction when an intervention would have been likely to have the effect expected.

The significance level (p value) of the statistical test to be used represents the extent to which the observed data matches the null model to be tested (Wasserstein and Lazar 2016). In most trials the null model will be a measure of 'no difference' between control and intervenion groups. By convention, the p value *threshold* for rejecting the null model (the risk of a Type I error) is generally set to 0.05 (5%) although this choice is entirely subjective. In commercial or policy terms an action taken on a larger p value (e.g. setting the p value threshold to 10%) would increase the risk of making a Type I error and thus implementing a potentially costly intervention that is unlikely to have the effect desired. However, as we discuss in more detail below, this is not necessarily *bad practice* as it may reflect the potential magnitude of an effect, the decision-maker's tolerance of Type I error risk and the urgency of action.

Statistical power is normally set to 0.8 (80%) by convention and represents the pre-study risk of making a Type II error (Greenland et al. 2016). From a commercial or policy perspective reducing power (e.g. to 0.7 or 70%) will therefore increase the risk of taking no action when in fact the intervention would probably have had the effect desired. Statistical power calculations enable the investigator to estimate the sample size that would be needed to robustly detect an experimental effect with a given risk of a false positive (Type I error) or false negative (Type II error) result. This prevents a study from recruiting too few participants to be able to robustly detect the hypothesised intervention effect (Delmas, Fischlein, and Asensio 2013) or wasting resources by recruiting a larger sample than needed.

Previous work has suggested that sample sizes in most energy efficiency studies may be too low to provide adequate power and so statistically robust conclusions cannot be drawn at conventional thresholds (Frederiks et al. 2016) while a more recent review focusing on demand response studies reaching a similar conclusion (Srivastava, Van Passel, and Laes 2018). It is therefore hardly surprising that a number of studies report effect sizes which are not statistically significant at conventional thresholds (Srivastava, Van Passel, and Laes 2018), choose to use lower statistical significance thresholds (Institute 2006, AECOM (2011), CER (2012), Schofield et al. (2015)) or both lower statistical power values *and* statistical significance thresholds (UKPN 2017,UKPN (2018)).

However it would be wrong to conclude that this is *necessarily* bad practice. Recent discussions of the role of p values in inference (Greenland et al. 2016, Wasserstein and Lazar (2016)) should remind us that decisions should never be based only on statistical significance thresholds set purely by convention. Rather, inference and thus decision making should be based on:

- statistic effect size - is it 2% or 22% (i.e. is the result *important* or *useful*, "What is the estimated *bang for buck*?");
- statistic confidence intervals - (i.e. is there *uncertainty* or *variation* in response, "How uncertain is the estimated bang?");
- statistic 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?");

Only then can a contextually appropriate decision be taken as to whether the effect is large enough, certain enough and has a low enough risk of being a false positive result to warrant action.

In the following sections we apply these principles to the design and analysis of a hypothetical New Zealand household electricity demand response trial and to the use of a simple statistical test of difference between two groups.

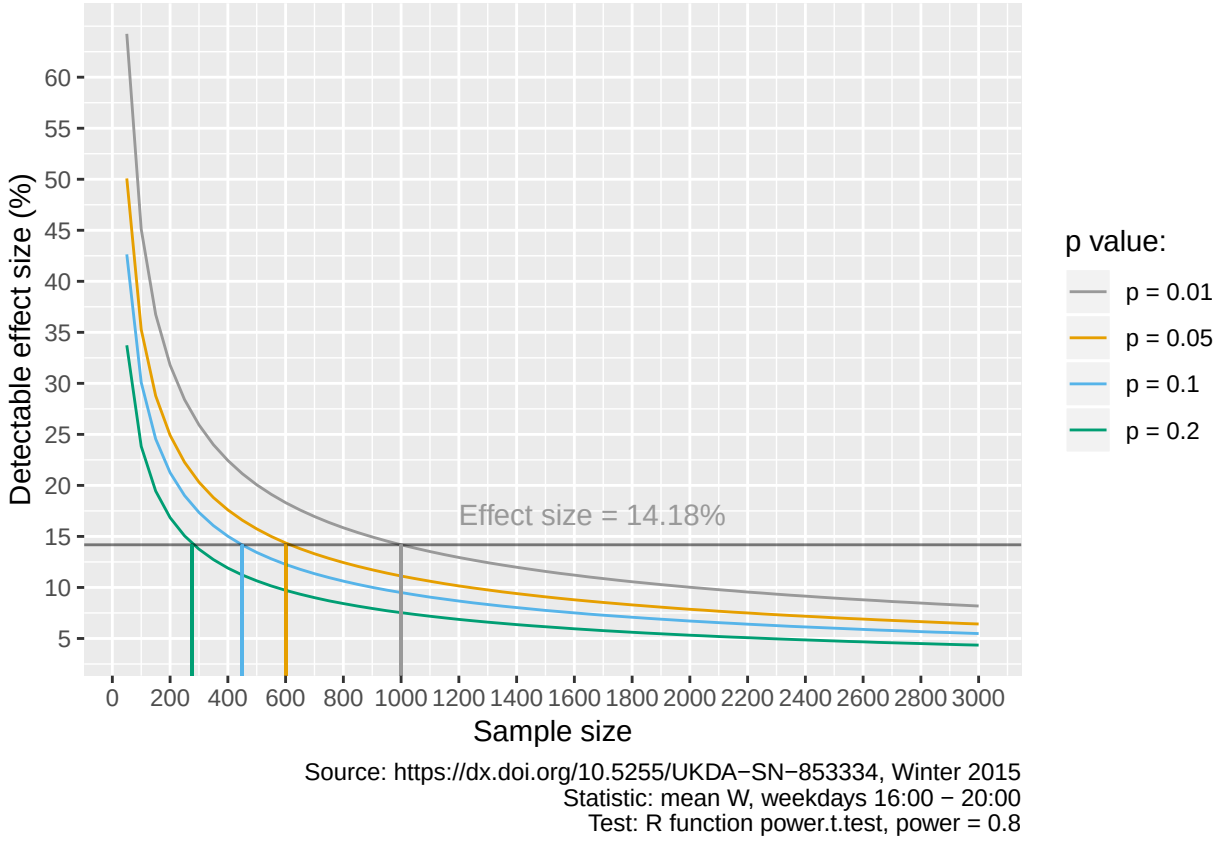
4 Sample design: statistical power

To return to the discussion of statistical power, we need to establish the probably size of the control and intervention groups we will require. This is an aid to resource budgeting (*"How many households and thus $ do I need?"*) and to ensure good study design practice ("*Will I be able to answer my research question?*") (Frederiks et al. 2016).

Calculation of the required sample size for a control and intervention group requires the estimation of the probable intervention effect size, agreement on the significance level (p value threshold or Type I error risk) of the statistical test to be used and agreement on the level of statistical power (Type II error risk). Given any three of these values the fourth can be calculated if an estimate of the mean and standard deviation of the outcome to be measured is known. In the case of DSR interventions the effect size comprises a given % reduction in energy demand or consumption in a given time period and estimates of the likely reduction can be derived from previous studies or data.

As we have noted the choice of significance level (p value threshold) and statistical power are subjective and normative. Most academic researchers will struggle to justify relaxing from the conventional $p = 0.05$ and power $= 0.8$. However as we have discussed there may be good reason in applied research to take action on results of studies that use less conservative thresholds. Nevertheless there is a strong argument for designing such studies using the more conservative conventional levels but acknowledging that making inferences from the results may require a more relaxed approach to Type I or Type II error risks than is considered 'normal' in academic research.

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```




```
## Saving 6.5 x 4.5 in image
```

As an illustration, **??**fig:ggHPSampleSizeFig) shows sample size calculations using 'Heat Pump' electricity demand extracted from the publicly available New Zealand Green Grid household electricity demand data (Anderson et al. 2018) for winter 2014 for the peak demand period (16:00 - 20:00) on weekdays.

As a guide, these results suggest that a trial comprising a control and intervention sample of 1000 households (each) would be able to detect an effect size of XXX with p = 0.05 and power = 0.8. Were a study to be less risk averse in it's decision making then p = 0.1 may be acceptable in which case only ~ XXX households would be needed in each group (see **??**fig:sampleSizeFig)) but the risk of a Type I error would increase. Reducing the statistical power used would also reduce the sample required for a given effect size tested at a given p value. However in this case the risk of a Type II error would increase.

5 Testing for differences: confidence intervals and p values

As an example, consider the a study which collected electricity power demand data for two different groups of households. The data shows that the mean W for group 1 was 35.14 and for group 2 was 162.67. This is a (very) large difference in the mean of 127.53.

A t-test of the difference between the groups produces the result shown below.

```
##
##  Welch Two Sample t-test
##
## data:  testDT[group == "S"]$meanW and testDT[group == "W"]$meanW
## t = -1.9907, df = 31.47, p-value = 0.05526
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -258.110005    3.050644
## sample estimates:
## mean of x mean of y
##  35.13947 162.66915
```

In this case we have:

- effect size = 127.5296803W or 78.4% representing a *substantial bang for buck* for whatever caused the difference;
- 95% confidence interval for the test = -258.11 to 3.05 representing *considerable* uncertainty/variation;
- p value of 0.055 representing a *relatively low* risk of a false positive results but which (just) fails the conventional $p < 0.05$ threshold.

What would we have concluded? We have a large effect size, substantial uncertainty and a slightly raised risk of a false positive or Type I error when compared to conventional p value levels. From a narrow and conventional 'p value testing' perspective we would have concluded that there was no statistically signficant difference between the groups. However this misses the crucial point that an organisation with a higher risk tolerance might conclude that the large effect size justifies implementing the intervention even though the risk of a false positive is slightly higher. If the p value had been 0.25 then this would have still been the case but would have warranted even further caution. As the recent discussions of the role of the p value in decision making have made clear (Wasserstein and Lazar 2016) statistical analysis needs to report all of these elements to enable contextually appropriate and defensible evidence-based decisions to be taken. Simply dismissing results on the basis of failure to meet conventional statistical levels of significance risks throwing both the baby and the bath water out of the window.

6 Summary and recomendations

7 Runtime

Analysis completed in 54.12 seconds (0.9 minutes) using knitr in RStudio with R version 3.5.1 (2018-07-02) running on x86_64-apple-darwin15.6.0.

8 R environment

R packages used:

- base R - for the basics (R Core Team 2016)
- data.table - for fast (big) data handling (Dowle et al. 2015)
- lubridate - date manipulation (Grolemund and Wickham 2011)
- ggplot2 - for slick graphics (Wickham 2009)
- readr - for csv reading/writing (Wickham, Hester, and Francois 2016)
- dplyr - for select and contains (Wickham and Francois 2016)
- progress - for progress bars (Csárdi and FitzJohn 2016)
- kableExtra - to create this document & neat tables (Xie 2016)
- GREENGrid - for local NZ GREEN Grid project utilities

Session info:

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
```

```
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] kableExtra_0.9.0  SAVEr_0.0.1.9000  lubridate_1.7.4   readr_1.1.1
## [5] ggplot2_3.0.0     dplyr_0.7.6       data.table_1.11.4 GREENGrid_0.1.0
## [9] GREENGridData_1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.18      lattice_0.20-35   tidyr_0.8.1
##  [4] prettyunits_1.0.2 png_0.1-7         utf8_1.1.4
##  [7] assertthat_0.2.0  rprojroot_1.3-2   digest_0.6.15
## [10] R6_2.2.2          cellranger_1.1.0  plyr_1.8.4
## [13] backports_1.1.2   evaluate_0.11     httr_1.3.1
## [16] pillar_1.3.0      RgoogleMaps_1.4.2 rlang_0.2.2
## [19] progress_1.2.0    lazyeval_0.2.1    readxl_1.1.0
## [22] rstudioapi_0.7    geosphere_1.5-7   rmarkdown_1.10
## [25] proto_1.0.0       stringr_1.3.1     munsell_0.5.0
## [28] broom_0.5.0       compiler_3.5.1    modelr_0.1.2
## [31] xfun_0.3          pkgconfig_2.0.2   htmltools_0.3.6
## [34] openssl_1.0.2     tidyselect_0.2.4  tibble_1.4.2
## [37] bookdown_0.7      fansi_0.3.0       viridisLite_0.3.0
## [40] crayon_1.3.4      withr_2.1.2       grid_3.5.1
## [43] nlme_3.1-137      jsonlite_1.5      gtable_0.2.0
## [46] magrittr_1.5      scales_1.0.0      cli_1.0.0
## [49] stringi_1.2.4     mapproj_1.2.6     reshape2_1.4.3
## [52] bindrcpp_0.2.2    sp_1.3-1          tidyverse_1.2.1
## [55] xml2_1.2.0        rjson_0.2.20      tools_3.5.1
## [58] forcats_0.3.0     ggmap_2.6.1       glue_1.3.0
## [61] purrr_0.2.5       maps_3.3.0        hms_0.4.2
## [64] jpeg_0.1-8        yaml_2.2.0        colorspace_1.3-2
## [67] rvest_0.3.2       knitr_1.20.13     bindr_0.1.1
## [70] haven_1.1.2
```

References

AECOM. 2011. “Energy Demand Research Project: Final Analysis.” St Albans: AECOM.

Anderson, Ben, David Eyers, Rebecca Ford, Diana Giraldo Ocampo, Rana Peniamina, Janet Stephenson, Kiti Suomalainen, Lara Wilcocks, and Michael Jack. 2018. “New Zealand GREEN Grid Household Electricity Demand Study 2014-2018,” September. doi:10.5255/UKDA-SN-853334.

CER. 2012. “Smart Meter Electricity Consumer Behaviour Trial data.” Dublin: Irish Social Science Data Archive. http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Energywise/.

Csárdi, Gábor, and Rich FitzJohn. 2016. *Progress: Terminal Progress Bars*. https://CRAN.R-project.org/

package=progress.

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. 2013. "Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012." *Energy Policy* 61 (October): 729–39. doi:10.1016/j.enpol.2013.05.109.

Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. 2015. *Data.table: Extension of Data.frame*. https://CRAN.R-project.org/package=data.table.

Frederiks, Elisha R., Karen Stenner, Elizabeth V. Hobman, and Mark Fischle. 2016. "Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers." *Energy Research & Social Science* 22 (December): 147–64. doi:10.1016/j.erss.2016.08.020.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (4): 337–50. doi:10.1007/s10654-016-0149-3.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. http://www.jstatsoft.org/v40/i03/.

Institute, Rocky Mountain. 2006. "Automated demand response system pilot: Final report." https://www.smartgrid.gov/files/Aumated_Demd_Response_System_Pilot_Volume_1_Intro_Exec_Summa.pdf.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schofield, James, Richard Carmichael, Simon Tindemans, Matt Woolf, Mark Bilton, and Goran Strbac. 2015. "Experimental validation of residential consumer responsiveness to dynamic time-of-use pricing." In *23 International Conference on Electricity Distribution*.

Srivastava, Aman, Steven Van Passel, and Erik Laes. 2018. "Assessing the Success of Electricity Demand Response Programs: A Meta-Analysis." *Energy Research & Social Science* 40 (June): 110–17. doi:10.1016/j.erss.2017.12.005.

UKPN. 2017. "The Final Energy Saving Trial Report." London: UK Power Networks. http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Energywise/.

———. 2018. "The Energy Shifting Trial Report." London: UK Power Networks. http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Energywise/.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The Asa's Statement on P-Values: Context, Process, and Purpose." *The American Statistician* 70 (2). Taylor & Francis: 129–33. doi:10.1080/00031305.2016.1154108.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Romain Francois. 2016. *Readr: Read Tabular Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2016. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.