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](mailto:b.anderson@soton.ac.uk), @dataknut)

Last run at: 2018-11-13 09:59:32

Table of Contents

# About

## License

This work is (c) the author(s).

License This work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/by/4.0/) unless otherwise marked.

For the avoidance of doubt and explanation of terms please refer to the full [license notice](https://creativecommons.org/licenses/by/4.0/) and [legal code](https://creativecommons.org/licenses/by/4.0/legalcode).

## Citation

If you wish to use any of the material from this paper 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), Southampton: University of Southampton.

This work is (c) 2018 the authors.

## History

Code & report history:

* [Paper history](https://github.com/dataknut/weGotThePower/commits/master)

## Data:

This report uses circuit level extracts for ‘Heat Pumps’ from 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 <https://github.com/CfSOtago/GREENGridData/blob/master/examples/code/extractCleanGridSpy1minCircuit.R>

## Acknowledgements

This work was supported by:

* The [University of Otago](https://www.otago.ac.nz/);
* The [University of Southampton](https://www.southampton.ac.uk/);
* The New Zealand [Ministry of Business, Innovation and Employment (MBIE)](http://www.mbie.govt.nz/) through the [NZ GREEN Grid](https://www.otago.ac.nz/centre-sustainability/research/energy/otago050285.html) grant (Contract ID: UOCX1203);
* The UK Office of Gas and Electricity Markets through the [Low Carbon Network Fund](https://www.ofgem.gov.uk/press-releases/ofgem-announces-%C2%A359.4-million-funding-10-innovation-projects)-funded ‘[Solent Achieving Value from Efficiency](http://www.energy.soton.ac.uk/tag/save/)’ (SAVE) project;
* [SPATIALEC](http://www.energy.soton.ac.uk/tag/spatialec/) - a [Marie Skłodowska-Curie Global Fellowship](http://ec.europa.eu/research/mariecurieactions/about-msca/actions/if/index_en.htm) based at the University of Otago’s [Centre for Sustainability](http://www.otago.ac.nz/centre-sustainability/staff/otago673896.html) (2017-2019) & the University of Southampton’s Sustainable Energy Research Group (2019-202).

# Introduction

This report contains the analysis for a paper of the same name. The text is stored elsewhere for ease of editing.

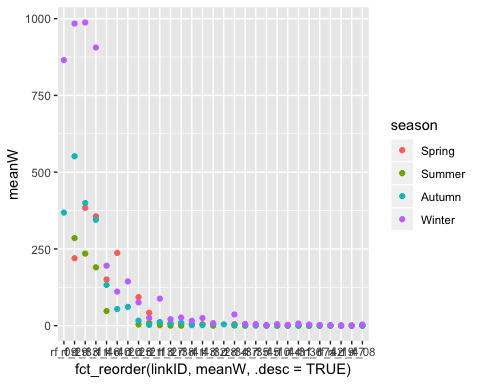
# Error, power, significance and decision making

# Sample design: statistical power

## Means

Table 1 Summary of mean consumption per household by season

|  |  |  |
| --- | --- | --- |
| season | meanMeanW | sdMeanW |
| Spring | 58.80597 | 113.53102 |
| Summer | 35.13947 | 83.90258 |
| Autumn | 68.37439 | 147.37279 |
| Winter | 162.66915 | 325.51171 |



Observations are summarised to mean W per household during 16:00 - 20:00 on weekdays for year = 2015.

## Warning: replacing previous import 'data.table::melt' by 'reshape2::melt'  
## when loading 'weGotThePower'

## Warning: replacing previous import 'data.table::dcast' by 'reshape2::dcast'  
## when loading 'weGotThePower'

Figure 1 shows the initial p = 0.01 plot.

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

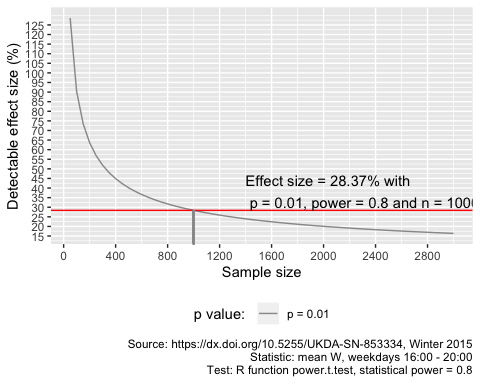


Figure 1 Power analysis results (p = 0.01, power = 0.8)

## Saving 5 x 4 in image

Effect size at n = 1000: 28.37.

Figure 2 shows the plot for all results.

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

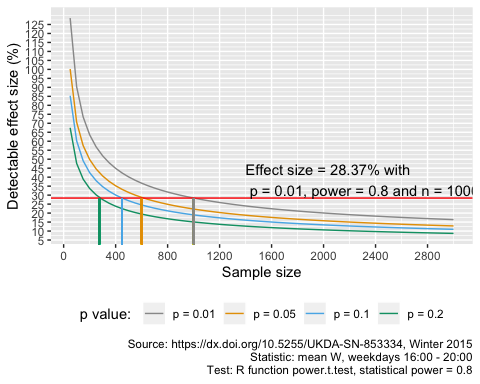


Figure 2 Power analysis results (power = 0.8)

## Saving 5 x 4 in image

Full table of results:

## Using 'effectSize' as value column. Use 'value.var' to override

Table 2 Power analysis for means results table (partial)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sampleN | p = 0.01 | p = 0.05 | p = 0.1 | p = 0.2 |
| 50 | 128.57 | 100.21 | 85.33 | 67.49 |
| 100 | 90.27 | 70.61 | 60.21 | 47.68 |
| 150 | 73.53 | 57.58 | 49.13 | 38.92 |
| 200 | 63.61 | 49.84 | 42.53 | 33.70 |
| 250 | 56.86 | 44.56 | 38.03 | 30.14 |
| 300 | 51.88 | 40.67 | 34.71 | 27.51 |
| 350 | 48.01 | 37.65 | 32.14 | 25.47 |
| 400 | 44.90 | 35.21 | 30.06 | 23.82 |
| 450 | 42.33 | 33.20 | 28.34 | 22.46 |
| 500 | 40.15 | 31.49 | 26.88 | 21.31 |
| 550 | 38.27 | 30.02 | 25.63 | 20.31 |
| 600 | 36.64 | 28.74 | 24.54 | 19.45 |
| 650 | 35.20 | 27.61 | 23.57 | 18.69 |
| 700 | 33.92 | 26.61 | 22.72 | 18.01 |
| 750 | 32.77 | 25.71 | 21.95 | 17.40 |
| 800 | 31.72 | 24.89 | 21.25 | 16.84 |
| 850 | 30.77 | 24.14 | 20.61 | 16.34 |
| 900 | 29.91 | 23.46 | 20.03 | 15.88 |
| 950 | 29.11 | 22.84 | 19.50 | 15.46 |
| 1000 | 28.37 | 22.26 | 19.00 | 15.06 |

## Proportions

Does not require a sample. As a relatively simple example, suppose we were interested in the adoption of heat pumps in two equal sized samples. Suppose we thought in one sample (say, home owners) we thought it might be 40% and in rental properties it would be 25% (ref BRANZ 2015). What sample size would we need to conclude a significant difference with power = 0.8 and at various p values?

pwr::pwr.tp.test() (ref pwr) can give us the answer…

Table 3 Samples required if p1 = 40% and p2 = 25%

|  |  |  |  |
| --- | --- | --- | --- |
| n | sig.level | power | props |
| 224.94 | 0.01 | 0.8 | p1 = 0.4 p2 = 0.25 |
| 151.17 | 0.05 | 0.8 | p1 = 0.4 p2 = 0.25 |
| 119.07 | 0.10 | 0.8 | p1 = 0.4 p2 = 0.25 |
| 86.73 | 0.20 | 0.8 | p1 = 0.4 p2 = 0.25 |

We can repeat this for other values of p1 and p2. For example, suppose both were much smaller (e.g. 10% and 15%)… Clearly we need *much* larger samples.

Table 4 Samples required if p1 = 10% and p2 = 15%

|  |  |  |  |
| --- | --- | --- | --- |
| n | sig.level | power | props |
| 1012.35 | 0.01 | 0.8 | p1 = 0.1 p2 = 0.15 |
| 680.35 | 0.05 | 0.8 | p1 = 0.1 p2 = 0.15 |
| 535.89 | 0.10 | 0.8 | p1 = 0.1 p2 = 0.15 |
| 390.31 | 0.20 | 0.8 | p1 = 0.1 p2 = 0.15 |

The above used an arcsine transform.

As a double check, using eqn to assess margin of error…

If:

* p = 0.4 (40%)
* n = 151

then the margin of error = +/- 0.078 (7.8%). So we could quote the Heat Pump uptake for owner-occupiers as 40% (+/- 7.8% [or 32.2 - 47.8] with p = 0.05).

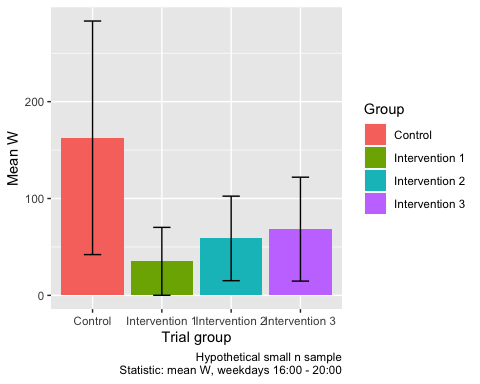
This may be far too wide an error margin for our purposes so we may instead have recruited 500 per sample. Now the margin of error is +/- 0.043 (4.3%) so we can now quote the Heat Pump uptake for owner-occupiers as 40% (+/- 4.3% [or 35.7 - 44.3] with p = 0.05).

# Testing for differences: effect sizes, confidence intervals and p values

## Getting it ‘wrong’

Table 5 Number of households and summary statistics per group

|  |  |  |  |
| --- | --- | --- | --- |
| group | mean W | sd W | n households |
| Control | 162.66915 | 325.51171 | 28 |
| Intervention 1 | 35.13947 | 83.90258 | 22 |
| Intervention 2 | 58.80597 | 113.53102 | 26 |
| Intervention 3 | 68.37439 | 147.37279 | 29 |



T test group 1

Table 6 T test results (Group 1 vs Control)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Control mean | Group 1 mean | Mean difference | statistic | p.value | conf.low | conf.high |
| 162.6691 | 35.13947 | -127.5297 | -1.990661 | 0.0552626 | -258.11 | 3.050644 |

The results show that the mean power demand for the control group was 162.67W and for Intervention 1 was 35.14W. This is a (very) large difference in the mean of 127.53. The results of the t test are:

* effect size = 128W or 78% 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 result but which (just) fails the conventional p < 0.05 threshold.

T test Group 2

Table 7 T test results (Group 2 vs Control)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Control mean | Group 2 mean | Mean difference | statistic | p.value | conf.low | conf.high |
| 162.6691 | 58.80597 | -103.8632 | -1.587604 | 0.1216582 | -236.8285 | 29.10212 |

Now:

* effect size = 104W or 63.85% representing a still *reasonable bang for buck* for whatever caused the difference;
* 95% confidence interval for the test = -236.83 to 29.1 representing *even greater* uncertainty/variation;
* p value of 0.122 representing a *higher* risk of a false positive result which fails the conventional p < 0.05 threshold and also the less conservative p < 0.1.

To detect Intervention Group 2’s effect size of 63.85% would have required control and trial group sizes of 47 respectively.

## Getting it ‘right’

Table 8 Number of households and summary statistics per group

|  |  |  |  |
| --- | --- | --- | --- |
| group | mean W | sd W | n households |
| Control | 152.24120 | 314.08715 | 1159 |
| Intervention 1 | 36.31760 | 83.62904 | 857 |
| Intervention 2 | 61.78741 | 112.03607 | 1017 |
| Intervention 3 | 73.82184 | 150.41622 | 1167 |

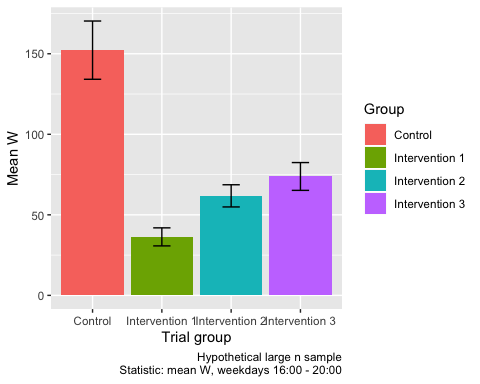


Figure 3 Mean W demand per group for large sample (Error bars = 95% confidence intervals for the sample mean)

re-run T tests Group 1

Table 9 T test results (Group 1 vs Control)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Control mean | Group 1 mean | Mean difference | statistic | p.value | conf.low | conf.high |
| 152.2412 | 61.78741 | -90.45379 | -9.16252 | 0 | -109.8187 | -71.08892 |

In this case:

* effect size = 90.4537895W or 59.41% representing a still *reasonable bang for buck* for whatever caused the difference;
* 95% confidence interval for the test = -109.82 to -71.09 representing *much less* uncertainty/variation;
* p value of 0 representing a *very low* risk of a false positive result as it passes all conventional thresholds.

# Summary and recommendations

## Statistical power and sample design

## Reporting statistical tests of difference (effects)

## Making inferences and taking decisions

# Acknowledgments

# Runtime

Analysis completed in 48.72 seconds ( 0.81 minutes) using [knitr](https://cran.r-project.org/package=knitr) in [RStudio](http://www.rstudio.com) with R version 3.5.1 (2018-07-02) running on x86\_64-apple-darwin15.6.0.

# 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)
* knitr - to create this document & neat tables (Xie 2016)
* pwr - non-base power analysis (Champely 2018)
* dkUtils - for local dataknut utilities :-) devtools::install\_github("dataknut/dkUtils")

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\_NZ.UTF-8/en\_NZ.UTF-8/en\_NZ.UTF-8/C/en\_NZ.UTF-8/en\_NZ.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] knitr\_1.20 pwr\_1.2-2 forcats\_0.3.0   
## [4] broom\_0.5.0 lubridate\_1.7.4 readr\_1.1.1   
## [7] ggplot2\_3.1.0 dplyr\_0.7.7 data.table\_1.11.8   
## [10] dkUtils\_0.0.0.9000  
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.19 highr\_0.7 pillar\_1.3.0   
## [4] compiler\_3.5.1 plyr\_1.8.4 bindr\_0.1.1   
## [7] tools\_3.5.1 digest\_0.6.18 lattice\_0.20-35   
## [10] nlme\_3.1-137 evaluate\_0.12 tibble\_1.4.2   
## [13] gtable\_0.2.0 pkgconfig\_2.0.2 rlang\_0.3.0.1   
## [16] cli\_1.0.1 yaml\_2.2.0 xfun\_0.4   
## [19] bindrcpp\_0.2.2 withr\_2.1.2 stringr\_1.3.1   
## [22] hms\_0.4.2 rprojroot\_1.3-2 grid\_3.5.1   
## [25] tidyselect\_0.2.5 glue\_1.3.0 R6\_2.3.0   
## [28] fansi\_0.4.0 rmarkdown\_1.10 bookdown\_0.7   
## [31] reshape2\_1.4.3 weGotThePower\_0.1 tidyr\_0.8.1   
## [34] purrr\_0.2.5 magrittr\_1.5 backports\_1.1.2   
## [37] scales\_1.0.0 htmltools\_0.3.6 assertthat\_0.2.0   
## [40] colorspace\_1.3-2 labeling\_0.3 utf8\_1.1.4   
## [43] stringi\_1.2.4 lazyeval\_0.2.1 munsell\_0.5.0   
## [46] crayon\_1.3.4

# References

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](https://doi.org/10.5255/UKDA-SN-853334).

Champely, Stephane. 2018. *Pwr: Basic Functions for Power Analysis*. <https://CRAN.R-project.org/package=pwr>.

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

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>.

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/>.

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

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>.