

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: 2018-09-20 18:00:10

Contents

1	About	2
1.1	Paper circulation:	2
1.2	License	2
1.3	Citation	2
1.4	History	2
1.5	Data:	2
1.6	Support	2
2	Introduction	3
3	Error, power, significance and decision making	3
4	Sample design: statistical power	4
5	Testing for differences: effect sizes, confidence intervals and p values	7
5.1	Getting it ‘wrong’	7
5.2	Getting it ‘right’	9
6	Summary and recommendations	10
6.1	Statistical power and sample design	10
6.2	Reporting statistical tests of difference (effects)	10
6.3	Making inferences and taking decisions	10
7	Acknowledgements	11
8	Runtime	11
9	R environment	11
	References	12

1 About

1.1 Paper 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.

This work is (c) 2018 the authors.

1.4 History

Code history is generally tracked via our 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 experience 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 significance in testing for differences and of both statistical significance 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 research where the efficacy of an intervention is to be tested and also in applied research where a decision may then be taken based on the results:

- 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.

Type I error: The significance level (p value) of the statistical test to be used to test the efficacy of an intervention 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 intervention 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.

Type II error: 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 reached 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 (Rocky Mountain Institute 2006, AECOM (2011), CER (2012), Schofield et al. (2015)) or both lower statistical power values *and* lower 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 trial groups to demonstrate and clarify these points.

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).

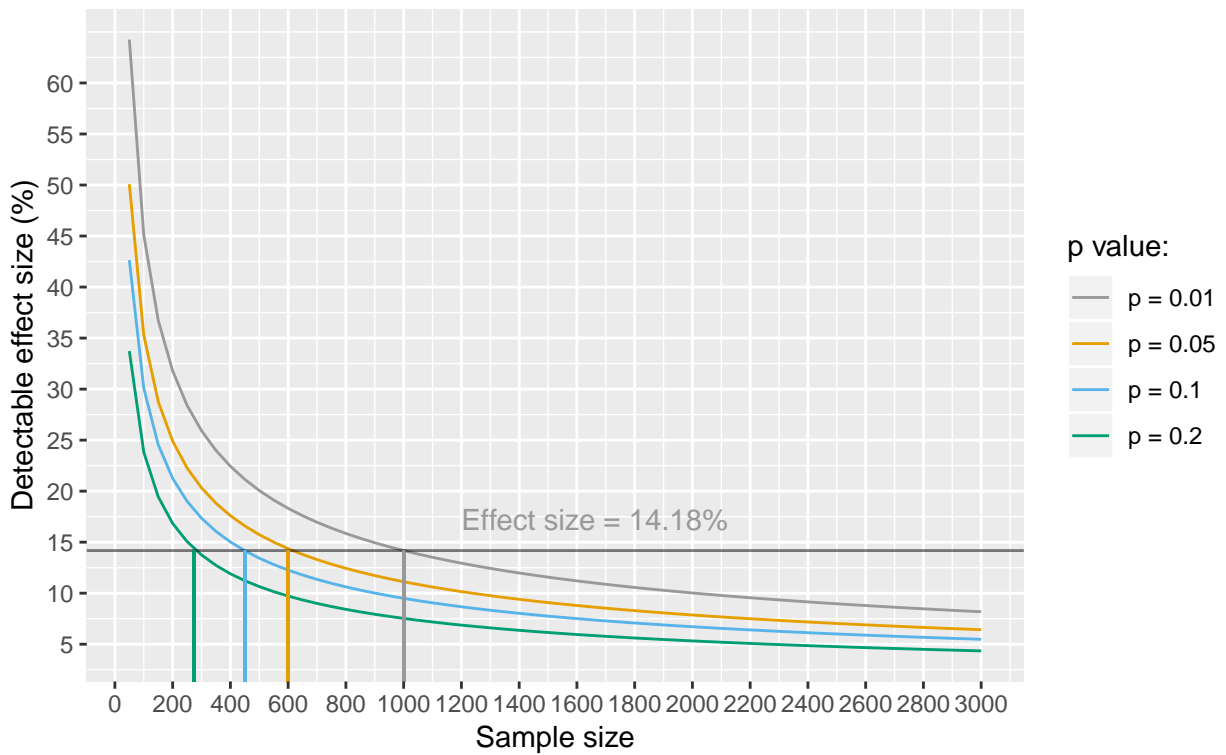
Calculation of the required sample size for a control and intervention group requires the estimation of the probable intervention effect size, agreement on the significance level (p value threshold or Type I error risk) of the statistical test to be used and agreement on the level of statistical power (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.

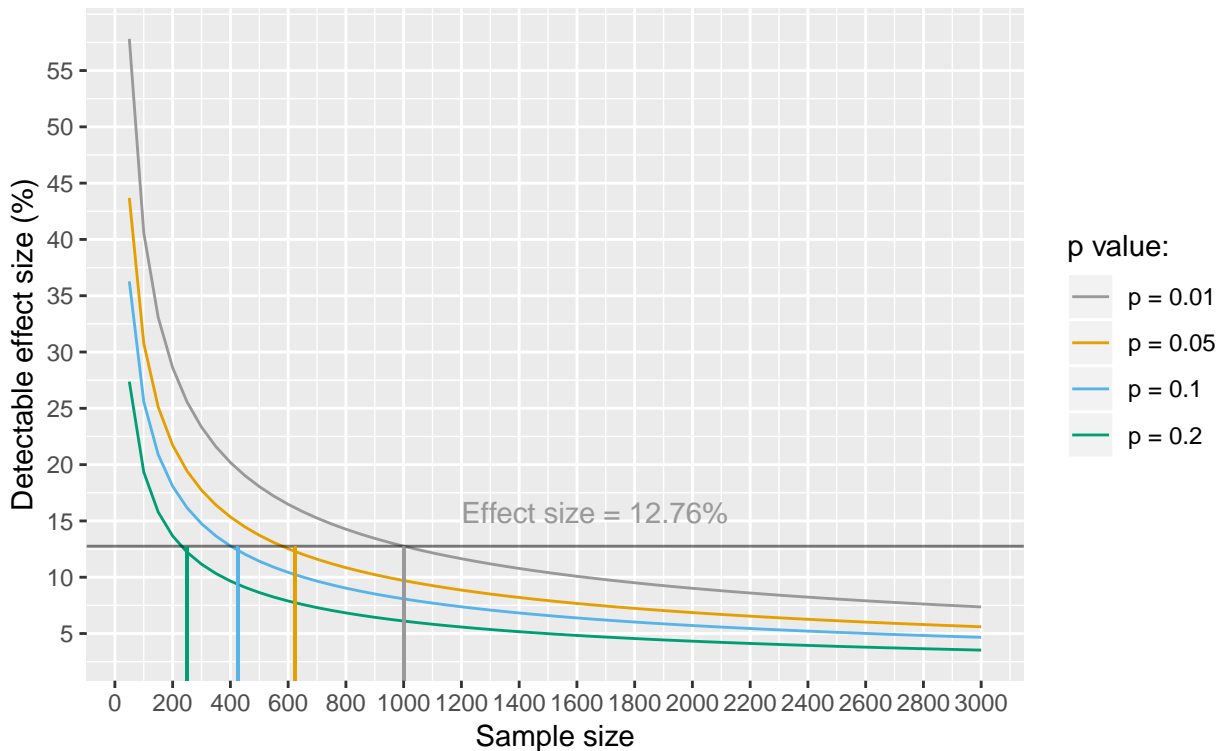
## Saving 6.5 x 4.5 in image
```

As an illustration,  shows sample size calculations for power = 0.8 (80%) using ‘Heat Pump’ electricity demand extracted from the publicly available New Zealand Green Grid household electricity demand data (Anderson et al. 2018) for winter 2015 for the peak demand period (16:00 - 20:00) on weekdays.



Source: <https://dx.doi.org/10.5255/UKDA-SN-853334>, Winter 2015
 Statistic: mean W, weekdays 16:00 – 20:00
 Test: R function power.t.test, statistical power = 0.8

Figure 1: Power analysis results (power = 0.8)



Source: <https://dx.doi.org/10.5255/UKDA-SN-853334>, Winter 2015
 Statistic: mean W, weekdays 16:00 – 20:00
 Test: R function power.t.test, statistical power = 0.7

Figure 2: Power analysis results (power = 0.7)

These results show that a trial comprising a control and intervention sample of 1000 households (each) would be able to detect an effect size of 14.18% with $p = 0.01$ and power = 0.8. Were a study to be less risk averse in its decision making then $p = 0.1$ may be acceptable in which case only ~ 425 households would be needed in each group (see [fig:ggHPSampleSizeFig80](#)) but the risk of a Type I error would increase.

```
## 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
```

Were we to reduce the statistical power to 0.7 then we would obtain the results shown in [fig:ggHPSampleSizeFig70](#)). In this case a trial comprising a control and intervention sample of 1000 households (each) would be able to detect an effect size of 12.76% with $p = 0.01$ and power = 0.7. Were a study to be less risk averse in its decision making then $p = 0.1$ may be acceptable in which case only ~ 425 households would be needed in each group (see [fig:ggHPSampleSizeFig70](#)) but as before the risk of a Type I error would increase. Similarly, reducing the statistical power used would also reduce the sample required for a given effect size tested at a given p value. However, as before the risk of a Type II error would increase.

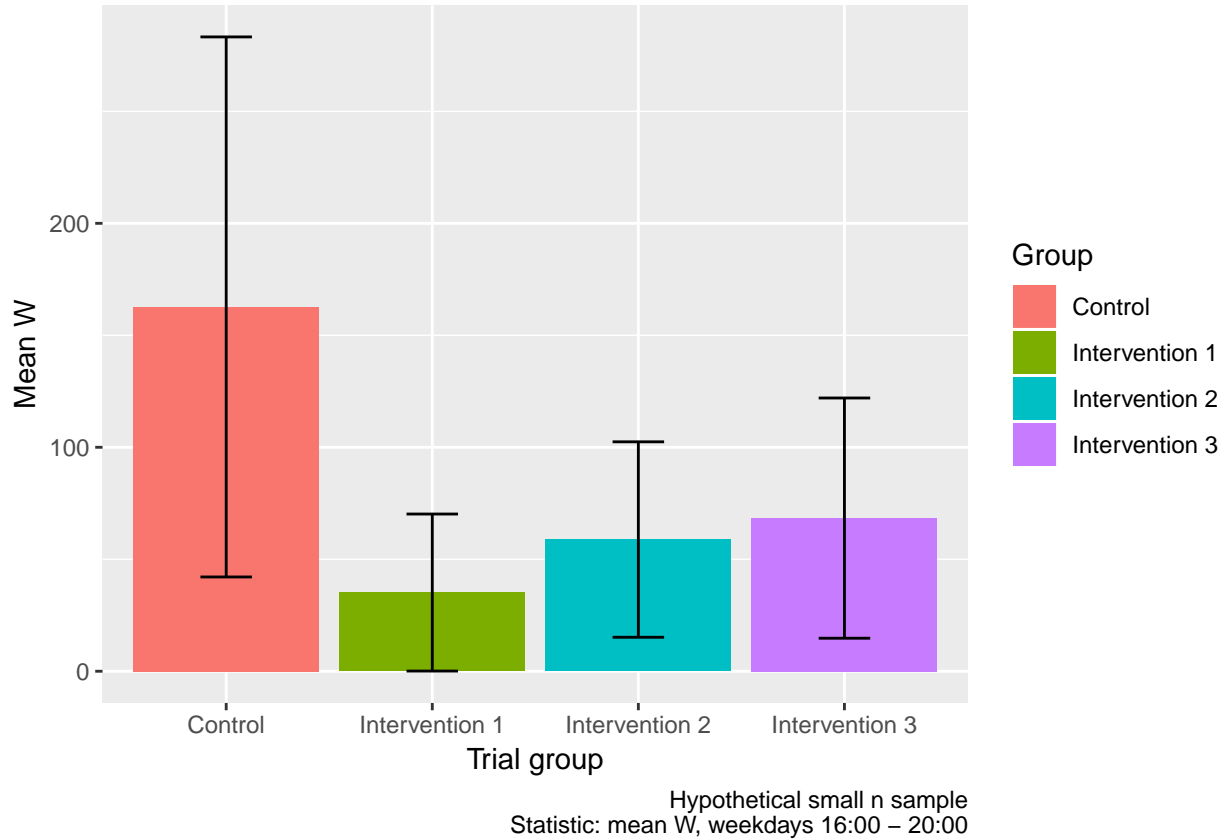
Table 1: Number of households and summary statistics per group

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

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

5.1 Getting it ‘wrong’

Let us imagine that we have not designed and implemented our sample recruitment according to [fig:ggHPSampleSizeFig80](#)) and instead decided, perhaps for cost reasons to recruit ~ 30 households per group. Now we wish to test for differences between the control and intervention groups.



As a first step we plot the differences using the mean and 95% confidence intervals as shown in [fig:ggMeanDiffs](#)). As we can see the interventions appear to have reduced demand quite substantially and the error bars indicate the uncertainty (variation) around the mean within each group. Based on this, we suspect that we are unlikely to see low p values when we use statistical tests of the differences as the error bars overlap substantially.

Suppose a t-test of the difference between the Control and Intervention 1 group produces the result shown below.

##

```
## Welch Two Sample t-test
##
## data: testDT[group == "Intervention 1"]$meanW and testDT[group == "Control"]$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
```

The data shows 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.

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 significant 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.

But what about Intervention Group 2? In this case the t.test results are slightly different:

```
##
## Welch Two Sample t-test
##
## data: testDT[group == "Intervention 2"]$meanW and testDT[group == "Control"]$meanW
## t = -1.5876, df = 33.909, p-value = 0.1217
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -236.82848 29.10212
## sample estimates:
## mean of x mean of y
## 58.80597 162.66915
```

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 $p < 0.1$.

As before, the subsequent action we take depends on our tolerance of Type I (false positive) risk. We still have a reasonably large effect size but we are less certain about it and we have a higher risk of it not being real. What do you think we should do?

In both cases our decision-making is rather hampered by the small sample size even though we have extremely large effect sizes. As we can see from [fig:ggHPSampleSizeFig80](#)), to detect Intervention Group 2’s effect size of 63.85% would have required control and trial group sizes of 47 respectively.

However, 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 the result elements to enable contextually appropriate

Table 2: Number of households and summary statistics per group

group	mean W	sd W	n households
Control	139.28889	292.78141	1137
Intervention 1	36.47888	84.60407	835
Intervention 2	60.40882	113.37202	1054
Intervention 3	73.64247	147.39290	1174

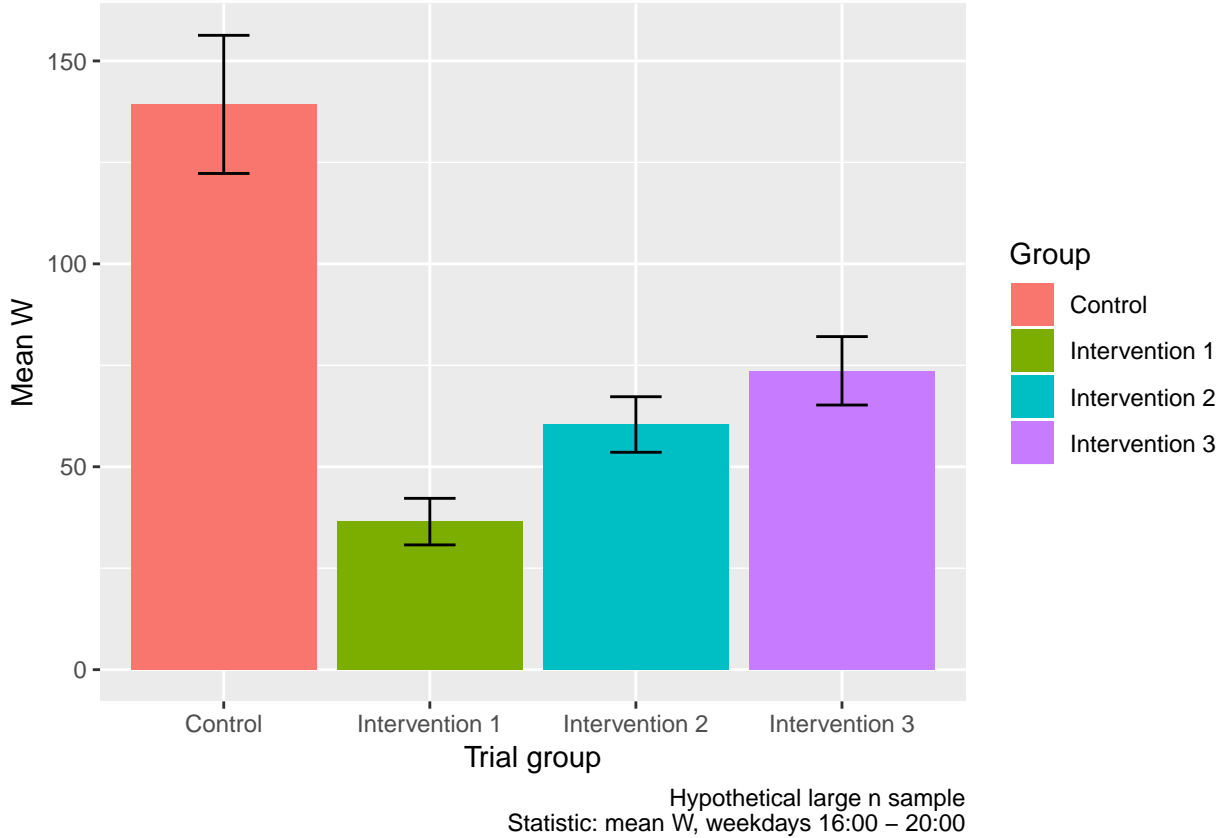


Figure 3: Mean W demand per group for large sample (Error bars = 95% confidence intervals for the sample mean)

and defensible evidence-based decisions to be taken. Simply dismissing results on the basis of a failure to meet conventional statistical levels of significance risks levitating babies and bathwater...

5.2 Getting it ‘right’

Suppose instead that we had designed and implemented our sample recruitment according to [fig:ggHPSampleSizeFig80](#)) so that we have a reasonable chance of detecting a difference of $\sim 14\%$ with power = 0.8 and at a significance level (p) of 0.05. This means we should have a sample of around 4000 households split equally (and randomly) between our trial and two intervention groups.

In comparison to [fig:ggMeanDiffs](#)) we can now see ([fig:largeNmeanDiffs](#))) that the 95% confidence intervals for the group means are much narrower. This is almost entirely due to the larger sample sizes. Re-running our previous test for differences now produces:

```
##
```

```
## Welch Two Sample t-test
##
## data: largeTestDT[group == "Intervention 2"]$meanW and largeTestDT[group == "Control"]$meanW
## t = -8.4284, df = 1491.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -97.23786 -60.52228
## sample estimates:
## mean of x mean of y
## 60.40882 139.28889
```

In this case:

- effect size = 78.880071W or 56.63% representing a still *reasonable bang for buck* for whatever caused the difference;
- 95% confidence interval for the test = -97.24 to -60.52 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.

So now we are able to be much more confident in our decision to implement Intervention 2 since the average effect is reasonably large, the expected variation in the effect size is reasonably narrow and the risk of a Type I (false positive) error is extremely small.

6 Summary and recommendations

6.1 Statistical power and sample design

Get it right *first time*: we should do the statistical power analysis before we start to make sure the study is even worth trying. If we don't have previous data to use, we *justify* our choices through power analysis based on defensible assumptions.

6.2 Reporting statistical tests of difference (effects)

Report all three elements *always*:

- average effect size
- effect size confidence intervals
- the p value (risk of Type I errors)

We should also report the statistical power used just to be clear on the risk of Type II errors.

6.3 Making inferences and taking decisions

Pay attention to all three elements *always*:

- average effect size: what is the *average bang for buck*?
- effect size confidence intervals: *how uncertain is the bang*?
- the p value: *what is the risk of a false positive*?

If we have ticked all the boxes so far then we have combined good study design based on statistical power analysis, with a nuanced understanding of what test statistic effect sizes, confidence intervals and p values can tell us. As a result we now have a robust, evidence-based, contextually meaningful and *defensible* strategy.

7 Acknowledgements

We would like to thank collaborators and partners on a number of applied research projects for prodding us into thinking about these issues more deeply and clearly than we otherwise would have done. We hope this paper helps to bring some clarity.

8 Runtime

Analysis completed in 67.65 seconds (1.13 minutes) using knitr in RStudio with R version 3.5.1 (2018-07-02) running on x86_64-apple-darwin15.6.0.

9 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      highr_0.7
```

## [16] httr_1.3.1	pillar_1.3.0	RgoogleMaps_1.4.2
## [19] rlang_0.2.2	progress_1.2.0	lazyeval_0.2.1
## [22] readxl_1.1.0	rstudioapi_0.7	geosphere_1.5-7
## [25] rmarkdown_1.10	labeling_0.3	proto_1.0.0
## [28] stringr_1.3.1	munsell_0.5.0	broom_0.5.0
## [31] compiler_3.5.1	modelr_0.1.2	xfun_0.3
## [34] pkgconfig_2.0.2	htmltools_0.3.6	openssl_1.0.2
## [37] tidyselect_0.2.4	tibble_1.4.2	bookdown_0.7
## [40] fansi_0.3.0	viridisLite_0.3.0	crayon_1.3.4
## [43] withr_2.1.2	grid_3.5.1	nlme_3.1-137
## [46] jsonlite_1.5	gtable_0.2.0	magrittr_1.5
## [49] scales_1.0.0	cli_1.0.0	stringi_1.2.4
## [52] mapproj_1.2.6	reshape2_1.4.3	bindrcpp_0.2.2
## [55] sp_1.3-1	tidyverse_1.2.1	xml2_1.2.0
## [58] rjson_0.2.20	tools_3.5.1	forcats_0.3.0
## [61] ggmap_2.6.1	glue_1.3.0	purrr_0.2.5
## [64] maps_3.3.0	hms_0.4.2	jpeg_0.1-8
## [67] yaml_2.2.0	colorspace_1.3-2	rvest_0.3.2
## [70] knitr_1.20.13	bindr_0.1.1	haven_1.1.2

References

- AECOM. 2011. “Energy Demand Research Project: Final Analysis.” St Albans: AECOM.
- Anderson, Ben, David Eysers, 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/>.
- R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rocky Mountain Institute. 2006. “Automated demand response system pilot: Final report.” <https://www.rockymountaininstitute.org/>.

smartgrid.gov/files/Aumated_Demd_Response_System_Pilot_Volume_1_Intro_Exec_Summa.pdf.

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>.