UNIVERSITY OF
Southampton

Ensuring statistics have power

Sample sizes, effect sizes and confidence intervals (and how to use them)

11th March 2021

Ben Anderson
@dataknut

The Menu

- What do we need to know?
    - Effect sizes, precision and the risk of getting it ‘wrong’
- Case studies:
    - Actual small sample
    - Simulated large(r) sample
- Decisions:
    - Before: Study design
    - After: Evidence, certainty and risk
- Summary

UNIVERSITY OF Southampton

Evaluation: we need to know

Is it 2% or 22%

Is it useful?

Difference or effect size

- Is the result *important* or *useful?*
- “What is the estimated *bang for buck*?”)

15-29% ?

Statistical Confidence Intervals

- Is there *uncertainty* or *variation* in response?
- “How uncertain is the estimated ba ?”

Are we sure enough?

Statistical *p* values

- Risk of a Type I error / *false positive?*
- “Risk the bang isn't real?”

p = 0.1?

We might waste £ on something that doesn’t work

Statistical power

- Risk of a Type II error / *false negative?*
- “Risk there is a bang when we concluded there wasn't?”

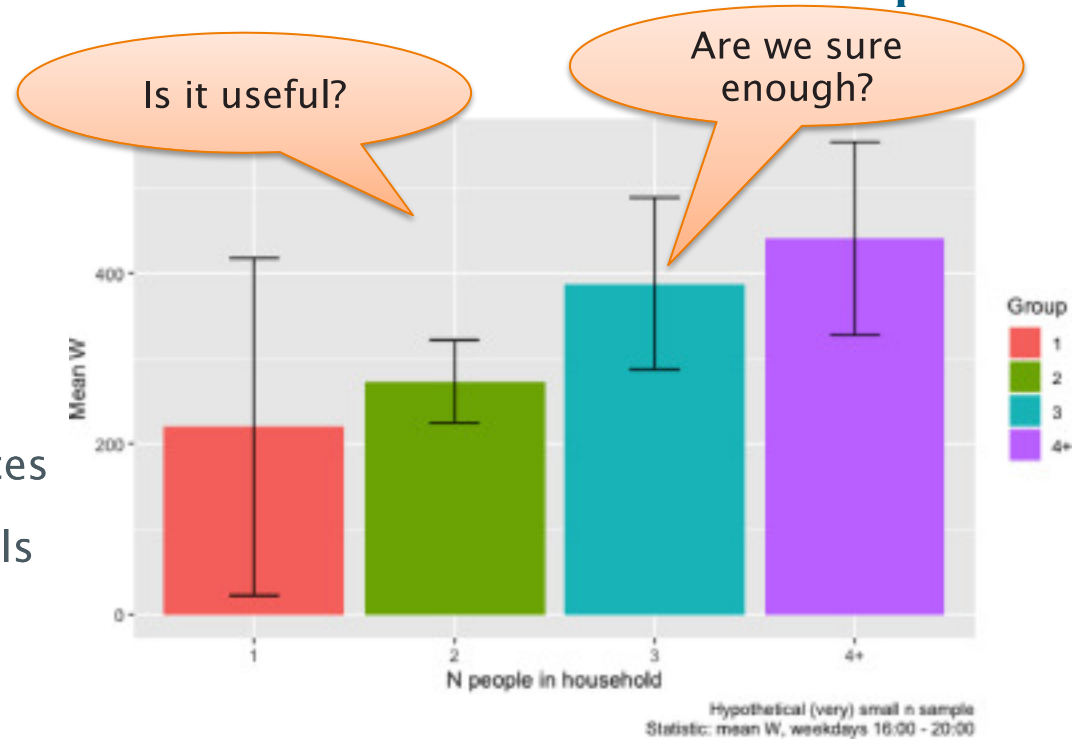
power = 0.8?

We might not do something that does work

UNIVERSITY OF Southampton

An example…

- Heat pump power demand*
- Total sample = 53
  - There are ‘useful’ differences
  - But 95% confidence intervals overlap
  - So none are ‘statistically significant’
  - And all are imprecise



N people	N households
1	3
2	6
3	20
4+	23

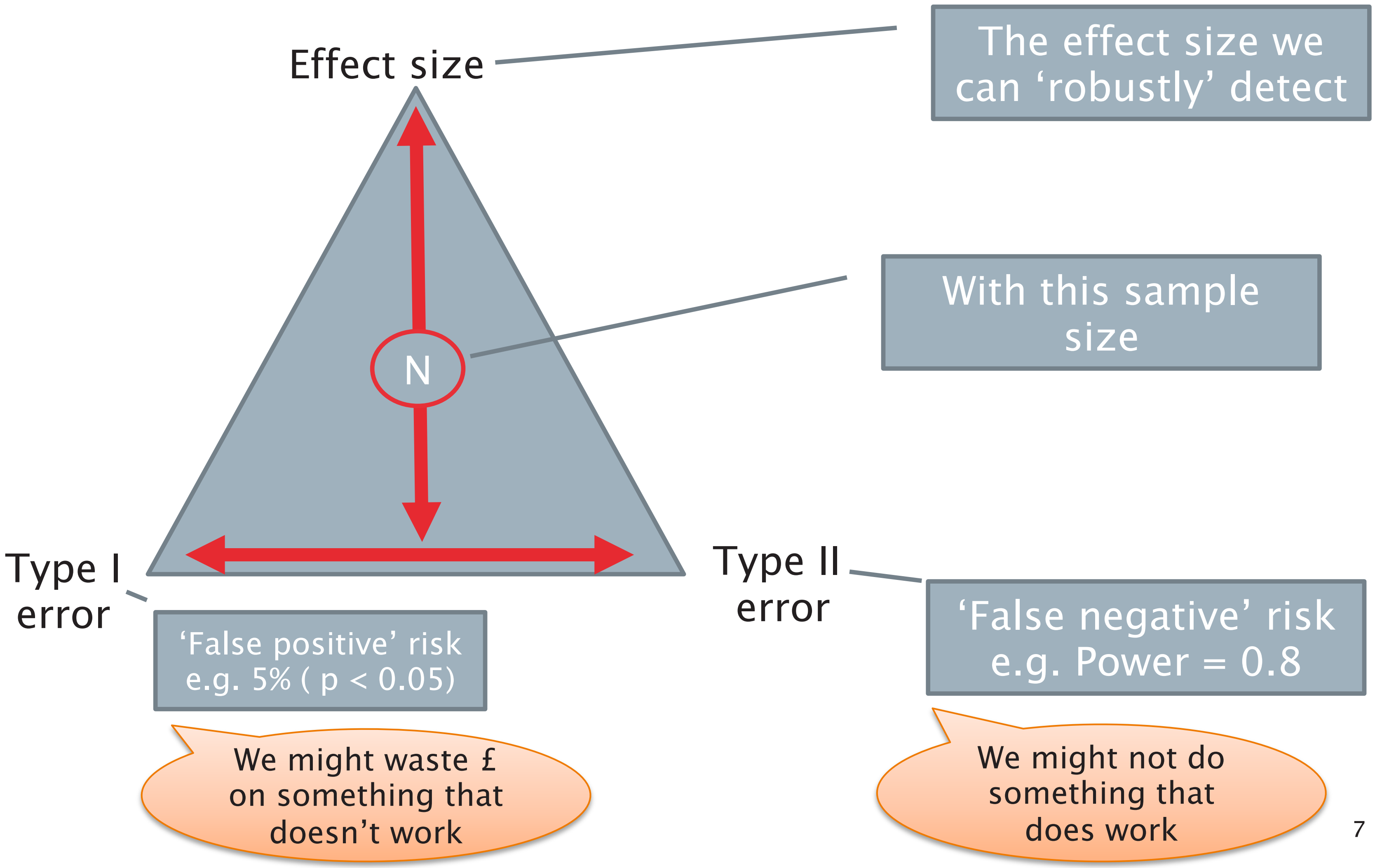
*Data source: B. Anderson et al., ‘New Zealand GREEN Grid household electricity demand study 2014-2018’, Sep. 2018

An example… 2

- Heat pump power demand*
- Simulated sample^ = 1,040
  - There are ‘very useful’ differences
  - 95% confidence intervals do not overlap
  - All are ‘statistically significant’
  - And all are much more precise

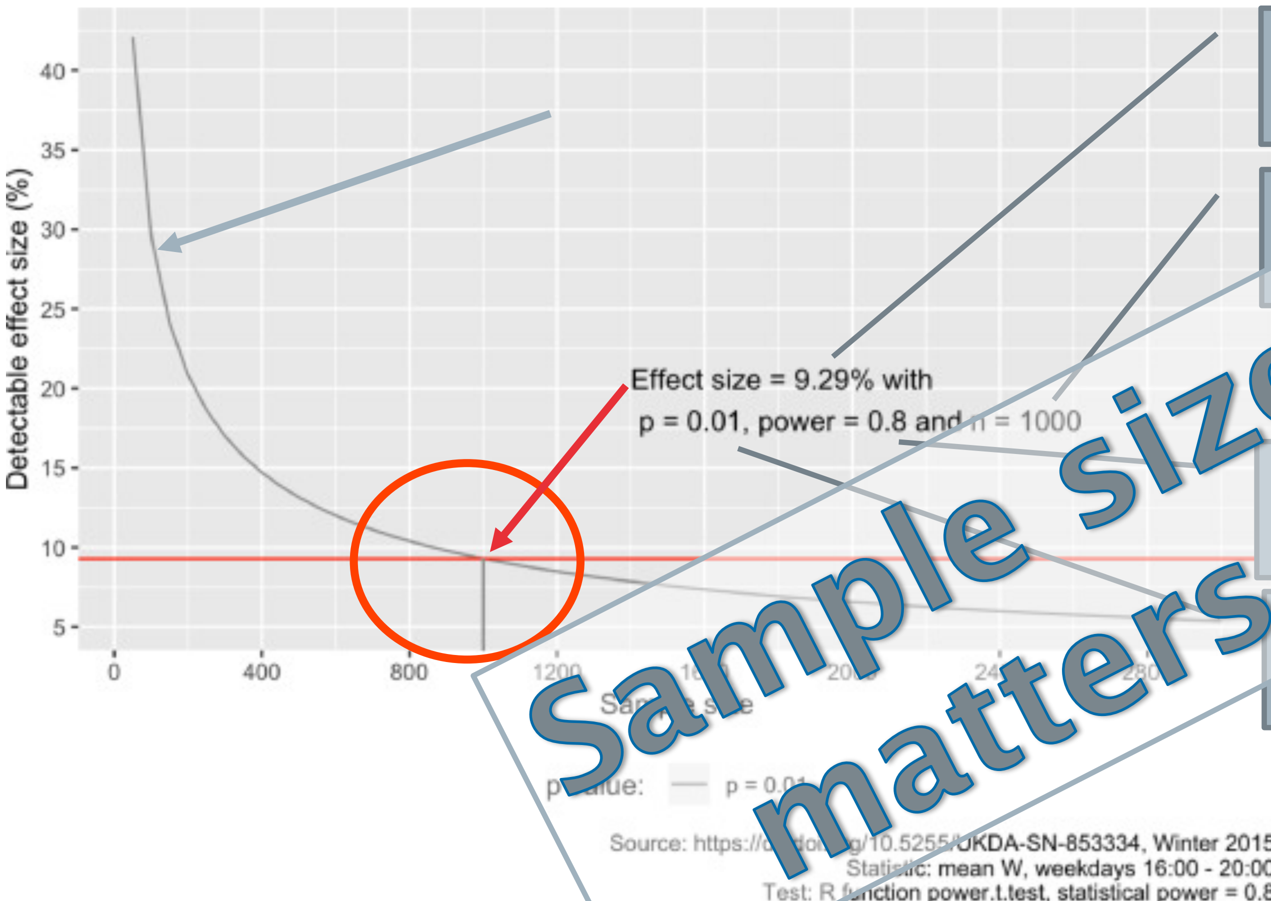
*Data source: B. Anderson et al., ‘[New Zealand GREEN Grid household electricity demand study 2014-2018](#)’, Sep. 2018
^Repeated random sampling from 53 with replacement

Decisions before: power analysis



UNIVERSITY OF Southampton

Power Analysis: Start here…



The effect size we can ‘robustly’ detect

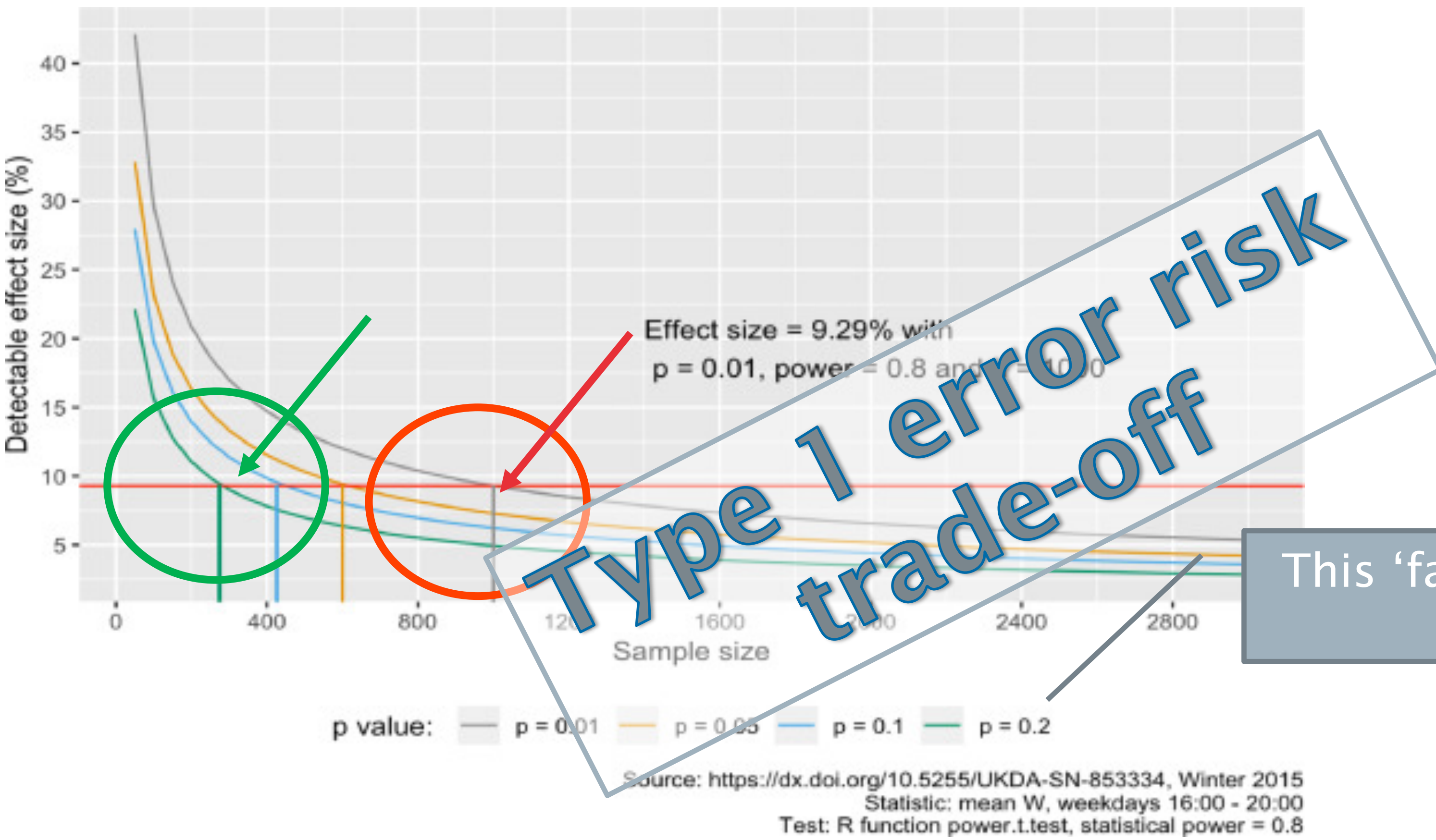
With this sample size…

This ‘false negative’ risk and…

This ‘false positive’ risk

Sample size matters

Power Analysis: depending on risk appetite



This 'false positive' risk

Decisions after: Evidence, certainty and risk

- Suppose:
  - Trial 1: needs 4% to be worthwhile
  - Trial 2: needs 18% to be worthwhile

	Trial 1	Trial 2
Mean effect size	6%	16%
95% Confidence Interval	-1% to 13%	10% to 22%
Test p value (Type I)	0.12	0.04
Power (Type II)	0.8	0.8

1. Mean effect size is large enough
2. 95% CI
   - include the target
   - are wide and include 0
3. The effect is n/s at p = 0.05 and p = 0.1

1. Mean effect size is not quite large enough
2. 95% CI
   - include the target
   - are wide but do not include 0
3. The effect is statistically significant at p = 0.05

Summary

Reporting evidence:

- Sample size -> is it big enough?
- Effect sizes -> is it useful enough?
- Confidence intervals -> is it precise enough?
- Statistical significance thresholds -> is it random chance?

Is it useful?

Are we sure enough?

Thresholds depend on your appetite for:

- Type I error (*test p value*)
  - You conclude it '*worked*' when (in fact) it didn't
- Type II error (*statistical power*)
  - You conclude it '*didn't work*' when (in fact) it did

We might waste £ on something that doesn't work

We might not do something that does work

Which depend on:

- The social, reputational and £ costs *if you're wrong*
- The benefits *if you're right*

YOUR QUESTIONS

b.anderson@soton.ac.uk
@dataknut
https://doi.org/10.1016/j.erss.2019.101260