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
Evaluation: we need to know

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

**Statistical Confidence Intervals**
- Is there *uncertainty* or *variation* in response?
- “How uncertain is the estimated bang?”

**Statistical p values**
- Risk of a Type I error / *false positive*?
- “Risk the bang isn't real?”
- p = 0.1?

**Statistical power**
- Risk of a Type II error / *false negative*?
- “Risk there is a bang when we concluded there wasn't?”
- power = 0.8?

- Is it useful?
- 15-29%?
- Are we sure enough?
- We might waste £ on something that doesn't work
- We might not do something that does work
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

An example… 2

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

\(^{\wedge}\)Repeated random sampling from 53 with replacement
Decisions before: power analysis

The effect size we can ‘robustly’ detect

With this sample size

We might waste £ on something that doesn’t work

We might not do something that does work

Effect size

Type I error

‘False positive’ risk e.g. 5% (p < 0.05)

Type II error

‘False negative’ risk e.g. Power = 0.8
Power Analysis: Start here…

The effect size we can ‘robustly’ detect

With this sample size…

This ‘false negative’ risk and…

This ‘false positive’ risk
Power Analysis: depending on risk appetite

This ‘false positive’ risk trade-off

Effect size = 9.29% with
p = 0.01, power = 0.8 approx.

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
Decisions after: Evidence, certainty and risk

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

<table>
<thead>
<tr>
<th></th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean effect size</strong></td>
<td>6%</td>
<td>16%</td>
</tr>
<tr>
<td><strong>95% Confidence Interval</strong></td>
<td>-1% to 13%</td>
<td>10% to 22%</td>
</tr>
<tr>
<td><strong>Test p value (Type I)</strong></td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Power (Type II)</strong></td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

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?

**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

**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