Power Analysis

Anne Segonds-Pichon
v2020-09
• **Definition of power**: probability that a statistical test will reject a false null hypothesis ($H_0$).
  • **Translation**: the probability of detecting an effect, given that the effect is really there.

• **In a nutshell**: the bigger the experiment (big sample size), the bigger the power (more likely to pick up a difference).

• **Main output of a power analysis**:
  • Estimation of an appropriate **sample size**
    • **Too big**: waste of resources,
    • **Too small**: may miss the effect ($p>0.05$) + waste of resources,
    • **Grants**: justification of sample size,
  • **Publications**: reviewers ask for power calculation evidence,
  • **Home office**: the 3 Rs: Replacement, **Reduction** and Refinement.
What does Power look like?
What does Power look like? Null and alternative hypotheses

- Probability that the observed result occurs if $H_0$ is true
  - $H_0$: Null hypothesis = absence of effect
  - $H_1$: Alternative hypothesis = presence of an effect
What does Power look like? Type I error $\alpha$

- **Type I error** ($\alpha$) is the failure to reject a **true** $H_0$
  - Claiming an effect which is not there.

- **p-value**: probability that the observed statistic occurred by chance alone
  - probability that a difference as big as the one observed could be found even if there is no effect.

- **Statistical significance**: comparison between $\alpha$ and the **p-value**
  - $p$-value < 0.05: reject $H_0$
  - $p$-value > 0.05: fail to reject $H_0$
What does Power look like? Power and Type II error $\beta$

- **Type II error** ($\beta$) is the failure to reject a false $H_0$
  - Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there.

- Direct relationship between **Power** and **type II error**: $\beta$
  - **Power** = $1 - \beta$
What does Power look like? Power = 80%

• **General convention:** 80% but could be more
  • if Power = 0.8 then $\beta = 1 - \text{Power} = 0.2$ (20%)

• Hence a true difference will be missed 20% of the time

• Jacob Cohen (1962):
  • For most researchers: Type I errors are four times more serious than Type II errors so: $0.05 \times 4 = 0.2$
    • Compromise: 2 groups comparisons:
      • 90% = +30% sample size
      • 95% = +60% sample size
Critical value = size of difference + sample size + significance
What does Power look like? Example with the $t$-test

- **In hypothesis testing:**
  - **test statistic** is compared to the **critical value** to determine significance
  - Example of test statistic: $t$-value

- If **test statistic** > **critical value**: statistical significance and rejection of the null **hypothesis**
  - Example: $t$-value > critical $t$-value
To recapitulate:

- The null hypothesis \((H_0)\): \(H_0 = \text{no effect}\)
- The aim of a statistical test is to reject or not \(H_0\).

<table>
<thead>
<tr>
<th>Statistical decision</th>
<th>True state of (H_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(H_0\ True (no effect))</td>
</tr>
<tr>
<td>Reject (H_0)</td>
<td>Type I error (\alpha)</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
</tr>
<tr>
<td>Do not reject (H_0)</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**

[https://github.com/allisonhorst/stats-illustrations#other-stats-artwork](https://github.com/allisonhorst/stats-illustrations#other-stats-artwork)
The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **variability** in the data (standard deviation)
- the **significance level** (5%)
- the desired **power** of the experiment (80%)
- the **sample size**
- the alternative hypothesis (ie **one or two-sided test**)

Effect size

Sample Size: Power Analysis
The difference of biological interest

- This is to be determined scientifically, not statistically.
  - minimum meaningful effect of biological relevance
  - the larger the effect size, the smaller the experiment will need to be to detect it.

- **How to determine it?**
  - Previous research, pilot study ...

The Standard Deviation (SD)

- Variability of the data

- **How to determine it?**
  - Data from previous research on WT or baseline ...
The effect size: what is it?

- The **effect size**: Absolute difference + variability

- How to determine it?
  - Substantive knowledge
  - Previous research
  - Conventions

- **Jacob Cohen**
  - Defined small, medium and large effects for different tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Relevant effect size</th>
<th>Effect Size Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>t-test for means</td>
<td>d</td>
<td>0.2</td>
</tr>
<tr>
<td>F-test for ANOVA</td>
<td>f</td>
<td>0.1</td>
</tr>
<tr>
<td>t-test for correlation</td>
<td>r</td>
<td>0.1</td>
</tr>
<tr>
<td>Chi-square</td>
<td>w</td>
<td>0.1</td>
</tr>
<tr>
<td>2 proportions</td>
<td>h</td>
<td>0.2</td>
</tr>
</tbody>
</table>
The effect size: how is it calculated?

The absolute difference

- It depends on the type of difference and the data
- Easy example: comparison between 2 means
  
  \[
  \text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}
  \]

- The bigger the effect (the absolute difference), the bigger the power
  = the bigger the probability of picking up the difference

http://rpsychologist.com/d3/cohend/
The effect size: how is it calculated?

The standard deviation

• The bigger the variability of the data, the smaller the power

\[
\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}
\]
Power Analysis

The power analysis depends on the relationship between 6 variables:

• the **difference** of biological interest

• the **standard deviation**

• the **significance level** (5%) \((p< 0.05)\) \(\alpha\)

• the **desired power of the experiment** (80%) \(\beta\)

• the **sample size**

• the alternative hypothesis (ie one or two-sided test)
The sample size

• Most of the time, the output of a power calculation.

• The bigger the sample, the bigger the power
  • but how does it work actually?

• In reality it is difficult to reduce the variability in data, or the contrast between means,
  • most effective way of improving power:
    • increase the sample size.
Infinite number of samples

Samples means = $\bar{x}$

Sample size

Population

Continuous variable

Big samples ($n=30$)

Small samples ($n=3$)

Sample means
The sample size

Probability distribution under $H_0$: small samples

- Observed result must be in this range to be significant
- True value = 40
- Significant results: 21% of the time

Probability distribution under $H_0$: big samples

- Observed result must be in this range to be significant
- True value = 40
- Significant results: 90% of the time
The sample size

![Probability distribution under H₀: small samples](image)

- True value = 40
- Significant results: 23% of the time

![Probability distribution under H₀: big samples](image)

- True value = 40
- Significant results: 90% of the time

- $d = 1$
- $n = 1$
- Power = 0.26

- $n = 3$
- Power = 0.53

- $n = 7$
- Power = 0.84
The sample size: the bigger the better?

- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

- What if the tiny difference is meaningless?
  - Beware of *overpower*
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - What is the effect size of biological interest?
Power Analysis

The power analysis depends on the relationship between 6 variables:

• the **effect size** of biological interest
• the **standard deviation**
• the **significance level** (5%)
• the **desired power** of the experiment (80%)
• the **sample size**
• the **alternative hypothesis** (ie one or two-sided test)
The alternative hypothesis: what is it?

- One-tailed or 2-tailed test? One-sided or 2-sided tests?

- Is the question:
  - Is there a difference?
  - Is it bigger than or smaller than?

- Can rarely justify the use of a one-tailed test
- Two times easier to reach significance with a one-tailed than a two-tailed
  - Suspicious reviewer!
Hypothesis

Experimental design

Choice of a Statistical test

Power analysis

Sample size

Experiment(s)

(Stat) analysis of the results
• Fix any five of the variables and a mathematical relationship can be used to estimate the sixth.

e.g. What sample size do I need to have a 80% probability (power) to detect this particular effect (difference and standard deviation) at a 5% significance level using a 2-sided test?
• **Good news:**
there are packages that can do the power analysis for you ... providing you have some prior knowledge of the key parameters!

\[
\text{difference} + \text{standard deviation} = \text{effect size}
\]

• **Free packages:**
  • **R**
  • **G*Power**
  • **InVivoStat**

• **Cheap package:** **StatMate** (~ $95)

• **Not so cheap package:** **MedCalc** (~ $495)
Power Analysis
Let’s do it

• Examples of power calculations:
  • Comparing 2 proportions: Exercise 1
  • Comparing 2 means: Exercise 2
Exercises 1 and 2

• Use the functions below to answer the exercises
  • Clue: exactly one of the parameters must be passed as NULL, and that parameter is determined from the others.

• Use R Help to find out how to use the functions
  • e.g. ?power.prop.test in the console

**Exercise 1**

```r
power.prop.test(n=NULL, p1=NULL, p2=NULL,
  sig.level=NULL, power=NULL, alternative=c("two.sided", "one.sided"))
```

**Exercise 2**

```r
power.t.test(n=NULL, delta=NULL, sd=1, sig.level=NULL, power=NULL,
  type=c("two.sample", "one.sample", "paired"),
  alternative=c("two.sided", "one.sided"))
```
Exercise 1:

- Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa: painting eyes on cows’ bottoms.
- Early trials suggest that lions are less likely to attack livestock when they think they’re being watched
  - Fewer livestock attacks could help farmers and lions co-exist more peacefully.
- Pilot study over 6 weeks:
  - 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

Questions:
- Do you think the observed effect is meaningful to the extent that such a ‘treatment’ should be applied? Consider ethics, economics, conservation ...
- Run a power calculation to find out how many cows should be included in the study.
  - Clue 1: `power.prop.test()`
  - Clue 2: exactly one of the parameters must be passed as NULL, and that parameter is determined from the others.

http://www.sciencealert.com/scientists-are-painting-eyes-on-cows-butts-to-stop-lions-getting-shot
Exercise 1: Answer

- Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa:
  - Painting eyes on the butts of cows
- Early trials suggest that lions are less likely to attack livestock when they think they’re being watched
  - Less livestock attacks could help farmers and lions co-exist more peacefully.

- Pilot study over 6 weeks:
  - 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

```r
power.prop.test(p1 = 3/39, p2 = 0, sig.level = 0.05, power = 0.8, alternative="two.sided")
```

Two-sample comparison of proportions power calculation

- n = 96.92364
- p1 = 0.07692308
- p2 = 0
- sig.level = 0.05
- power = 0.8
- alternative = two.sided

NOTE: n is number in *each* group
Exercise 2:

- Pilot study: 10 arachnophobes were asked to perform 2 tasks:
  - **Task 1**: Group1 (n=5): to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes.
  - **Task 2**: Group 2 (n=5): to look at pictures of the same hairy tarantula.
- Anxiety scores were measured for each group (0 to 100).
- Use R to calculate the values for a power calculation
  - Get the data in R (spider.csv)
  - Hint: you can use `group_by()` and `summarise()`
  - Or you can do it in Excel!
- Run a power calculation (assume balanced design and parametric test)
  - Clue 1: `power.t.test()`
  - Clue 2: choose the sd that makes more sense.
Exercise 2: Answer

spider.data %>%
  group_by(Group) %>
  summarise(mean=mean(Scores), sd=sd(Scores))

<table>
<thead>
<tr>
<th>Group</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture</td>
<td>39</td>
<td>9.617692</td>
</tr>
<tr>
<td>Real</td>
<td>52</td>
<td>9.746794</td>
</tr>
</tbody>
</table>

2 rows

power.t.test(delta = 52 - 39, sd = 9.75, sig.level = 0.05, power = 0.8, type = "two.sample", alternative = "two.sided")

Two-sample t test power calculation

  n = 9.889068
delta = 13
  sd = 9.75
sig.level = 0.05
  power = 0.8
alternative = two.sided

NOTE: n is number in *each* group

• To reach significance with a t-test, providing the preliminary results are to be trusted, and be confident in a difference between the 2 groups, we need about 10 arachnophobes in each group.
Unequal sample sizes

- Scientists often deal with unequal sample sizes
- No simple trade-off:
  - if one needs 2 groups of 30, going for 20 and 40 will be associated with decreased power.
- **Unbalanced design = bigger total sample**
- Solution:
  
  **Step 1**: power calculation for equal sample size
  **Step 2**: adjustment

\[
N = \frac{2n(1+k)^2}{4k}
\]
\[
n_1 = \frac{N}{(1+k)}
\]
\[
n_2 = \frac{kN}{(1+k)}
\]

- **Cow example**: balanced design: \(n = 97\)
  but this time: unpainted group: 2 times bigger than painted one (\(k=2\)):
  - Using the formula, we get a total:
    \[N=2*97*(1+2)^2/4*2 = 219\]
  Painted butts \((n_1)=73\) Unpainted butts \((n_2)=146\)

- **Balanced design**: \(n = 2*97 = 194\)
- **Unbalanced design**: \(n = 70+140 = 219\)
Non-parametric tests

- **Non-parametric tests**: do not assume data come from a Gaussian distribution.
  - Non-parametric tests are based on ranking values from low to high
  - Non-parametric tests almost always less powerful

- Proper power calculation for non-parametric tests:
  - Need to specify which kind of distribution we are dealing with
    - Not always easy

- Non-parametric tests never require more than 15% additional subjects providing that the distribution is not too unusual.

- **Very crude rule of thumb for non-parametric tests**:  
  - Compute the sample size required for a parametric test and **add 15%**.
• What happens if we ignore the power of a test?
  • Misinterpretation of the results

• p-values: never ever interpreted without context:
  • **Significant p-value (<0.05)**: exciting! Wait: what is the difference?
    • $\geq$ smallest meaningful difference: exciting
    • $<$ smallest meaningful difference: not exciting
      • very big sample, too much power

• **Not significant p-value (>0.05)**: no effect! Wait: how big was the sample?
  • Big enough = enough power: no effect means no effect
  • Not big enough = not enough power
    • Possible meaningful difference but we miss it