Day 1
Power Analysis
Anne Segonds-Pichon
v2019-03
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?

- 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
• \( \alpha \): the threshold value that we measure p-values against.
  • For results with 95% level of confidence: \( \alpha = 0.05 \)
  • \( = \) probability of type I error

• **p-value**: probability that the observed statistic occurred by chance alone

• **Statistical significance**: comparison between \( \alpha \) and the p-value
  • p-value \(< 0.05\): reject \( H_0 \) and p-value \( > 0.05 \): fail to reject \( H_0 \)

What does Power look like?
What does Power look like?

- **Type II error ($\beta$)** is the failure to reject a false $H_0$
  - Direct relationship between **Power** and type II error:

  - $\beta = 0.2$ and **Power** = $1 - \beta = 0.8$ (80%)
The desired power of the experiment: 80%

- **Type II error** ($\beta$) is the failure to reject a false $H_0$
  - Direct relationship between **Power** and type II error:
    - if $\beta = 0.2$ and **Power** = $1 - \beta = 0.8$ (80%)
  - Hence a true difference will be missed 20% of the time
  - General convention: 80% but could be more or less

- Cohen (1988):
  - For most researchers: Type I errors are four times more serious than Type II errors: $0.05 \times 4 = 0.2$
  - Compromise: 2 groups comparisons: 90% = +30% sample size, 95% = +60%
In hypothesis testing, a critical value is a point on the test distribution that is compared to the test statistic to determine whether to reject the null hypothesis.

- Example of test statistic: t-value

If the absolute value of your test statistic is greater than the critical value, you can declare statistical significance and reject the null hypothesis.

- Example: t-value > critical t-value

What does Power look like?
To recapitulate:

- The null hypothesis (H₀): H₀ = no effect
- The aim of a statistical test is to reject or not H₀.

<table>
<thead>
<tr>
<th>Statistical decision</th>
<th>True state of H₀</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H₀ True (no effect)</td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I error α</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
</tr>
<tr>
<td>Do not reject H₀</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- Traditionally, a test or a difference are said to be “significant” if the probability of type I error is: α <= 0.05
- High specificity = low False Positives = low Type I error
- High sensitivity = low False Negatives = low Type II error
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 (i.e., one or two-sided test)
The effect size: what is it?

- The **effect size**: minimum meaningful effect of biological relevance.
  - Absolute difference + variability

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

- **Jacob Cohen**
  - Author of several books and articles on power
  - 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>Z 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

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

• The standard deviation of the sample distribution = Standard Error of the Mean: $SEM = \frac{SD}{\sqrt{N}}$
  • SEM decreases as sample size increases

![Diagram showing sample distribution and standard error of the mean](image)
The sample size

A population

![Histogram and scatter plot]
The sample size

Small samples (n=3)

Big samples (n=30)

‘Infinite’ number of samples
Samples means = \bar{X}
The sample size

Small samples (n=3)

Big samples (n=30)

Probability distribution under $H_0$: small samples

Probability distribution under $H_0$: big samples

Observed result must be in this range to be significant

Observed result must be in this range to be significant

True value = 40

Significant results: 21% of the time

True value = 40

Significant results: 90% of the time
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: 23% 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

$d = 1$
$n = 1$

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** (i.e., 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!

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

• Free packages:
  • R
  • G*Power and InVivoStat
  • Russ Lenth's power and sample-size page:
    • [http://www.divms.uiowa.edu/~rlenth/Power/](http://www.divms.uiowa.edu/~rlenth/Power/)

• 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.

```
power.prop.test(p1 = 0.08, p2 = 0, sig.level = 0.05, power = 0.8, alternative="two.sided")
```

```
Two-sample comparison of proportions power calculation

n = 92.99884
p1 = 0.08
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
  - Enter the data in R
  - Hint: you can use `data.frame()` and `apply()`

- Run a power calculation (assume balanced design and parametric test)
  - Clue 1: `power.t.test()`
  - Clue 2: change sd for the one that makes more sense.
Exercise 2: Answer

scores.means <- tapply(spider$Scores, spider$Group, mean)
scores.sd <- tapply(spider$Scores, spider$Group, sd)

stripchart(spider$Scores~spider$Group, vertical=TRUE, method="jitter")
segments(x0 = 1:2 -0.15, y0=scores.means, x1 = 1:2 +0.15)

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

• 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.
• What if the difference is a bit bigger or smaller than expected?
  • difference = \( \delta = 52-39 = 13 \) in the pilot study, let’s go for 10 to 15 in increments of 0.2

\[
difference.values \leftarrow \text{seq}(10, 15, 0.2)
\]
\[
sample.sizes \leftarrow \text{sapply}(difference.values, \text{function}(x) \{
  \text{return(\text{power.t.test}(\delta = x, \text{sd} = 9.75, \text{sig.level} = 0.05, \text{power} = 0.8)$n)}
\})
\]
\[
\text{plot(difference.values, sample.sizes, type = "b", las = 1)}
\]
• What if the difference is the same but the data are noisier than expected?
  
  • $sd = 9.75$ in the pilot study, let’s go for 8 to 12 in increments of 0.2

```r
sd.values <- seq(8, 12, 0.2)
sample.sizes <- sapply(sd.values, function(x) {
  return(power.t.test(delta = 92 - 87.4, sd = x, sig.level = 0.05, power = 0.8)$n)
})
plot(sd.values, sample.sizes, type = "b", las = 1)
```
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: \( n = 93 \)
  - but this time: unpainted group: 2 times bigger than painted one (\( k=2 \)):
  - Using the formula, we get a total:
    \[
    N=2*93*(1+2)^2/4*2 = 210
    \]
  - Painted butts \((n_1)=70\) Unpainted butts \((n_2)=140\)

- **Balanced design**: \( n = 2*93 = 186 \)
- **Unbalanced design**: \( n= 70+140 = 210 \)
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 not 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?
    - >= 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