Power and Sample Size

In epigenetic epidemiology studies
Overview

• Pros and cons

• Working examples

• Concerns for epigenetic epidemiology
• Power is the probability of detecting an effect, given that the effect is really there

• Or likewise, the probability of rejecting the null hypothesis when it is in fact false

• An example;
  • Power of 0.8 = if we performed a study 1000 times, we would see a statistically significant difference 80% of the time
Why perform them

• Ideally:
  • To determine the sample size required to confidently observe an anticipated effect

• Or, at least:
  • To determine if there is sufficient power to detect a meaningful difference in a given sample size

• Required as part of a grant proposal

• Part of planning and designing good quality research
  • Familiarise yourself with the data and study design
  • Implement changes to improve the power and design
Limitations

- They are not universal but depend on;
  - Purpose, methodology, statistical design and procedure

- Provide the minimum number of samples required following the ‘best case scenario’

- Based on statistical assumptions and data characteristics,
  - Which if incorrect (or unknown) will lead to inaccurate estimates

- They are not intuitive;
  - E.g. they may suggest a number of subjects that is inadequate for the statistical procedure

- Hence, power should not be the only consideration when deciding on your sample size
What you need to know

• Core elements
  • Power
  • Sample size
  • Significance
  • Effect size*

• These elements are all inter-related such that;
  • If you know three you can estimate the fourth
  • Manipulating one influences the others
A note on effect size

- There are many ways to define and calculate effect size
  - Difference in means
  - Variance explained
  - Odds ratio

- Standardised vs. unstandardised measures
  - If possible use unstandardized measures
    - Raw difference between group means
    - Raw regression coefficients
  - Use standardised effect sizes as a last resort
    - Standardised difference (d): difference in means/SD
    - Pearson’s correlation coefficient (r)

<table>
<thead>
<tr>
<th>Effect</th>
<th>d</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>≥0.2</td>
<td>≥0.1</td>
</tr>
<tr>
<td>Medium</td>
<td>≥0.5</td>
<td>≥0.3</td>
</tr>
<tr>
<td>Large</td>
<td>≥0.8</td>
<td>≥0.5</td>
</tr>
</tbody>
</table>
Deciding on levels of $\alpha$ and $\beta$

- **Power (sensitivity) $[1-\beta]$**
  - Probability of finding a true effect when one does exist
  - **Type 2 error $[\beta]$**: incorrectly accepting the null hypothesis (false negative)
  - Aim to minimise the risk of failing to detect a real effect
  - Typical values for power are 80%, 90% and 95%

- **Significance (p-value) $[\alpha]$**
  - Probability that an effect occurred by chance alone
  - **Type 1 error $[\alpha]$**: incorrectly rejecting the null hypothesis (false positive)
  - Aim to minimise the risk of detecting a non-real/spurious effect
  - Typical values are 0.05, 0.01

- **Reducing the risk of type 1 error $\Leftrightarrow$ increased risk of type 2 error (i.e. reduced power)**
Available Software

• Standard statistical packages
  • Stata, Minitab, SPSS Sample Power, R

• Online web calculators

• Freely available software
  • G*Power [wwwpsychouni-duesseldorfdetaabteilungsaapgpower3](http://www.psycho.uni-duesseldorfdetaabteilungsaapgpower3)
  • Quanto [http://hydra.usc.edu/gxe](http://hydra.usc.edu/gxe)

• Different packages only perform specific power calculations so you will need to find one relevant to the statistical model you are planning
Example of an independent two-sample t-test

- Breitling et al, AJHG 2011
- CpG site mapping to F2RL3 was associated with smoking behaviour
- Average methylation in smokers was 83% compared to 95% in never smokers

How many samples do we need to detect this effect?

- Power = 90%
- Significance = 0.05
- Methylation characteristics: means = 83% & 95%, SD = 10%
  - Effect size: \( \frac{\text{difference in means}}{\text{SD}} = \frac{95-83}{10} = 1.2 \)
1. Select the statistical test
2. Select the type of power analysis
3. Input the data characteristics to determine the effect size
4. Input power parameters
F2RL3 methylation & smoking
5. Draw plot for a range of values
5. Draw plot for a range of values
5. Draw plot for a range of values
6. Produce of table of values
Dialog box
1. Select statistical test and input data characteristics
2. Input power parameters
F2RL3 methylation & smoking

Estimated sample size for two-sample comparison of means
Test Ho: m1 = m2, where m1 is the mean in population 1
and m2 is the mean in population 2

Assumptions:
alpha = 0.0500 (two-sided)
power = 0.9000
m1 = .95
m2 = .83
sd1 = .1
sd2 = .1
n2/n1 = 1.00

Estimated required sample sizes:
n1 = 15
n2 = 15

Sample size requirements
(Power = 0.9, Alpha = 0.05)
Estimated power for two-sample comparison of means

Test Ho: \( m_1 = m_2 \), where \( m_1 \) is the mean in population 1
and \( m_2 \) is the mean in population 2

Assumptions:
- \( \alpha = 0.0500 \) (two-sided)
- \( m_1 = 0.9 \)
- \( m_2 = 0.83 \)
- \( sd_1 = 0.15 \)
- \( sd_2 = 0.2 \)
- \( n_2/n_1 = 1.00 \)

Estimated power:
- \( power = 0.7996 \)
Other statistical tests

- **G* Power**
  - Correlations & regressions (univariate, multiple variate, logistic)
  - Means (one, two, many groups, un/paired, non-parametric)
  - Proportions (one, two groups, un/paired)
  - Variances (one, two groups)

- **STATA**
  - sampsi (one, two groups, un/paired, means, proportions)
  - fpower (one-way anova)
  - powerreg (regression)
Challenges

Non-normality of DNA methylation data

• Can try transform the data
  • Popular transformations don’t always work
  • They make interpretation of results more difficult
  • Transformations that modify the data too much can actually lose more power

• Can categorise the data
  • Requires more samples given less power

• Can perform non-parametric tests
  • Few programs perform non-parametric power calculations
  • Those that do still assume the data is normally distributed
Non-normality of DNA methylation data

“there is minimal power loss associated with the non-parametric tests even when the data are distributed normally, while the power gains of these tests when normality is violated are substantial” (Kitchen, Am J Ophthalmol. 2009)

• If the data is normal, the Mann-Whitney test has been estimated to be ~0.96 times as powerful as the t-test

• If the data isn’t normal, the Mann-Whitney test is more powerful than the t-test

• [Therefore, if you have enough power for a t-test, you'll have enough power for a Mann-Whitney.]
Challenges

Lack of prior knowledge

• Public databases
  • Detailing spectrum of genome-wide DNA methylation distributions across multiple populations, ages, tissues, and cells are not yet available

• Literature
  • Specific sample & tissue populations may be different
  • Don't always give the relevant data characteristics

• Pilot data
  • Not always possible
  • [Small sample sizes in pilot data can be misleading]
Multiple testing

- Alpha inflation = the more tests you perform, the more likely you are to see a false-positive effect (Type 1 error)
- P-value ($\alpha$) = probability that an effect occurred by chance
  - P-value of 0.05 = 5% of all tests performed (or 1 in 20)
- Could perform alpha adjustments
  - Bonferroni correction [0.05/number of tests]
  - Genome-wide significance is estimated at $p = 10^{-6}$ and $10^{-8}$ for GWAS
  - Overly stringent
  - Limited data regarding correlation across genome-wide CpG sites
Best practice

• Perform a range of power calculations covering a range of scenarios

<table>
<thead>
<tr>
<th>ABSOLUTE CHANGE IN METHYLATION</th>
<th>TWO GROUP COMPARISON (N=280 VS. 990)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD of methylation = 8%</td>
<td>POWER FOR P=1.0x10^-8</td>
</tr>
<tr>
<td>3%</td>
<td>0.4245</td>
</tr>
<tr>
<td>4%</td>
<td>0.9512</td>
</tr>
<tr>
<td>5%</td>
<td>0.9998</td>
</tr>
<tr>
<td>10%</td>
<td>1.0000</td>
</tr>
<tr>
<td>SD of methylation = 10%</td>
<td></td>
</tr>
<tr>
<td>4%</td>
<td>0.5710</td>
</tr>
<tr>
<td>5%</td>
<td>0.9512</td>
</tr>
<tr>
<td>10%</td>
<td>1.0000</td>
</tr>
<tr>
<td>SD of methylation = 12%</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.6646</td>
</tr>
<tr>
<td>10%</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Blurb:** There is more than 66% power to detect an absolute change in methylation of 5% between the two groups (n=280 vs 990), assuming the variance in methylation is 12% or less, at the $p \times 10^{-8}$ significance level. (Power calculations were performed in STATA and are based on standard unpaired t-tests, which assume normality of data and equal variances between groups.)
Summary

• Unfortunately there is no single solution
• They are only an estimation, and even then, of the “best case scenario”
• They should not be the only factor involved when deciding on sample size
• Try a range of scenarios and consider other factors, e.g.
  • The purpose of your study
  • Potential errors in the model parameters
  • Restraints on the statistical model
  • Common sense
• Power calculations are a good thing to do when planning your study.
  • Help you to familiarise yourself with your data and study design
  • Enable you to identify any limitations
  • Implement changes to your study design in order to get the best out of it
References

Jacob Cohen
Cohen, Statistical Power Analysis for the Behavioral Sciences, 1988

Russell V. Lenth


G*Power www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/