Statistical experimental design principles for biological studies

Alexander (Alec) Zwart
10 July 2014
An apology, and a caveat

• Apology: I am not a genetics/genomics expert!
  • But I DO know a bit about design...

• Caveat: Experimental design is a science! (as is statistical analysis...)
  • Every experiment is different
  • This talk is necessarily simplistic, and very incomplete...
  • Intended as a reminder of some concepts, some bits of advice, and a plea...
An experiment – recap:

- Apply treatments (experimental conditions) or treatment combinations to subjects (or experimental units).
- Carefully control/remove other influences as much as possible.
- Observe one or more responses (observed or measured results)

- A way to infer cause => effect.

- Genotype/Species/Variety is a treatment? Yes...
What is statistical experimental design concerned with?

• Everything! The entire process leading to the final dataset...

• Recognising sources of variation (systematic, random) that influence your results and hence controlling for them

• How to ensure that we obtain data that provides ‘honest’ and accurate answers to our questions
Key point:

• Good statistics cannot rescue a bad dataset!

• And a bad dataset may not be able to answer your questions (honestly, that is)

• Do the experiment right => get a good dataset

• Consider design from the very beginning of planning the experiment
Fundamentals

Randomisation

Blocking

Replication
Randomisation

• Allocate treatments to subjects **randomly** (ideally using a proper random number generator, not your own personal judgement)

• Also randomise:
  • Spatial layout
  • Order of processing

• **Avoid systematic patterns** (But: see Blocking)

  • Protects against **unforeseen** biases
Blocking

• Identify groupings that may effect your results
  • Choice of operator/technician/machine
  • Cage/Table/shelf/assay plate/batch

• Try to assign (ideally) one of each treatment to each group! (Randomised Block Design or RBD)
  • (Example coming later...)
Replication

• Sample size
  • G*Power ([http://www.gpower.hhu.de/en.html](http://www.gpower.hhu.de/en.html))
  • Russ Lenth’s applets ([http://homepage.stat.uiowa.edu/~rlenth/Power/](http://homepage.stat.uiowa.edu/~rlenth/Power/))
  • Simulation studies in the literature? (Google scholar)

• Biological/internal/technical replicates
  • Don’t confuse these – Internal/technical replicates are not a substitute for biological replication.
  • (more later...)
12 mice, three treatments (Colour)

Completely randomised design
Balanced (equal info on each treatment, treatments compared with equal accuracy)
But, the mice must be housed...

Here, treatments are **confounded** with cages

Cannot separate Cage effects from Trt effects!
Cages as Blocks
Animals (especially) can pose a problem...
Cages as Experimental Units / Subjects...

...& Cage = Mouse
(Aside-env. trend across columns of cages?)

=> Block by Columns!
Not enough cages? A compromise:

But, Cages are the Experimental units

Only two replicates of each Trt!
More on replication:

• The replication used determines the population(s) being sampled...

  ...hence...

• The replication used determines the conclusions that may be validly drawn from a statistical test
Comparison between populations of plants

<table>
<thead>
<tr>
<th>Wild Type</th>
<th>Mutant</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Wild Type Images]</td>
<td>![Mutant Images]</td>
</tr>
<tr>
<td>2.1 1.4 1.9 1.6</td>
<td>4.3 3.9 4.1 4.4</td>
</tr>
</tbody>
</table>
Comparisons between two plants

Wild Type

Mutant

2.1  1.9  1.6

4.3  3.9  4.1  4.4
Recording

• Details of treatments and associated randomisation, blocking, types of replication should always be recorded

• Check what information is required for the analysis, or talk to the statistician about what is required.

• This is important meta data when distributing your data...
More on Blocking...

• ‘One of each treatment per block’ is not always feasible – blocks may be too small
  • ‘Incomplete blocks’

• Think about:
  • Balance – treatments occur together in the same blocks the same number of times.
  • Linkage – common treatments in different blocks provide links between blocks – ensure you have plenty of linkage.
Balanced Incomplete Block Design (BIBD)

- Trts occur together the same number of times
- Plenty of treatment-linkages between blocks
Notes on incomplete blocks

• BIBDs are only available for certain combinations of parameters.
• However, the general principles are still relevant
  • ‘Unbalanced incomplete blocks’
• Note that teasing apart the treatment effects from block effects in unbalanced situations is dependent on the capabilities of the analysis method
  • Check on said capabilities, consider simpler experiment if necessary...
### Treatment structure

- Sir Ronald Fisher => factorial treatment structure

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Environment</th>
<th>I</th>
<th>II</th>
<th>II</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

(3x4 factorial)

- 5 = Number of replicates of each treatment
Factorial treatment structure

• Separate **main effects** and **interactions**
• Can include more factors (in theory)
• Unbalanced numbers of treatments usually OK:

![Factorial Treatment Structure Table]

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
</tr>
</tbody>
</table>
But beware **missing** treatment combinations

- The more that entire treatments are missing from the factorial, the harder it is to tease apart main effects and interactions and the more **confounded** different treatment effects become.
Don’t get too ambitious!

• Two-way factorials are much easier to interpret than three-way, four way... etc.
• The higher the factorial structure, the more likely the structure will be incomplete
  • (missing values, treatment combos which don’t make sense, treatment combos you aren’t interested in and don’t include).
• Statisticians and their tools can’t work miracles!
• Simpler is better...
• Talk to a statistician!
Plea:

• If you don’t already know how to design and conduct good experiments...

...then...

• ...involve a statistician in the early stages of planning your experiment!

• And note – the statistician may need time...
Sir Ronald Fisher, ‘the father of modern statistics’…

“To consult the statistician after an experiment is finished is often merely to ask them to conduct a post mortem examination.

They can perhaps say what the experiment died of.”

(R. Fisher)
Thank you