Experimental design, basic statistics, and sample size determination

Karl W Broman
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

http://www.biostat.jhsph.edu/~kbroman

Experimental design
## Basic principles

1. Formulate question/goal in advance
2. Comparison/control
3. Replication
4. Randomization
5. Stratification (aka blocking)
6. Factorial experiments

## Example

**Question:** Does salted drinking water affect blood pressure (BP) in mice?

**Experiment:**
1. Provide a mouse with water containing 1% NaCl.
2. Wait 14 days.
3. Measure BP.
Comparison/control

Good experiments are comparative.

- Compare BP in mice fed salt water to BP in mice fed plain water.
- Compare BP in strain A mice fed salt water to BP in strain B mice fed salt water.

Ideally, the experimental group is compared to concurrent controls (rather than to historical controls).
Why replicate?

- Reduce the effect of uncontrolled variation (i.e., increase precision).
- Quantify uncertainty.

A related point:
An estimate is of no value without some statement of the uncertainty in the estimate.

Randomization

Experimental subjects ("units") should be assigned to treatment groups at random.

At random does not mean haphazardly.

One needs to explicitly randomize using
- A computer, or
- Coins, dice or cards.
Why randomize?

• Avoid bias.
  – For example: the first six mice you grab may have intrinsically higher BP.

• Control the role of chance.
  – Randomization allows the later use of probability theory, and so gives a solid foundation for statistical analysis.

Stratification

• Suppose that some BP measurements will be made in the morning and some in the afternoon.
• If you anticipate a difference between morning and afternoon measurements:
  – Ensure that within each period, there are equal numbers of subjects in each treatment group.
  – Take account of the difference between periods in your analysis.
• This is sometimes called “blocking”.

Example

• 20 male mice and 20 female mice.
• Half to be treated; the other half left untreated.
• Can only work with 4 mice per day.

Question: How to assign individuals to treatment groups and to days?

An extremely bad design

<table>
<thead>
<tr>
<th>Week One</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Week Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Tu</td>
<td>W</td>
<td>Th</td>
<td>F</td>
<td>M</td>
<td>Tu</td>
<td>W</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

T = treated, C = control, pink = female, blue = male
Randomized

A stratified design
Randomization and stratification

• If you can (and want to), fix a variable.
  – e.g., use only 8 week old male mice from a single strain.

• If you don’t fix a variable, stratify it.
  – e.g., use both 8 week and 12 week old male mice, and stratify with respect to age.

• If you can neither fix nor stratify a variable, randomize it.

Factorial experiments

Suppose we are interested in the effect of both salt water and a high-fat diet on blood pressure.

Ideally: look at all 4 treatments in one experiment.

Plain water   Normal diet
Salt water    High-fat diet

Why?
  – We can learn more.
  – More efficient than doing all single-factor experiments.
Interactions

Additive

Interactive

Other points

• Blinding
  – Measurements made by people can be influenced by unconscious biases.
  – Ideally, dissections and measurements should be made without knowledge of the treatment applied.

• Internal controls
  – It can be useful to use the subjects themselves as their own controls (e.g., consider the response after vs. before treatment).
Other points

- Representativeness
  - Are the subjects/tissues you are studying really representative of the population you want to study?
  - Ideally, your study material is a random sample from the population of interest.

Summary

Characteristics of good experiments:

- Unbiased
  - Randomization
  - Blinding
- High precision
  - Uniform material
  - Replication
  - Stratification
- Simple
  - Protect against mistakes
- Wide range of applicability
  - Deliberate variation
  - Factorial designs
- Able to estimate uncertainty
  - Replication
  - Randomization
Basic statistics

What is statistics?

“We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.”

— Sir R. A. Fisher
What is statistics?

- Data exploration and analysis
- Inductive inference with probability
- Quantification of uncertainty

Example

- We have data on the blood pressure (BP) of 6 mice.
- We are not interested in these particular 6 mice.
- Rather, we want to make inferences about the BP of all possible such mice.
Sampling

BP of all possible mice

true average

Blood pressure

BP of 6 mice

Blood pressure

Several samples

BP of all possible mice

true average

Blood pressure

Several possible samples
We observe the BP of 6 mice, with average = 103.6 and standard deviation (SD) = 9.7.

We assume that BP in the underlying population follows a normal (aka Gaussian) distribution.

On the basis of these data, we calculate a 95% confidence interval (CI) for the underlying average BP:

\[
103.6 \pm 10.2 = (93.4 \text{ to } 113.8)
\]
What is a CI?

• The plausible values for the underlying population average BP, given the data on the six mice.
• In advance, there is a 95% chance of obtaining an interval that contains the population average.
CI for difference

BP of all possible mice

BP of 6 mice per group

Blood pressure

Blood pressure

95% CI for treatment effect = 12.6 ± 11.5

Significance tests

Confidence interval:
The plausible values for the effect of salt water on BP.

Test of statistical significance:
Answer the question, “Does salt water have an effect?”

Null hypothesis (H₀):
Salt water has no effect on BP.

Alt. hypothesis (Hₐ):
Salt water does have an effect.
Two possible errors

- Type I error ("false positive")
  Conclude that salt water has an effect on BP when, in fact, it does not have an effect.

- Type II error ("false negative")
  Fail to demonstrate the effect of salt water when salt water really does have an effect on BP.

Type I and II errors

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>No effect</th>
<th>Has an effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
<td>✓</td>
</tr>
<tr>
<td>Fail to reject $H_0$</td>
<td>✓</td>
<td>Type II error</td>
</tr>
</tbody>
</table>
Conducting the test

- Calculate a test statistic using the data. (For example, we could look at the difference between the average BP in the treated and control groups; let’s call this D.)

- If this statistic, D, is large, the treatment appears to have some effect.

- How large is large?
  - We compare the observed statistic to its distribution if the treatment had no effect.

Significance level

- We seek to control the rate of type I errors.

- Significance level (usually denoted $\alpha$) = chance you reject $H_0$, if $H_0$ is true; usually we take $\alpha = 5\%$.

- We reject $H_0$ when $|D| > C$, for some C.

- C is chosen so that, if $H_0$ is true, the chance that $|D| > C$ is $\alpha$. 
If salt has no effect

BP of all possible mice

Plain water

Salt water

Blood pressure

Diff. between group aves.

2.5%

2.5%

Change in ave. BP

If salt has an effect

BP of all possible mice

Plain water

Salt water

Blood pressure

Diff. between group aves.
P-values

• A P-value is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (i.e., if the treatment has no effect).

• If your P-value is smaller than your chosen significance level (\(\alpha\)), you reject the null hypothesis.

• We seek to reject the null hypothesis (we seek to show that there is a treatment effect), and so small P-values are good.

Summary

• Confidence interval
  – Plausible values for the true population average or treatment effect, given the observed data.

• Test of statistical significance
  – Use the observed data to answer a yes/no question, such as “Does the treatment have an effect?”

• P-value
  – Summarizes the result of the significance test.
  – Small P-value \(\alpha\) conclude that there is an effect.

Never cite a P-value without a confidence interval.
Data presentation

Good plot

Bad plot

Data presentation

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean</th>
<th>(SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11.3</td>
<td>(0.6)</td>
</tr>
<tr>
<td>B</td>
<td>13.5</td>
<td>(0.8)</td>
</tr>
<tr>
<td>C</td>
<td>14.7</td>
<td>(0.6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean</th>
<th>(SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11.2965</td>
<td>(0.63)</td>
</tr>
<tr>
<td>B</td>
<td>13.49</td>
<td>(0.7913)</td>
</tr>
<tr>
<td>C</td>
<td>14.727</td>
<td>(0.6108)</td>
</tr>
</tbody>
</table>
Sample size determination

Fundamental formula

\[ n = \frac{\$ \text{ available}}{\$ \text{ per sample}} \]
Listen to the IACUC

Too few animals  a total waste

Too many animals  a partial waste

Significance test

- Compare the BP of 6 mice fed salt water to 6 mice fed plain water.
- : true difference in average BP (the treatment effect).
- \( H_0: \bar{D} = 0 \) (i.e., no effect)
- Test statistic, \( D \).
- If \(|D| > C\), reject \( H_0 \).
- \( C \) chosen so that the chance you reject \( H_0 \), if \( H_0 \) is true, is 5%
Statistical power

Power = The chance that you reject $H_0$ when $H_0$ is false (i.e., you [correctly] conclude that there is a treatment effect when there really is a treatment effect).

Power depends on...

- The structure of the experiment
- The method for analyzing the data
- The size of the true underlying effect
- The variability in the measurements
- The chosen significance level ($\alpha$)
- The sample size

Note: We usually try to determine the sample size to give a particular power (often 80%).
Effect of sample size

6 per group:

Power = 70%

12 per group:

Power = 94%

Effect of the effect

$\Delta = 8.5$:

Power = 70%

$\Delta = 12.5$:

Power = 96%
A formula

\[ n = \frac{2}{\hat{D}} \]

Various effects

- Desired power \( \uparrow \) \( \Rightarrow \) sample size \( \uparrow \)
- Stringency of statistical test \( \uparrow \) \( \Rightarrow \) sample size \( \uparrow \)
- Measurement variability \( \downarrow \) \( \Rightarrow \) sample size \( \downarrow \)
- Treatment effect \( \uparrow \) \( \Rightarrow \) sample size \( \downarrow \)
Determining sample size

The things you need to know:

- Structure of the experiment
- Method for analysis
- Chosen significance level, $\alpha$ (usually 5%)
- Desired power (usually 80%)
- Variability in the measurements
  - if necessary, perform a pilot study, or use data from prior publications
- The smallest meaningful effect

Reducing sample size

- Reduce the number of treatment groups being compared.
- Find a more precise measurement (e.g., average time to effect rather than proportion sick).
- Decrease the variability in the measurements.
  - Make subjects more homogeneous.
  - Use stratification.
  - Control for other variables (e.g., weight).
  - Average multiple measurements on each subject.
Final conclusions

• Experiments should be designed.
• Good design and good analysis can lead to reduced sample sizes.
• Consult an expert on both the analysis and the design of your experiment.

Resources

  – An excellent introductory text.
  – Includes a more advanced treatment of experimental design.
• Course: *Statistics for Laboratory Scientists* (Biostatistics 140.615-616, Johns Hopkins Bloomberg Sch. Pub. Health)
  – Introductory statistics course, intended for experimental scientists.
  – Greatly expands upon the topics presented here.