Designing Efficient Sampling Plans for Ecology and Conservation

Edward F. Connor
Department of Biology
San Francisco State University
Issues in Study Design

- Sampling intensity (how much to sample)
- Choosing among different sampling designs (between and within-subjects designs)
- Determining sampling frequency (how often to sample)
- Determining which study design is most cost effective
Choosing a Sample Size

• Obtain a preliminary estimate of $\sigma$
• Conjecture the minimally interesting difference in treatment means, or effect size, $\delta$
• Set the Type I error rate, $\alpha$
• Set the desired power for the test ($1 - \beta$)
• Calculate necessary sample size by approximation, web-based calculators, software, or simulation
Figure 2 - Distribution of $t$ under $H_0$ and $H_a$.
## Recommended Calculators for Sample Size Determination

### Web Based

<table>
<thead>
<tr>
<th>Author</th>
<th>Location</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. Lenth</td>
<td>University of Iowa</td>
<td><a href="http://www.stat.uiowa.edu/~rlenth/Power/index.html">http://www.stat.uiowa.edu/~rlenth/Power/index.html</a></td>
</tr>
</tbody>
</table>

### Software

<table>
<thead>
<tr>
<th>Author</th>
<th>Package</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>W. Dupont &amp; W. Plummer</td>
<td>PS</td>
<td><a href="http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize">http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize</a></td>
</tr>
<tr>
<td>R. O’Brien</td>
<td>UnifyPow</td>
<td><a href="http://www.bio.ri.ccf.org/power.html">http://www.bio.ri.ccf.org/power.html</a></td>
</tr>
<tr>
<td>J. Hintze</td>
<td>PASS</td>
<td><a href="http://www.ncss.com/pass.html">http://www.ncss.com/pass.html</a> (for purchase)</td>
</tr>
</tbody>
</table>
Choosing an Experimental Design

ANOVA Designs
- Between Subjects*
- Within Subjects*

Regression Designs
- Between Subject

Mixed Model Designs
- Between Subject
- Within Subject
- Covariates
Difference between Separate Groups and Paired t-test

Separate Groups t-test

\[
t = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{\sqrt{S^2_{\bar{x}_a-\bar{x}_b}}}
\]

\[
df = 2(n-1)
\]

Paired t-test

\[
t = \frac{\bar{d} - u}{\sqrt{S^2_d}}
\]

\[
s^2_d = s^2_{\bar{x}_a-\bar{x}_b} = \frac{s^2_d}{n} = \frac{s^2_a}{n} + \frac{s^2_b}{n} - \frac{2r_{ab}S_aS_b}{n}
\]

\[
df = (n-1)
\]
Paired versus Separate Groups \( t \)-test

\[
\alpha = 0.05, \quad \beta = 0.2
\]

\( k = 2 \)

\( \delta/\sigma = 0.5 \)

\( \delta/\sigma = 0.75 \)

\( \delta/\sigma = 1.0 \)

\( \delta/\sigma = 1.25 \)

\( \delta/\sigma = 1.5 \)
Estimating flowering plant species richness in upland regions and seasonal stream channels in Freeman Meadow (Lakes Basin, California)
Preliminary Data Collection
Meadow Example

WS= Within Subjects
BS = Between Subjects
O = out of stream channel
I = In stream channel
Sample Size Calculations for Meadow Example
Using PASS

Within Subjects

Power vs N with Mean0=0.5 Mean1=2.5 S=0.7
Alpha=0.05 T Test

<table>
<thead>
<tr>
<th>N</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Sample size with Power = 0.8

Number of measurements required = 6

Between Subjects

Power vs N1 with M1=0.5 M2=2.5 S1=1.8 S2=1.8
Alpha=0.05 N2=N1 1-Sided T Test

<table>
<thead>
<tr>
<th>N1</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Sample size with Power = 0.8

Number of measurements required = 22
Difference between Within-subjects and Between-subjects ANOVA

Between Subjects

\[
F = \frac{MS_{treat}}{MS_{error}} \quad dfs = \frac{k - 1}{k(n - 1)} \quad MS_{error} = \bar{\text{var}}
\]

Within Subjects

\[
F = \frac{MS_{treat}}{MS_{error}} \quad dfs = \frac{k - 1}{(n - 1)(k - 1)} \quad MS_{error} = \bar{\text{var}} - \bar{\text{cov}}
\]

\[
r' = \frac{\bar{\text{cov}}}{\bar{\text{var}}} \quad MS_{error} = \bar{\text{var}}(1 - r')
\]
Within versus Between Subjects ANOVA

\[ k = 3 \]
\[ \varepsilon = 1 \]

\[ \frac{\delta}{\sigma} = 0.5 \]
\[ \frac{\delta}{\sigma} = 0.75 \]
\[ \frac{\delta}{\sigma} = 1.0 \]
\[ \frac{\delta}{\sigma} = 1.25 \]
\[ \frac{\delta}{\sigma} = 1.5 \]

\[ k = 4 \]
\[ \varepsilon = 1 \]

\[ \frac{\delta}{\sigma} = 0.5 \]
\[ \frac{\delta}{\sigma} = 0.75 \]
\[ \frac{\delta}{\sigma} = 1.0 \]
\[ \frac{\delta}{\sigma} = 1.25 \]
\[ \frac{\delta}{\sigma} = 1.5 \]

\[ \alpha = 0.05, \beta = 0.2 \]
Advantage of Within-subjects Design

For modest levels of correlation within a subject, *within subjects designs* will be more powerful than the equivalent between subjects design.

When testing the same $H_o$ against the same $H_a$ ($\delta$ fixed), with the same $\alpha$, $\beta$, *within subjects designs* are more efficient designs since $\sigma$ will most likely be sufficiently smaller.
Northern Spotted Owl
Northern Spotted Owl
Point Reyes National Seashore, Muir Woods and Golden Gate National Recreation Area


Fledglings/nest

1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

1999 2000 2001 2002 2003 2004 2005

Year
Spotted Owl Fecundity Sample Size Calculation
Effect of varying sampling frequency

Effect of Sampling Frequency ($\alpha = 0.05$)

Effect of Sampling Frequency ($\alpha = 0.20$)

10% Annual Decline

Power vs. Sample Size ($n$) for different sampling frequencies and sample sizes.
Trade off between Sampling Intensity and Frequency

- Achieve approximately equal power using fewer sites sampled more years
- Or more sites sampled fewer years
- For Spotted Owl more cost effective to sample more sites fewer years (20/yr x 12 yrs = 240 surveys, 25/yr x 4 yrs = 100 surveys)
In Search of the Optimal Quadrat Size and Shape – What is Optimal?

- Sampling variance
- Sampling time/replicate
- Transit time/replicate
- Minimizes bias
Field effort = \((p \times n) + (d \times t)\)

**p** = processing time per quadrat  
**n** = sample size (number of quadrats)  
**d** = total distance traveled between quadrats  
**t** = rate of movement between quadrats
Transit rate through vegetation (1.1125 m/sec)

Processing time for each size quadrat

Estimate $\sigma$ from simulated population
Use $\sigma$ to calculate estimate of required sample size

For estimated $n$ determine shortest distance between quadrats
Field effort = \((p \times n) + (d \times t)\)

- \(p\) = processing time per quadrat
- \(n\) = sample size (number of quadrats)
- \(d\) = total distance traveled between quadrats
- \(t\) = rate of movement between quadrats
Field

- Using a within-subjects design (n = 4) we estimated the time required to set-up and sample a variety of quadrat sizes and shapes (p).
- We estimated the time required to walk a variety of distances through vegetation on a fixed bearing (t).
Time Costs for Different Size and Shape Quadrats – Carmen Valley, CA
• We estimated the between quadrat variance in plant density for each size and shaped quadrat for each simulated plant population

• We used the estimated variance to calculate the sample size (n) required to place a (1 - α) confidence interval on the mean population density that was no greater than $\pm 0.3 \times \mu$
Simulated plant populations with a mean density of 0.8 plants/m² in a 50 x 100 m region with different spatial structures

Aggregation

Random = 1  Low aggregation = 5  High aggregation = 10
Shape - Rectangularity

Length/Width
Square = 1
0.25 x 50m = 200

Rotation
0° along gradient
90° perpendicular to gradient
From Zar p. 105

\[
P\left(\bar{x} - t_{(1-\alpha/2,n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(1-\alpha/2,n-1)} \frac{s}{\sqrt{n}} \right) = 1 - \alpha
\]

\[
n = \frac{s^2 t^2_{(1-\alpha/2,n-1)}}{d^2}
\]

\[
n = \frac{s^2 t^2_{(1-\alpha/2,n-1)}}{0.3\bar{x}^2}
\]
Conditional Probability of having the correct width interval \( (W) \), given that it is a \( 1 - \alpha \) interval \( (V) \) (Jiroutek et al 2003)

\[
Pr(W | V) = Pr\left\{ \left[ (U - L) \leq \delta \right] \mid (L \leq \theta \leq U) \right\} = \frac{Pr(W \cap V)}{Pr(V)}
\]

Where

\[
Pr(V) \geq \{1 - \alpha\}
\]

\[
Pr(W) = \{(U - L) \leq \delta\}
\]

• Let \( U \) and \( L \) indicate the upper and lower confidence interval bounds, respectively

• Let \( \delta \) indicate the desired Confidence Interval width
Pr(W | V) ≥ \int_0^{x_1} \left[ \Phi \left( c_1 \sqrt{x} \right) - \Phi \left( -c_1 \sqrt{x} \right) \right] \frac{f_{x^2}(x; \nu_e)}{(1 - \alpha)} \, dx

Where

\[ \Phi(\bullet) \] indicates the CDF of a standard normal variate

\[ f_{x^2}(x; \nu_e) \] indicates a central chi-square density function with \( \nu_e \) degrees of freedom

\[ c_1 = \frac{\sqrt{F_{\text{crit}}}}{\nu_e} \]

\[ x_1 = \frac{\nu_e \delta^2}{\left( 4 \sigma^2 F_{\text{crit}} \right) \sqrt{N}} \]

\[ F_{\text{crit}} \]

\[ \nu_e = N - r \]

\[ N \]

indicates the critical value from a central \( F \) distribution with 1 numerator and \( \nu_e \) denominator degrees of freedom

where \( N \) indicates the estimated sample size and \( r = 2 \) for a two-sided test.
Nominal v/m = 1
Gradient = 0
Rotation = 0

Effort (hrs)

Rectangularity (L/W)

Area
Nominal $v/m = 5$
Gradient = 0
Rotation = 0
Nominal v/m = 10
Gradient = 0
Rotation = 0

Effort (hrs)

Area

Rectangularity (L/W)

4 x 50m
Nominal v/m = 1
Gradient = steep
Rotation = 0

Effort (hrs) vs. Rectangularity (L/W) vs. Area

0.25 x 1m
0.25 x 50m
0.5 x 50m
1 x 50m
2 x 50m
4 x 50m
4 x 50m
Nominal v/m = 1
Gradient = steep
Rotation = 15
Nominal v/m = 1
Gradient = steep
Rotation = 30

1 x 5m
Summary

- Preliminary data is essential to designing cost effective sampling programs.
- Within-subject and reduced-frequency sampling designs are generally more cost effective.
- Field data combined with computer simulations can provide concrete guidance on sampling methods and study design.
Think and calculate before you go to the field
Acknowledgements

Lucas Bohnett   Anastasia Chavez   Jen Carah
Numerous students from Biology 315 - Field Methods in Ecology