Sample Size Estimation for Research Grant Applications and Institutional Review Boards

Will G Hopkins · AUT University · Auckland NZ
Co-presented at the 2008 annual meeting of the American College of Sports Medicine by Stephen W Marshall, University of North Carolina

Background
Sample Size for Statistical Significance: How It Works
Sample Size for Clinical Outcomes: How It Works
Sample Size for Precise Estimates: How It Works
General Issues
Sample size in other studies; smallest effects; big effects, on the fly and suboptimal sizes; design, drop-outs, confounding; validity and reliability; comparing groups; subgroup comparisons and individual differences; mixing unequal sexes; multiple effects; case series; single subjects; measurement studies; simulation

Conclusions
Click on these topics to link to the slides.

Sample Size for Statistical Significance

- In this old-fashioned approach, you decide whether an effect is 'real'; that is, statistically significant (non-zero).
  - If you get significance and you're wrong, it's a false-positive or Type I statistical error.
  - If you get non-significance and you're wrong, it's a false negative or Type II statistical error.
  - The defaults for acceptably low error rates are 5% and 20%.
  - The false-negative rate is for the smallest important value of the effect, or the "minimum clinically important difference".
  - Solve for the sample size by assuming a sampling distribution for the effect.

Sample Size for Clinical Outcomes

- In the first new approach, the decision is about whether to use the effect in a clinical or practical setting.
  - If you decide to use a harmful effect, it's a false-positive or Type I clinical error.
  - If you decide not to use a beneficial effect, it's a false-negative or Type 2 clinical error.
  - Suggested defaults for acceptable error rates are 0.5% and 25%.
  - Benefit and harm are defined by the smallest clinically important effects.
  - Solve for the sample size by assuming a sampling distribution.
  - Sample sizes are ~1/4, those for statistical significance.
  - The traditional approach is too conservative?
    - P=0.05 with the traditional sample size implies one chance in a million of the effect being harmful.

Sample Size for Statistical Significance: How It Works

- The Type I error rate (5%) defines a critical value of the statistic.
  - If observed value > critical value, the effect is significant.
  - If observed value < critical value, the effect is not significant.
  - When true value = smallest important value, the Type II error rate (20%) = chance of observing a non-significant value.
  - Solve for the sample size (via the critical value).

Sample Size for Clinical Outcomes: How It Works

- The smallest clinically important effects define harmful, beneficial and trivial values.
  - At some decision value, Type 1 clinical error rate = 0.5%.
  - Type 2 clinical error rate = 25%

Background

- We study an effect in a sample, but we want to know about the effect in the population.
- The larger the sample, the closer we get to the population.
- Too large is unethical, because it's wasteful.
- Too small is unethical, because the outcome will be indecisive.
  - And you are less likely to get your study funded and published.
  - The traditional approach is based on statistical significance.
  - New approaches are needed for those who are moving away from statistical significance.
  - We will present the traditional approach, two new approaches, and some useful stuff that applies to all approaches.
  - A spreadsheet for all three approaches is available at sportsci.org.

Sample Size for Statistical Significance

- We study an effect in a sample, but we want to know about the effect in the population.
- The larger the sample, the closer we get to the population.
- Too large is unethical, because it's wasteful.
- Too small is unethical, because the outcome will be indecisive.
  - And you are less likely to get your study funded and published.
  - The traditional approach is based on statistical significance.
  - New approaches are needed for those who are moving away from statistical significance.
  - We will present the traditional approach, two new approaches, and some useful stuff that applies to all approaches.
  - A spreadsheet for all three approaches is available at sportsci.org.

Sample Size for Clinical Outcomes

- In the first new approach, the decision is about whether to use the effect in a clinical or practical setting.
  - If you decide to use a harmful effect, it's a false-positive or Type I clinical error.
  - If you decide not to use a beneficial effect, it's a false-negative or Type 2 clinical error.
  - Suggested defaults for acceptable error rates are 0.5% and 25%.
  - Benefit and harm are defined by the smallest clinically important effects.
  - Solve for the sample size by assuming a sampling distribution.
  - Sample sizes are ~1/4, those for statistical significance.
  - The traditional approach is too conservative?
    - P=0.05 with the traditional sample size implies one chance in a million of the effect being harmful.

Sample Size for Statistical Significance: How It Works

- The Type I error rate (5%) defines a critical value of the statistic.
  - If observed value > critical value, the effect is significant.
  - If observed value < critical value, the effect is not significant.
  - When true value = smallest important value, the Type II error rate (20%) = chance of observing a non-significant value.
  - Solve for the sample size (via the critical value).

Sample Size for Clinical Outcomes: How It Works

- The smallest clinically important effects define harmful, beneficial and trivial values.
  - At some decision value, Type 1 clinical error rate = 0.5%.
  - Type 2 clinical error rate = 25%

Background

- We study an effect in a sample, but we want to know about the effect in the population.
- The larger the sample, the closer we get to the population.
- Too large is unethical, because it's wasteful.
- Too small is unethical, because the outcome will be indecisive.
  - And you are less likely to get your study funded and published.
  - The traditional approach is based on statistical significance.
  - New approaches are needed for those who are moving away from statistical significance.
  - We will present the traditional approach, two new approaches, and some useful stuff that applies to all approaches.
  - A spreadsheet for all three approaches is available at sportsci.org.

Sample Size for Statistical Significance

- We study an effect in a sample, but we want to know about the effect in the population.
- The larger the sample, the closer we get to the population.
- Too large is unethical, because it's wasteful.
- Too small is unethical, because the outcome will be indecisive.
  - And you are less likely to get your study funded and published.
  - The traditional approach is based on statistical significance.
  - New approaches are needed for those who are moving away from statistical significance.
  - We will present the traditional approach, two new approaches, and some useful stuff that applies to all approaches.
  - A spreadsheet for all three approaches is available at sportsci.org.
Sample Size for Precise Estimates

- In the second new approach, the decision is about whether the effect has adequate precision in a mechanistic setting.
- "Precision" is defined by the confidence interval: the uncertainty in the true effect.
- "Adequate" implies a confidence interval that does not permit substantial values of the effect in a positive and negative sense.
- Positive and negative are defined by the smallest mechanistically important effects.
- Solve for the sample size by assuming a sampling distribution.
- Sample sizes are similar to those for the first new approach.

Sample Size for Precise Estimates: How It Works

- The smallest substantial positive and negative values define ranges of substantial values.
- Precision is unacceptable if the confidence interval overlaps substantial positive and negative values.

<table>
<thead>
<tr>
<th>SUBSTANTIAL NEGATIVE</th>
<th>TRIVIAL</th>
<th>SUBSTANTIAL POSITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>unacceptable</td>
<td>acceptable</td>
<td>acceptable</td>
</tr>
<tr>
<td>smallest substantial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative value</td>
<td></td>
<td>positive value</td>
</tr>
</tbody>
</table>

- Solve for sample size in the acceptable worst-case scenario.

General Issues

- Check your assumptions and sample-size estimate by comparing with those in published studies.
- But be skeptical about the justifications you see in Methods sections.
- Most are seriously flawed.
- Most either do not mention the smallest important effect or choose a large one to make the sample size acceptable.
- You can justify a sample size on the grounds that it is similar to those in similar studies that produced clear outcomes.
- But effects are clear often because they are substantial.
- If yours turns out to be smaller, it may need a larger sample.

- Bigger effects need smaller samples for decisive outcomes.
  - So start with a smallish cohort, then add more if necessary.
  - Aka "group-sequential design", or "sample size on the fly".
  - But this approach produces upward bias in effect magnitudes that needs sophisticated analysis to fix.
  - An unavoidably suboptimal sample size is ethically defensible...
    - ...if the true effect is large enough for the outcome to be conclusive.
    - And if it turns out inconclusive, argue that it will still set useful limits on the likely magnitude of the effect...
    - ...and should be published, so it can contribute to a meta-analysis.
  - Even optimal sample sizes can produce inconclusive outcomes, thanks to sampling variation.
  - The risk of such an outcome, estimated by simulation, is a maximum of ~10%, when the true effect = critical, decision and null values for the traditional, clinical and precision approaches.
  - Increasing the sample size by ~25% virtually eliminates the risk.

- Sample size depends on the design.
  - Non-repeated measures studies (cross-sectional, prospective, case-control) usually need hundreds or thousands of subjects.
  - Repeated-measures studies (controlled trials and crossovers) usually need scores of subjects.
  - Crossovers need less than parallel-group controlled trials (down to ~¼), provided subjects are stable during the washout.
  - Sample-size estimates for prospective studies and controlled trials should be inflated by 10-30% to allow for drop-outs...
  - ...depending on the demands placed on the subjects, the duration of the study, and incentives for compliance.
  - The problem of unadjusted confounding in observational studies is NOT overcome by increasing sample size.

• Sample size depends on validity and reliability.
  • Effect of validity of a dependent or predictor variable:
    • Sample size is proportional to \( 1/v^2 = 1+e^2/SD^2 \), where
      – \( v \) is the validity correlation of the dependent variable,
      – \( e \) is the error of measurement, and
      – SD is the between-subject standard deviation of the criterion variable in the validity study.
    • So \( r = 0.7 \) implies twice as many subjects as for \( r = 1 \).
  • Effect of reliability of a repeated-measures dependent variable:
    • Sample size is proportional to \( (1 – r) = e^2/SD^2 \), where
      – \( r \) is the test-retest reliability correlation coefficient,
      – \( e \) is the error of measurement, and
      – SD is the observed between-subject standard deviation.
    • So really small sample sizes are possible with high \( r \) or low \( e \).
    • But <10 in any group might misrepresent the population.

• With designs involving comparison of effects in subgroups…
  • You will need twice as many subjects in each subgroup.
  • Example: a controlled trial that would give adequate precision with 20 subjects would need 40 females and 40 males for comparison of the effect between females and males.
  • So don’t go there as a primary aim without adequate resources.
  • But you should be interested in the contribution of subject characteristics to individual differences and responses.
    • The characteristic effectively divides the sample into subgroups.
    • So you need 4x as many subjects to do the job properly!
    • This bigger sample also gives adequate precision for the standard deviation representing individual responses to a treatment.

• With more than one effect, you need a bigger sample size to constrain the overall chance of error.
  • For example, suppose you got chances of harm and benefit…
    …for Effect #1: 0.4% and 72%.
    …for Effect #2: 0.3% and 56%.
    • If you use both, chances of harm = 0.7% (> the 0.5% limit).
    • But if you don’t use #2 (say), you fail to use an effect with a good chance of benefit (> the 25% limit).
    • Solution: increase the sample size…
    …to keep total chance of harm <0.5% for effects you use,
    …and total chance of benefit <25% for effects you don’t use.
  • For \( n \) independent effects, set the Type 1 error rate (%) to 0.5/\( n \) and the Type 2 error rate to 25/\( n \).
  • The spreadsheet shows you need 50% more subjects for \( n=2 \) and more than twice as many for \( n=5 \).
  • For interdependent effects there is no simple formula.

• Make any compared groups equal in size for smallest total sample size.
  • If the size of one group is limited by availability of subjects, recruit more subjects for the comparison group.
  • But >5x more gives no practical increase in precision.
  • Example: 100 cases plus 10,000 controls is little better than 100 cases plus 500 controls.
    • Both are equivalent to 200 cases plus 200 controls.

• Mixing unequal numbers of females and males (or other substantially different subgroups) can decrease the effective sample size.
  • The effect under study has to be estimated separately in females and males, then averaged. Here is an example of the resulting effective sample size (for 90% conf. limits):
    | No. of males | No. of females | Total sample size | Effective sample size |
    |--------------|----------------|-------------------|----------------------|
    | 10           | 10             | 20                | 20                   |
    | 10           | 5              | 15                | 13                   |
    | 10           | 4              | 14                | 10                   |
    | 10           | 3              | 13                | 7                    |
    • Less than the number of males!

• Sample size for a case series defines norms adequately, via the mean and SD of a given measure.
  • The default smallest difference in the mean is 0.2 SD, so the uncertainty (90% confidence interval) needs to be <0.2 SD.
  • Resulting sample size is ¼ that of a cross-sectional study, ~70.
  • Resulting uncertainty in the SD is \( \times 1.15 \), which is OK.
  • Smaller sample sizes will lead to less confident characterization of future cases.
  • Larger sample sizes needed to characterize percentiles, especially for non-normally distributed measures.
• For single-subject studies, “sample size” is the number of repeated observations on the single subject.
• Use the sections of the spreadsheet for cross-sectional studies.
• Use the value for the smallest important difference that applies to sample-based studies.
• Use the subject’s within-subject SD as the “between-subject SD”.
  • The within is often << the between, so sample size is often small.
• Assume any trend-related autocorrelation will be accounted for by your model and will therefore not entail a bigger sample.

• Sample size for measurement studies is not included in available software for estimating sample size.
• Very high reliability and validity can be characterized with as few as 10 subjects.
• More modest validity and reliability (correlations ~0.7-0.9; errors ~2-3× the smallest important effect) need samples of 50-100 subjects.
• Studies of factor structure need many hundreds of subjects.

• Try simulation to estimate sample size for complex designs.
• Make reasonable assumptions about errors and relationships between the variables.
• Generate data sets of various sizes using appropriately transformed random numbers.
• Analyze the data sets to determine the sample size that gives acceptable width of the confidence interval.

Conclusions
• You can base sample size on acceptable rates of clinical errors or adequate precision.
• Both make more sense than sample size based on statistical significance and both lead to smaller samples.
• These methods are innovative and not yet widely accepted.
• So we recommend using the traditional approach in addition to the new approaches.
• Remember to ramp up sample size for:
  • measures with low validity
  • multiple effects
  • comparison of subgroups
  • individual differences.
• If short of subjects, do an intervention with a reliable dependent.

Preparation, article and spreadsheets:
SPORTSCIENCE  sportsci.org
A Peer-Reviewed Site for Sport Research
See Sportscience 10, 63-70, 2006