How to Work with a Biostatistician for Power and Sample Size Calculations

Amber W. Trickey, PhD, MS, CPH
Senior Biostatistician, S-SPIRE
1070 Arastradero #225
atrickey@stanford.edu
Goal: Effective Statistical Collaboration

Essentially all models are wrong but some are useful

George E. P. Box
Topics

Research Fundamentals
- Questions & Measures
- Hypothesis Testing

Power and Sample Size
- Estimation Parameters
- Assumptions

Statistical Collaboration
- Communication
- Ethical Considerations
  ★ Pro Tips
Health Services/Surgical Outcomes Research

**Patient/Disease**
- Comparative Effectiveness
- Meta-analyses
- Patient-centered Outcomes
- Decision Analysis

**Hospital**
- Disparities Research
- Quality Measurement
- Quality Improvement
- Patient Safety
- Implementation Science

**Health Policy**
- Policy Evaluation
- Cost Effectiveness Analysis
- Workforce

[Ban 2016: Is Health Services Research Important to Surgeons?]
Where to Begin?

Research Question!
Research Question (PICO)

**Patient population**
- Condition / disease, demographics, setting, time

**Intervention**
- Procedure, policy, process, treatment

**Comparison/Control group**
- No treatment, standard of care, non-exposed

**Outcome of interest**
- Treatment effects, patient-centered outcomes, healthcare utilization
Example Research Question

• Do hospitals with 200+ beds perform better than smaller hospitals?
  ❖ More developed question: specify population & outcome

• Do large California hospitals with 200+ beds have lower surgical site infection rates for adults undergoing inpatient surgical procedures?
  • Population: California adults undergoing inpatient surgical procedures with general anesthesia in 2017
  • Intervention (structural characteristic): 200+ beds
  • Comparison: smaller hospitals with <200 beds
  • Outcome: surgical site infections within 30 days post-op
Internal & External Validity

External validity: generalizability to other patients & settings

- Study design
  - Which patients are included
  - How the intervention is implemented
  - Real-world conditions

Internal validity: finding a true cause-effect relationship

- Study design + analysis
  - Specific information collected (or not)
  - Data collection definitions
  - Data analysis methods
Variable Types

Independent Variable (primary IV)
- Exposure (Intervention)
- Occurring first
- Causal relationship (?)

Dependent Variable (DV)
- Outcome
- Response variable
- Occurring after predictors

Confounder(s)
- Related to both outcome and exposure
- Must be taken into account for internal validity
## Variable Measurement Scales

<table>
<thead>
<tr>
<th>Type of Measurement</th>
<th>Characteristics</th>
<th>Examples</th>
<th>Descriptive Stats</th>
<th>Information Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Ranked spectrum; quantifiable intervals</td>
<td>Weight, BMI</td>
<td>Mean (SD) + all below</td>
<td>Highest</td>
</tr>
<tr>
<td>Ordered Discrete</td>
<td></td>
<td>Number of cigs / day</td>
<td>Mean (SD) + all below</td>
<td>High</td>
</tr>
<tr>
<td>Categorical Ordinal</td>
<td>Ordered categories</td>
<td>ASA Physical Status Classification</td>
<td>Median</td>
<td>Intermediate</td>
</tr>
<tr>
<td>(Polychotomous)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categorical Nominal</td>
<td>Unordered Categories</td>
<td>Blood Type, Facility</td>
<td>Counts, Proportions</td>
<td>Lower</td>
</tr>
<tr>
<td>(Polychotomous)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categorical Binary</td>
<td>Two categories</td>
<td>Sex (M/F), Obese (Y/N)</td>
<td>Counts, Proportions</td>
<td>Low</td>
</tr>
<tr>
<td>(Dichotomous)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Hulley 2007]
Measures of Central Tendency

1. Mean = average
   • Continuous, normal distribution
2. Median = middle
   • Continuous, nonparametric distribution
3. Mode = most common
   • Categorical
Variability

- Critical for describing & comparing populations.
- Example measures:
  - SD = “average” deviation from mean
  - Range = minimum – maximum
  - Interquartile range = 25\textsuperscript{th} - 75\textsuperscript{th} percentiles
- For skewed distributions, range / IQR are more representative of variability than SD.
  - E.g. $ and time
    (hospital charges, length of stay)
Hypothesis Testing
Hypothesis Testing

• **Null Hypothesis (H₀)**
  - Default assumption for superiority studies
    - Intervention/treatment **has NO effect**, i.e. no difference b/t groups
    - Acts as a “straw man”, assumed to be true so that it can be knocked down as false by a statistical test.

• **Alternative Hypothesis (Hₐ)**
  - Assumption being tested for superiority studies
    - Intervention/treatment **has an effect**

• **Non-inferiority study** hypotheses are reversed:
  alternative hypothesis = no difference (within a specified range)
Error Types

Type I Error $\alpha$: False positive
- Finding an effect that is not true
- Due to: Spurious association
- Solution: Repeat the study

Type II Error ($\beta$): False negative
- Do not find an effect when one truly exists
- Due to: Insufficient power, high variability / measurement error
- Solution: Increase sample size

Probability $\alpha = 0.05$
Hypothesis Testing

One- vs. Two-tailed Tests

One-sided

Evaluate association in one direction

Two-sided

Two-sided tests almost always required – higher standard, more cautious

Test Statistic

$H_A$ : $M_1 < M_2$

$H_0$ : $M_1 = M_2$

$H_A$ : $M_1 > M_2$
P-value Definition

The p-value represents the probability of finding the observed, or a more extreme, test statistic if the null hypothesis is true.

- Measures evidence against $H_0$
- Smaller p-value, larger evidence against $H_0$
- Reject $H_0$ if p-value $\leq \alpha$
P-Value Pitfalls

• P is highly dependent on sample size

• The *statistical* significance ...
  - does not equal *clinical* significance
  - does not equal *effect size*
  - Report descriptive statistics with p: n1, n2, %’s, means, SD...

• P is not dichotomous yes/no, but a continuum, <0.001 to >0.99
Which Statistical Test?

1. Number of IVs
2. IV Measurement Scale
3. Independent vs. Matched Groups
4. DV Measurement Scale

LEGEND:
IV = Independent Variable (i.e. predictor, exposure)
DV = Dependent Variable (i.e. response, outcome)
# Common Regression Models

<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Appropriate Regression</th>
<th>Model Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Linear Regression</td>
<td>Slope ($\beta$): How much the <strong>outcome</strong> increases for every 1-unit increase in the predictor</td>
</tr>
<tr>
<td>Binary / Categorical</td>
<td>Logistic Regression</td>
<td>Odds Ratio (OR): How much the <strong>odds</strong> for the outcome increases for every 1-unit increase in the predictor</td>
</tr>
<tr>
<td>Time-to-Event</td>
<td>Cox Proportional-Hazards Regression</td>
<td>Hazard Ratio (HR): How much the <strong>rate</strong> of the outcome increases for every 1-unit increase in the predictor</td>
</tr>
<tr>
<td>Count</td>
<td>Poisson Regression or Negative Binomial Regression</td>
<td>Incidence Rate Ratio (IRR): How much the <strong>rate</strong> of the outcome increases for every 1-unit increase in the predictor</td>
</tr>
</tbody>
</table>
Hierarchical / Mixed Effects Models

Correlated Data
- Grouping of subjects
- Repeated measures over time
- Multiple related outcomes

Can handle
- Missing data
- Nonuniform measures

Outcome Variable(s)
- Categorical
- Continuous
- Counts

Nested Data
Estimating Power
Error Types

Type I Error (\(\alpha\)): False positive
- Find an effect when it is truly not there
- Due to: Spurious association
- Solution: Repeat the study

Type II Error (\(\beta\)): False negative
- Do not find an effect when one truly exists
- Due to: Insufficient power, high variability / measurement error
- Solution: Increase sample size

Probability \(\beta = 0.20\)
Statistical Power

• A study with low power has a high probability of committing type II error.

★ Sample size planning aims to select a sufficient number of subjects to keep \( \alpha \) and \( \beta \) low without making the study too expensive or difficult.

Power = 1 − \( \beta \) (typically 1 − 0.2 = 0.8)

We want to know:
How many subjects do we need to find a statistically significant & **meaningful** effect size?
Statistical Power Tools

Three broad categories

1. Hypothesis-based
   - Formally testing a hypothesis to determine a statistically significant effect

2. Confidence interval-based
   - Estimating a number (e.g. prevalence) with a desired level of precision

3. Rules of thumb
   - Based on simulation studies, we estimate (ballpark) the necessary sample size
   - Interpret carefully & in conjunction with careful sample size calculation using method 1 or 2
Parameters of (1) Hypothesis-Based Power Calculations

- Outcome of interest
- Study design
- Effect Size
- Allocation ratio between groups
- Population variability
- Alpha (p-value, typically 0.05)
- Beta (1-power, typically 0.1-0.2)
- 1- vs. 2-tailed test
Effect Size

- Cohen’s d: comparison between two means
  - $d = m_1 - m_2 / \text{pooled SD}$
  - Small $d=0.2$; Medium $d=0.5$; Large $d=0.8$

- Expected meaningful differences in outcomes per group
  - e.g. complications: 10% open vs. 3% laparoscopic

- Minimal clinically important difference (e.g. 10% improvement)
  - What is the MCID that would lead a clinician to change his/her practice?

☆ Expect an inverse relationship between effect size & sample size
  - ↑ effect size, ↓ sample size
  - ↓ effect size, ↑ sample size
(2) Confidence Interval-Based Power

• How precisely can you estimate your measure of interest?

• Examples
  • Diagnostic tests: Sensitivity / Specificity
  • Care utilization rates
  • Treatment adherence rates

• Calculation components
  • N
  • Variability
  • $\alpha$ level
  • Expected outcomes
(3) Rule of Thumb Power Calculations

- Simulation studies
- Degrees of freedom (df) estimates
  - df: the number of IV factors that can vary in your regression model
  - Multiple linear regression: ~15 observations per df
  - Multiple logistic regression: df = # events/15
  - Cox regression: df = # events/15
- Best used with other hypothesis-based or confidence interval-based methods
Power Pitfalls

• Requires many assumptions

★ Conduct power sensitivity analyses to assess your assumptions

• Effect sizes can vary between populations/sites/studies

• Power estimates should focus on the minimum clinically important difference (MCID)

• If power calculation estimated effect size >> observed effect size, sample may be inadequate or observed effect may not be meaningful.
Collaboration with Biostatisticians
Biostatistics Collaboration

• 2001 Survey of BMJ & Annals of Internal Medicine re: statistical and methodological collaboration

• Stats/methodological support – how often?
  • Biostatistician 53%
  • Epidemiologist 32%

• Authorship outcomes given significant contribution
  • Biostatisticians 78%
  • Epidemiologists 96%

• Publication outcomes
  • Studies w/o methodological assistance more likely to be rejected w/o review: 71% vs. 57%, p=0.001

[Altman, 2002]
Power Questions from your Biostatistician

• What is the research question?
• What is the study design?
• What effect do you expect to observe?
• What other variables may affect your results?
• How many patients are realistic?
• Do you have repeated measures per individual/analysis unit?
• What are your expected consent and follow-up completion rates?
• Do you have preliminary data?
  • Previous studies / pilot data
  • Published literature
Stages of Power Calculation

1. **Study Design**
2. **Hypothesis**
3. **Sample Size**
4. **Similar Literature**
5. **Simulation/Rules of Thumb**
6. **Feasible? Important? Other Considerations?**

[Pye, 2016]
Statistical Power Tips

• Seek biostatistician feedback early

• Calculations take time and typically a few iterations

• Without pilot data, it is helpful to identify previous research with similar methods
  • If absolutely no information is available from a reasonable comparison study, you can estimate power from the minimum clinically important difference*

• Calculate power before the study is implemented
  • Post hoc power calculations are less useful, unless to inform the next study

• Report estimated power as a range, varying assumptions and conditions

*[Revicki, 2008]*
Authorship

International Committee of Medical Journal Editors (ICMJE):
“All those designated as authors should meet all four criteria for authorship, and all who meet the four criteria should be identified as authors.”

1. Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
2. Drafting the work or revising it critically for important intellectual content; AND
3. Final approval of the version to be published; AND
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

★ Epidemiologist/Biostatisticians typically qualify for authorship
(Sometimes an acknowledgement is appropriate)
• Authorship plan must be discussed
Ethical Principles

• Statisticians aim to be knowledgeable about the system under study, yet recognize the limitations of their subject-matter knowledge
  ❖ Real data are not “context-free”

• Honest statistical work has nothing to hide – it says what it says

• Admit where models are imperfect or conclusions are model-dependent

• Journal submissions should represent complete, high-quality, best effort

• Routinely take the advice of peer-review referees and editors
  ❖ Typically their advice is constructive and improves the article

• Give appropriate credit with co-authorship

[Vardeman, 2003: Statistics and Ethics]
How do I collaborate with SPİRE?
S-SPIRE Biostatisticians

Amber Trickey, PhD, MS, CPH
Senior Biostatistician

Qian Ding, MS
Biostatistician

Rui Chen, MS
Biostatistician

Nic Barreto, MPH
Biostatistician

Tong Meng, MPP
Biostatistician

Kate Arnow, MS
Biostatistician
Research Collaboration Program

The S-SPIRE team is happy to receive collaboration requests related to health services research projects. Collaboration consists of an initial in-person meeting at S-SPIRE Center to discuss short- and long-term quantitative, qualitative, and mixed methods research projects and follow-up assistance with research design, instrument development, analysis, interpretation, and write-up of methods/analysis for papers and grant proposals. Please plan for availability at least two weeks out.

Submit a Collaboration Request Form
Timelines for Initial Collaboration Meeting

1. Conference abstract deadlines
   • 4 weeks lead time with data ready for analysis (email 6 weeks out for appt)

2. Special issue or meeting paper deadlines
   • 6 weeks lead time with data ready for analysis (email 8 weeks out for appt)
   • Depending on the complexity of the analysis proposed, longer lead times may be necessary.

3. Grant application deadlines
   • 8-12 weeks lead time (email 10-14 weeks out for appt)
   • Statistical tests are tied to the research questions and design; earlier collaboration will better inform grant development
Summary

• Power calculations are complex, but biostatisticians can help
• Contact your biostatistician as early as possible
  • Power/sample calculations are iterative and take time
• Gather information prior to meeting
  1. Study design
  2. Expected effect size
  3. Feasible sample size
  4. Similar literature
  5. Pilot data
• Meet with us at 1070 Arastradero (PI must attend)
Thank you!

atrickey@stanford.edu
(650) 725-7003
References


