Research Design and Statistics

Amber W. Trickey, PhD, MS, CPH

Senior Biostatistician, S-SPIRE
atrickey@stanford.edu
Outline

1. Research Design
   • Research Question
   • Validity
   • Variables

2. Statistics
   • Measurement
   • Hypothesis testing
   • Power calculations
   • Statistical tests
Guiding Principles for Research Design

We must be careful not to confuse data with the abstractions we use to analyze them.

William James

Statistics are no substitute for judgment.

Henry Clay
Research Question (PICO)

1. Patient population
   • Condition / disease, demographics, setting, time

2. Intervention
   • Procedure / policy / test / process intervention

3. Comparison group
   • Control group: standard of care, non-exposed, medical management, no treatment

4. Outcome of interest
   • Treatment effects, side effects, patient-reported outcomes, proxy measures

* Call your friendly biostatistician now
External & Internal Validity

External validity: generalizability to other settings
- **Study design**
  - Which patients are included
  - How the treatment is implemented

Internal validity: finding a true cause-effect
- **Study design + analysis**
  - Randomized treatment assignment
  - Specific information collected (or not)
  - Data analysis methods
Study Designs

Evidence Pyramid:

- **Strength of Causal Evidence**

- **Descriptive Study Designs**
  - Editorials, Expert Opinion
  - Case Series, Case Reports

- **Analytic Study Designs**
  - Case-Control Studies
  - Cohort Studies
  - Randomized Controlled Trials
  - Systematic Reviews
  - Meta-Analyses
  - Reviews

- **Evidence Pyramid**

**s-SPiRE**
RCT Designs

- Individual randomization
  - Pros: Best power per n (optimal efficiency)
  - Considerations: Contamination within sites, may be infeasible

- Cluster randomization
  - Pros: Minimize contamination
  - Considerations: Individuals not independent, intracluster correlation coefficient (ICC), larger n needed to achieve power

- Stepped-wedge design
  - Pros: All sites get intervention, timing/order is randomized
  - Considerations: Individuals not independent (ICC), larger n needed to achieve power

[Rutterford 2015]
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 (Polychotomous)</td>
<td>Ordered categories</td>
<td>ASA Physical Status Classification</td>
<td>Median</td>
<td>Intermediate</td>
</tr>
<tr>
<td>Categorical Nominal (Polychotomous)</td>
<td>Unordered Categories</td>
<td>Blood Type, Facility</td>
<td>Counts, Proportions</td>
<td>Lower</td>
</tr>
<tr>
<td>Categorical Binary (Dichotomous)</td>
<td>Two categories</td>
<td>Sex (M/F), Obese (Y/N)</td>
<td>Counts, Proportions</td>
<td>Low</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:
  o SD = “average” deviation from mean
  o Range = minimum – maximum
  o Interquartile range = 25\textsuperscript{th} - 75\textsuperscript{th} percentiles

• For skewed distributions, range / IQR are more representative of variability than SD.
  o E.g. $ and time
    (hospital charges, length of stay)
Measures of Association: Risk Ratio

• Risk
  o Probability that a new event will occur
  o Risk in Exposed = a / (a+b)
  o Risk in Unexposed = c / (c+d)
    • AKA ‘Baseline Risk’

• Risk Ratio (RR)
  o Ratio of two risks

\[
RR = \frac{\frac{a}{a + b}}{\frac{c}{c + d}}
\]
Interpretation similar for other ratio measures

• RR=1
  - Risk in exposed = Risk in unexposed
  - No association

• RR > 1
  - Positive association, ? causal

• RR < 1
  - Negative association, ? protective
Measures of Association: Odds Ratio

• Odds
  • Ratio of the probability of an occurrence to the probability of nonoccurrence
  • Odds in Exposed = \( \frac{a}{b} \)
  • Odds in Unexposed = \( \frac{c}{d} \)
    • AKA ‘Baseline Odds’

• Odds Ratio (OR)
  • Ratio of two odds

\[
OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}
\]

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
</tr>
</tbody>
</table>
Guiding Principles for Statistical Analysis

Essentially all models are wrong but some are useful.

George E. P. Box

Models should be as simple as possible, but not more so.

(attributed) Albert Einstein
Hypothesis Testing
Hypothesis Testing

• Null Hypothesis ($H_0$)
  o Default assumption in superiority studies
    ➢ Treatment has NO effect, no difference b/t groups
    o Acts as a “straw man”, assumed to be true so that it can be knocked down as false by stats test.

• Alternative Hypothesis ($H_A$)
  o Assumption being tested in superiority studies
    ➢ Group/treatment has an effect

• Non-inferiority study design hypotheses are reversed.
Hypothesis Testing

One- vs. Two-tailed Tests

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Test Statistic</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_A: M_1 &lt; M_2$</td>
<td></td>
<td>$H_A: M_1 &lt; M_2$</td>
</tr>
<tr>
<td>$H_0: M_1 = M_2$</td>
<td></td>
<td>$H_0: M_1 = M_2$</td>
</tr>
<tr>
<td>$H_A: M_1 &gt; M_2$</td>
<td></td>
<td>$H_A: M_1 &gt; M_2$</td>
</tr>
</tbody>
</table>

Evaluate association in one direction

Two-sided tests almost always required – higher standard, more cautious
1- vs. 2-Sided Hypothesis Testing

1. One-tailed tests only look for association in one direction.
2. Two-tailed tests are almost always required.
   • Journal & conference audiences are highly skeptical of one-tailed tests.
3. Testing for associations at both ends of the spectrum is more cautious, conservative, and sets a higher standard.
4. The only valid reason to use a one-tailed test is if the direction of the association CAN ONLY go one way.
Error Types

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

Probability $\alpha = 0.05$

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
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)
Example Study

• Carotid Endarterectomy under general anesthesia (GA) vs. local/regional anesthesia (RA)
• Primary outcome of interest: 30-day postoperative myocardial infarction
• NSQIP covariates
Chi-Square or Fisher’s Exact (if any cell N<10)

Outcome Variable:
- Binary

Predictor Variable(s):
- Categorical
  - Independent groups
  - 2+ groups
  - 1 independent variable

2 x 2 Table:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MI +</td>
</tr>
<tr>
<td>GA</td>
<td>324</td>
</tr>
<tr>
<td>RA</td>
<td>20</td>
</tr>
<tr>
<td>Tot</td>
<td>344</td>
</tr>
</tbody>
</table>
Student’s T-test

Outcome Variable:

- Continuous
- Normally distributed

Predictor Variable(s):

- 1 Categorical IV
- 2 Independent groups

CEA: Age Comparison

![Box plot diagram showing age comparison between Gen Anes and Loc Anes groups.](image)
Analysis of Variance (ANOVA)

Outcome Variable:
- Continuous
- Normally distributed

Predictor Variable(s):
- 1 Categorical IV
- 3+ Independent groups

CEA: Op time by PGY

Operative Time (Minutes)

No Resident

1 2 3 4 5 6 7
Resident PGY Level
Paired T-test

Outcome Variable:

- Continuous
- Normally distributed

Predictor Variable(s):

- 1 Categorical IV
- 2 matched / dependent groups

1st vs. 2nd CEA Operative Time

Operative Time (Minutes)
Pearson’s Correlation

Outcome Variable:
- Continuous

Predictor Variable(s):
- 1 Continuous IV
- Normally distributed

CEA Op Time by BMI

Operative Time (Minutes)

BMI
Linear Regression

Outcome Variable:
- Continuous

Predictor Variable(s):
- Categorical OR continuous
- Number of IVs
  - 1 = simple linear regression
  - 2+ = multiple linear regression

CEA Op Time by BMI
Logistic Regression

Outcome Variable:
- Dichotomous (binary)

Predictor Variable(s):
- Categorical OR continuous
- Number of IVs
  - 1 = simple logistic regression
  - 2+ = multiple logistic regression
# Common Regression Models

<table>
<thead>
<tr>
<th>Outcome Variable</th>
<th>Appropriate Regression</th>
<th>Model Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous AND Normal</td>
<td>Linear Regression</td>
<td>Slope ($\beta$): How much the outcome 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 \textbf{odds} 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 \textbf{rate} 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
Power and Sample Size
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.

- Power = 1 – β
- Sample size planning aims to select a sufficient number of subjects to keep α and β low without making the study too expensive or difficult.
- Translation: How many subjects do I need to find a statistically & meaningful effect?
- Sample size calculation pitfalls:
  - Requires MANY assumptions
  - If power calculation estimated effect size >> observed effect size, sample may be inadequate or observed effect may not be meaningful.
Power Calculation Components

- Alpha (p-value, typically 0.05)
- Beta (1-power, typically 0.1-0.2)
- 1- vs. 2-tailed test
- Minimum important difference (clinically relevant)
- Population variability
- Outcome of interest
- Allocation ratio between groups
- Study design
Power Considerations

• Seek biostatistician feedback early
  ✶ power calculations are the perfect opportunity

• 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.

• Journals are increasingly requiring details of power calculations
Seven Habits of Highly Effective Data Users

1. **Check quality before quantity.** All data are not created equal; fancy statistics cannot salvage biased data.

2. **Describe before you analyze.** Special data require special tests; improper analysis gives deceptive results.

3. **Accept the uncertainty of all data.** All observations have some random error; interpretation requires estimating precision or confidence.

4. **Measure error with the right statistical test.** + results qualified by the chance of being wrong, neg. results by chance of having missed a true effect.

5. **Put clinical importance before statistical significance.** Statistical tests measure error, not importance; an appropriate measure of clinical importance must be checked.

6. **Seek the sample source.** Results from one dataset do not necessarily apply to others.

7. **View science as a cumulative process.** A single study is rarely definitive.

[Rosenfeld 2014]
Resource

- https://stats.idre.ucla.edu/other/dae/

<table>
<thead>
<tr>
<th>Regression Models</th>
<th>Stata</th>
<th>SAS</th>
<th>SPSS</th>
<th>Mplus</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models for Binary and Categorical Outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multinomial Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinal Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probit Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count Models</th>
<th>Stata</th>
<th>SAS</th>
<th>SPSS</th>
<th>Mplus</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Binomial Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-inflated Poisson Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-inflated Negative Binomial Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-truncated Poisson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-truncated Negative Binomial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Censored and Truncated Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobit Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truncated Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power Analysis / Sample Size</th>
<th>Stata</th>
<th>SAS</th>
<th>R</th>
<th>G*Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-sample t-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired-sample t-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent-sample t-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Independent Proportions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-way ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy in Parameter Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thanks!

Questions?

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


Well-written power example

“We are planning to compare the proportion of union between intramedullary nail with and without reaming in patients with a fractured tibia at 6-month follow-up using a ratio of 1:1. From our pilot study, the proportion of union was 0.85 for intramedullary nail without reaming and 0.95 for intramedullary nail with reaming within 6 months.

Assuming an MID of 0.1, we will need to enroll at least 140 patients per group to be able to reject a null hypothesis of no difference in proportions of union between the 2 groups with 80% power. The type-I error probability associated with this 2-sided test of this null hypothesis is 0.05.”

[Farrokhyar, 2013]