# Estimating Power and Sample Size

## (How to Help Your Biostatistician!)

**Amber W. Trickey, PhD, MS, CPH**
Senior Biostatistician
1070 Arastradero #225
atrickey@stanford.edu

S-SPIRE Stanford–Surgery Policy Improvement Research & Education

# Goal: Effective Statistical Collaboration

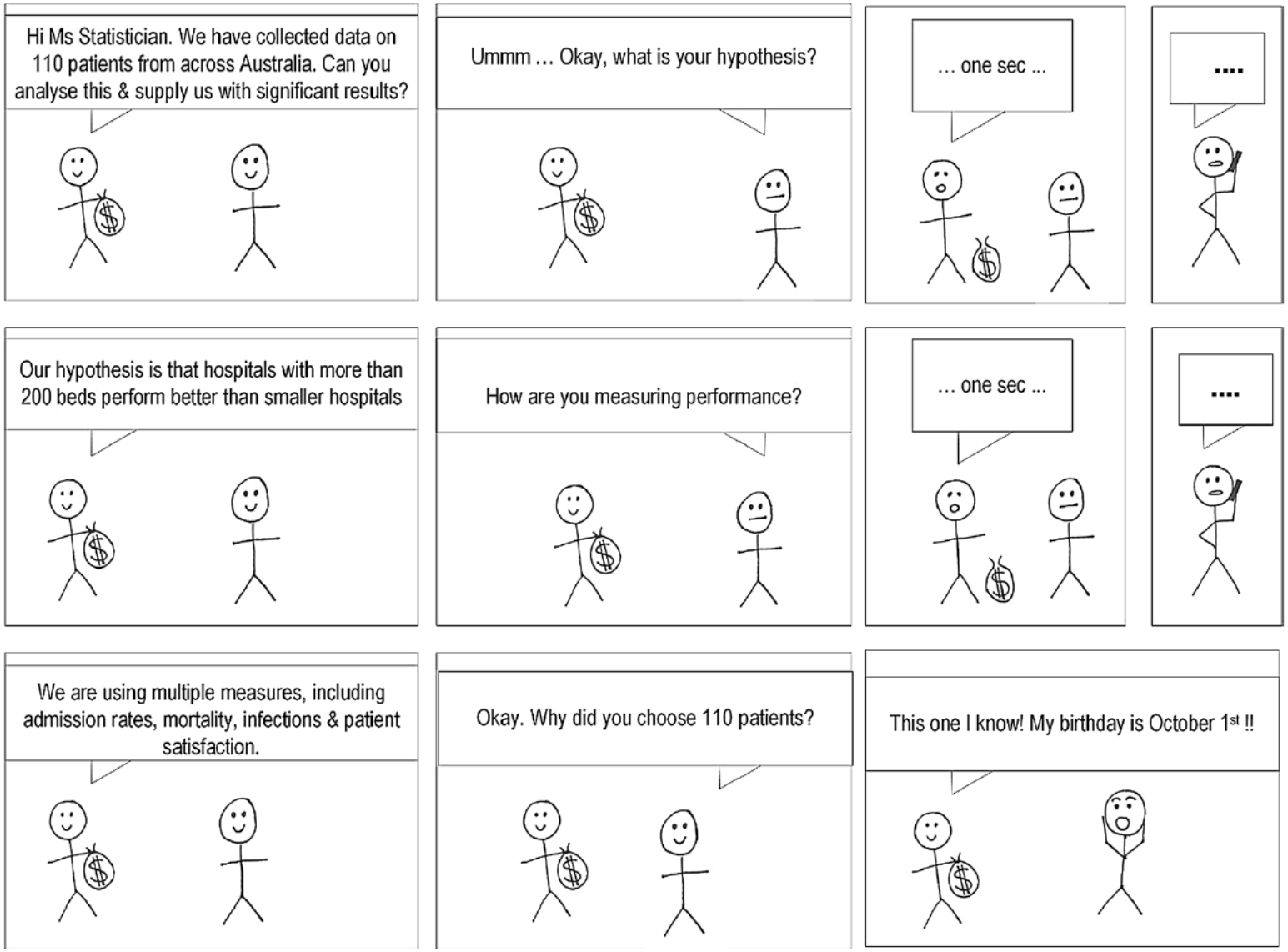



**Fig. 1** A statistician's dilemma

[Pye, 2016]

# Topics

| | |
|---|---|
| Research Data | • Questions & Measures<br>• Hypothesis Testing |
| Statistical Power | • Components<br>• Assumptions |
| Statistical Collaboration | • Consultation Process<br>• Timelines |

# Research Question (PICO)

1. Patient population
   - Condition / disease, demographics, setting, time
2. Intervention
   - Procedure, policy, process, treatment
3. Comparison group
   - Control group (e.g. no treatment, standard of care, non-exposed)
4. 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 >200 beds have lower surgical site infection rates for adults undergoing inpatient surgical procedures?
  - **P**opulation: California adults undergoing inpatient surgical procedures with general anesthesia in 2017
  - **I**ntervention (structural characteristic): 200+ beds
  - **C**omparison: smaller hospitals with <200 beds
  - **O**utcome: surgical site infections within 30 days post-op

S-SPIRE

# 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

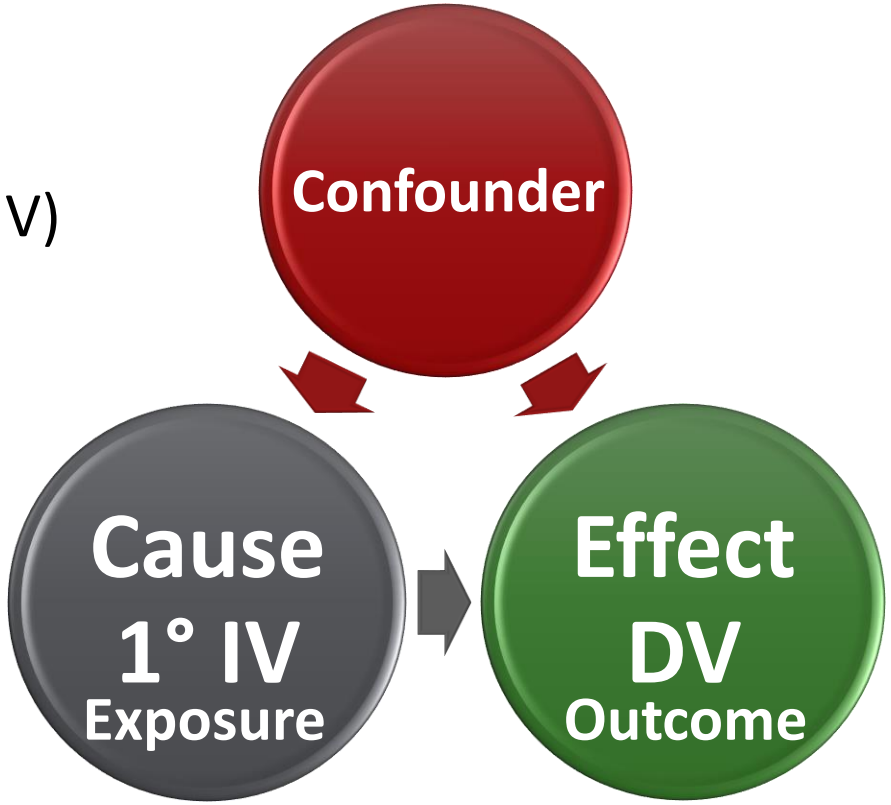
- Study design + analysis
  - Specific information collected (or not)
  - Data collection definitions
  - Data analysis methods

# Variable Types

1. Exposure (Intervention)
   - Predictor / Primary Independent variable (IV)
   - Occurring first
   - Causal relationship (?)

2. Outcome
   - Response / Dependent variable (DV)
   - Occurring after predictors

3. Confounders
   - Related to both outcome and exposure
   - Must be taken into account for internal validity



# Variable Measurement Scales

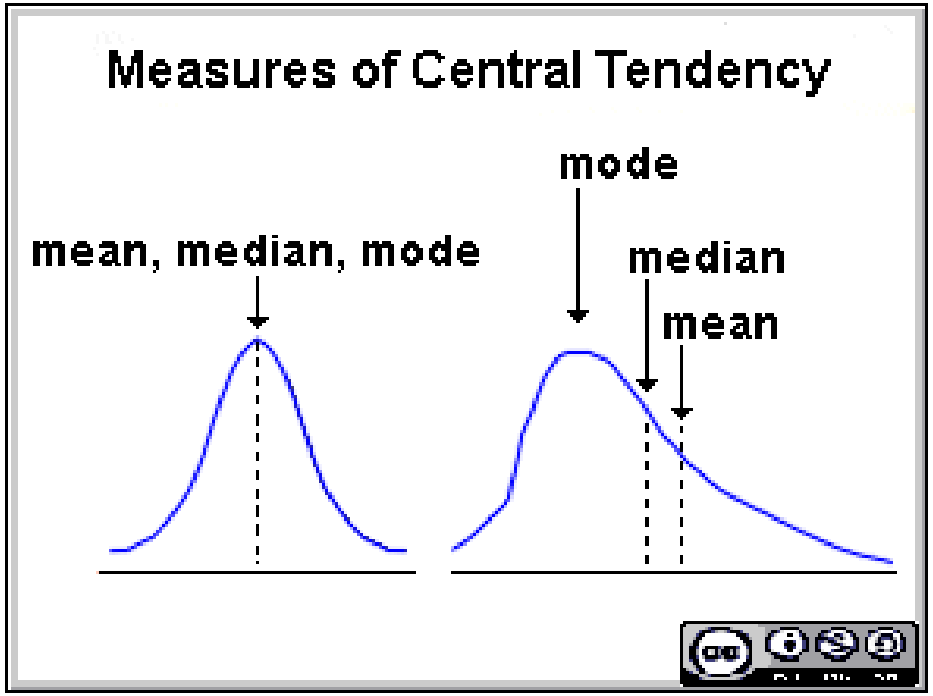
| Type of Measurement | Characteristics | Examples | Descriptive Stats | Information Content |
|---|---|---|---|---|
| Continuous | Ranked spectrum; quantifiable intervals | Weight, BMI | Mean (SD) + all below | Highest |
| Ordered Discrete | | Number of cigs / day | Mean (SD) + all below | High |
|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

[Hulley 2007]

S-SPIRE

# Measures of Central Tendency

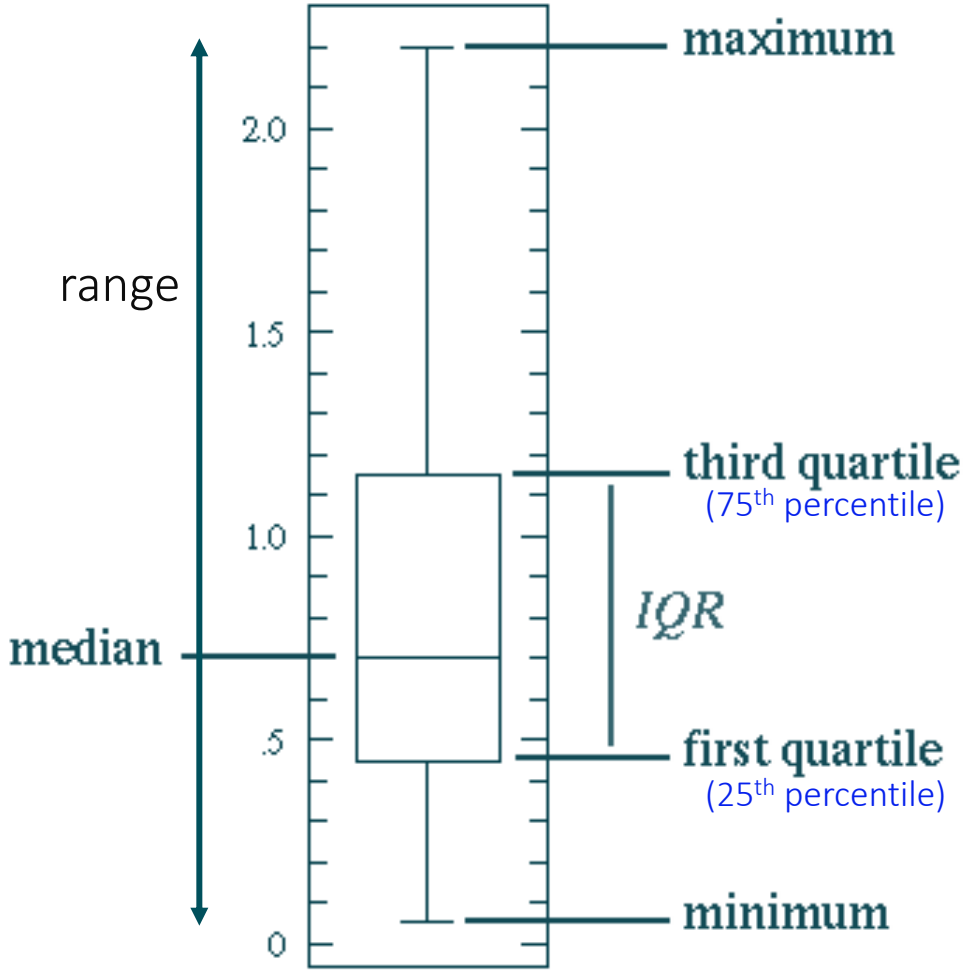
1. Mean = average
   - Continuous, normal distribution
2. Median = middle
   - Continuous, nonparametric distribution
3. Mode = most common
   - Categorical



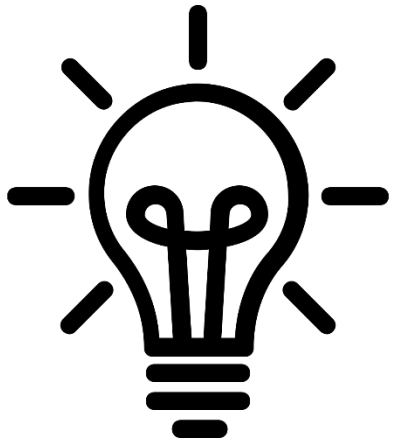
# Variability

- Averages are important, but variability is critical for describing & comparing populations.
- Example measures:
  - SD = “average” deviation from mean
  - Range = minimum – maximum
  - Interquartile range = 25$^{th}$ - 75$^{th}$ percentiles
- For skewed distributions (e.g. $, time), range or IQR are more representative measures of variability than SD.

## Box Plot Components:



# Hypothesis Testing



# Hypothesis Testing

- Null Hypothesis ($H_0$)
  - 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 ($H_A$)
  - 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

Probability $\alpha = 0.05$

**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

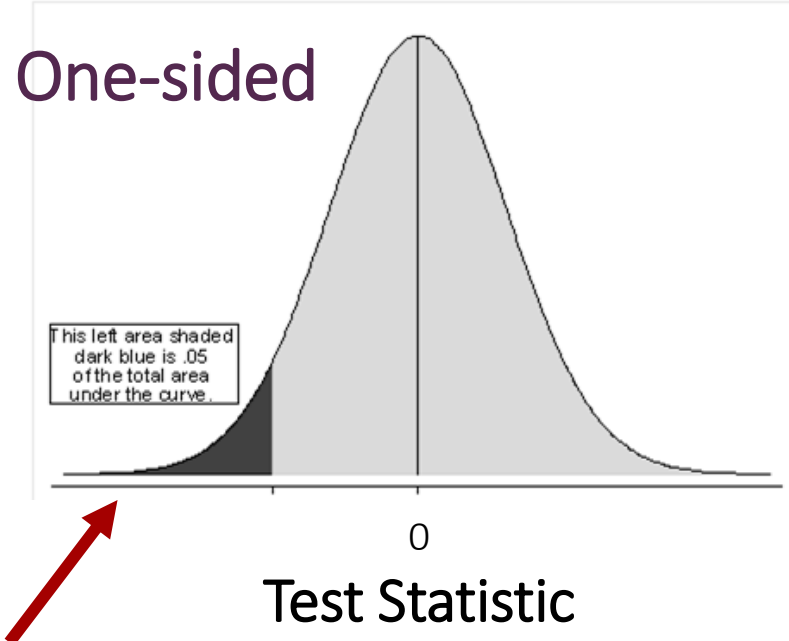
S-SPIRE

# Hypothesis Testing

## One- vs. Two-tailed Tests

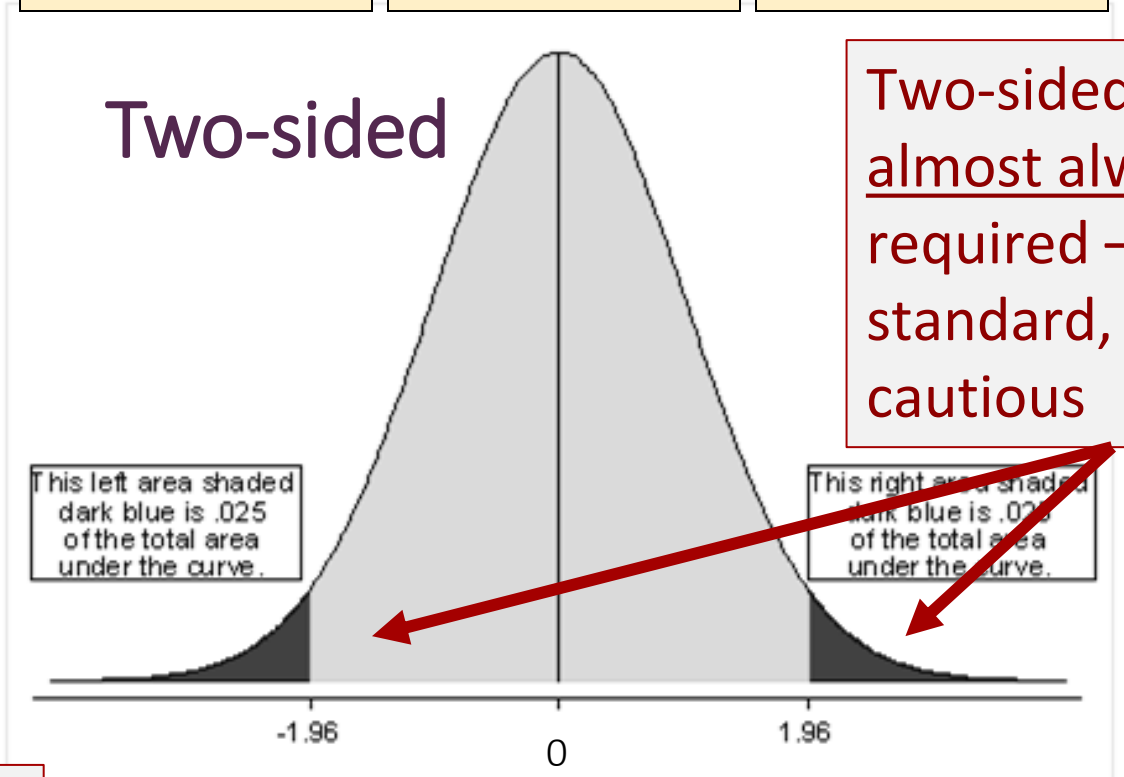
| $H_A$ : M1 < M2 | $H_0$ : M1 = M2 | $H_A$ : M1 > M2 |
|---|---|---|

One-sided




Evaluate association in one direction

Two-sided




Two-sided tests <u>almost always</u> required – higher standard, more cautious

S-SPIRE

# 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.

# P-Value

P-value measures evidence against $H_0$

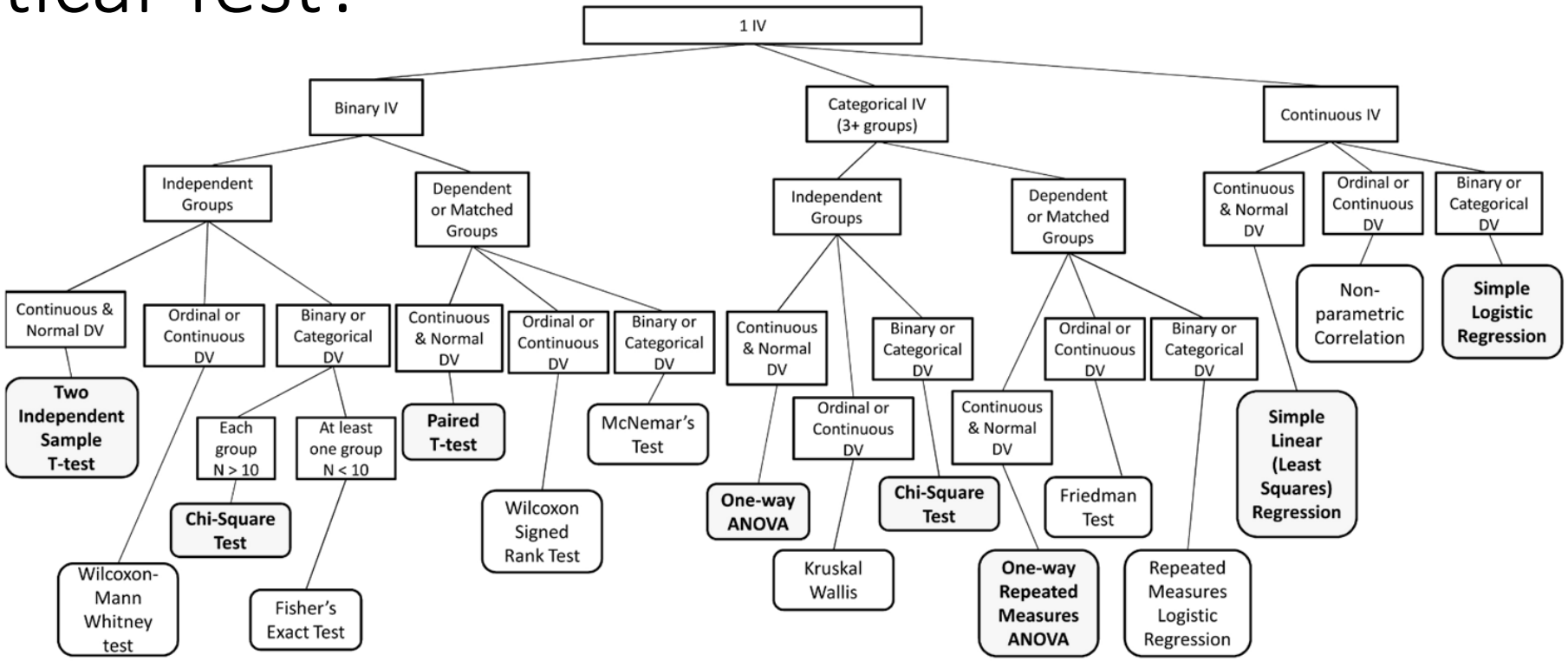
- Smaller the p-value, the larger the evidence against $H_0$
- Reject $H_0$ if p-value $\leq \alpha$

Pitfalls:

- The statistical significance of the effect does not explain the size of the effect
- Report descriptive statistics with p-values (N, %, means, SD, etc.)
- STATISTICAL significance does not equal CLINICAL significance
- P is not truly yes/no, all or none, but is actually a continuum
- P is highly dependent on sample size

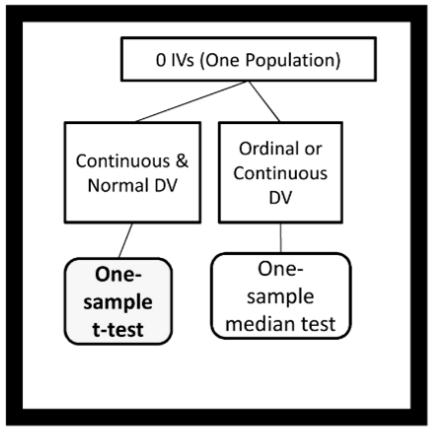
# Which Statistical Test?

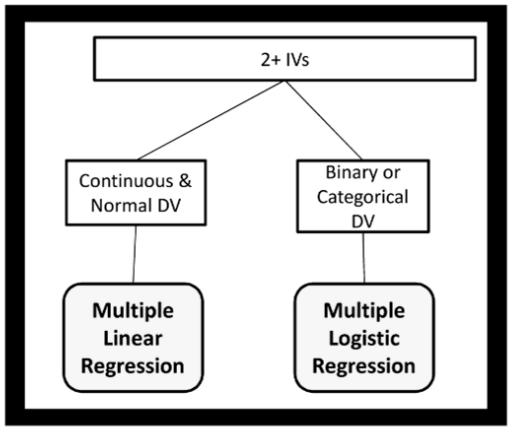
1. **Number of IVs**
2. **IV Measurement Scale**
3. **Independent vs. Matched Groups**
4. **DV Measurement Scale**




LEGEND:

IV = Independent Variable
(i.e. predictor, exposure)
DV = Dependent Variable
(i.e. response, outcome)



# Common Regression Models

| Outcome Variable | Appropriate Regression | Model Coefficient |
|---|---|---|
| Continuous | Linear Regression | Slope (β): How much the **outcome** increases for every 1-unit increase in the predictor |
| Binary / Categorical | Logistic Regression | Odds Ratio (OR): How much the **odds** for the outcome increases for every 1-unit increase in the predictor |
| Time-to-Event | Cox Proportional-Hazards Regression | Hazard Ratio (HR): How much the **rate** of the outcome increases for every 1-unit increase in the predictor |
| Count | Poisson Regression or Negative Binomial Regression | Incidence Rate Ratio (IRR): How much the **rate** of the outcome increases for every 1-unit increase in the predictor |

# Hierarchical / Mixed Effects Models

Correlated Data

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

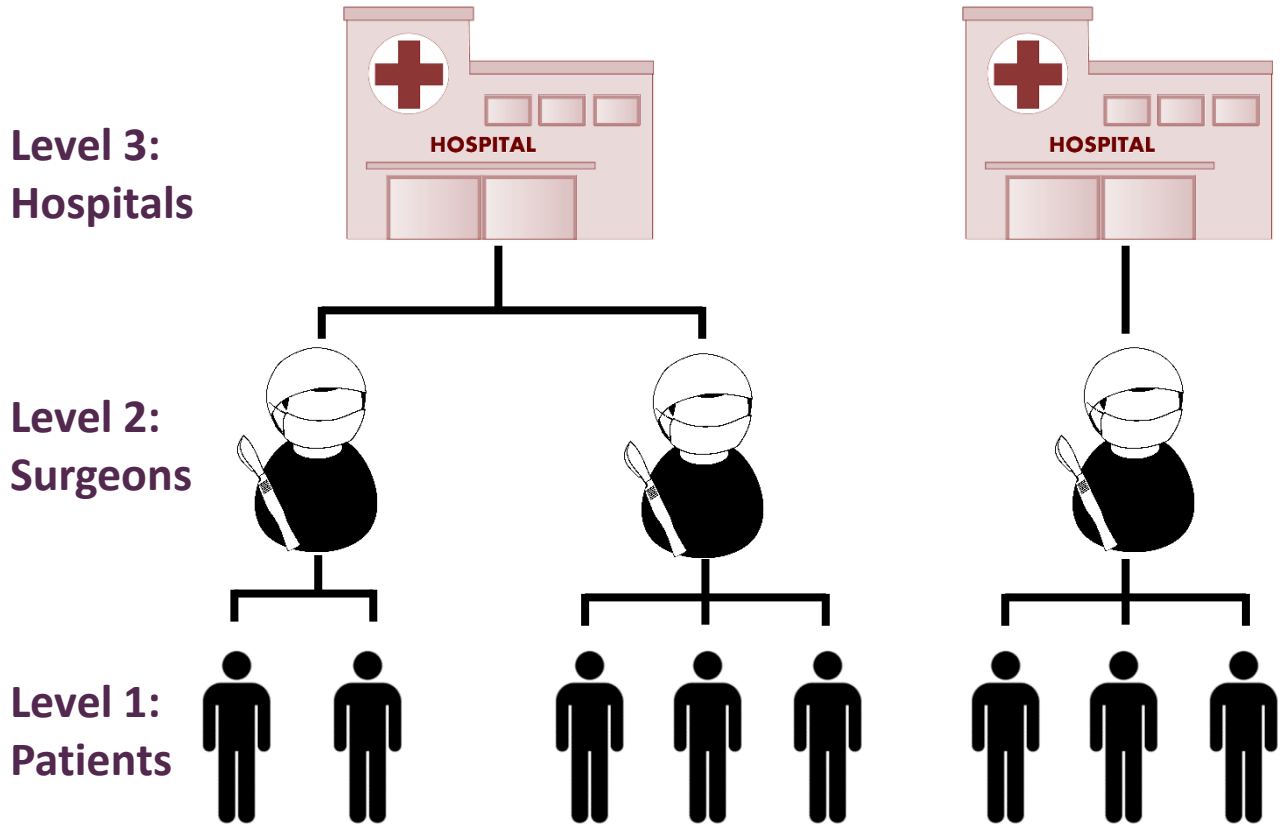
Can handle

- Missing data
- Nonuniform measures

Outcome Variable(s)

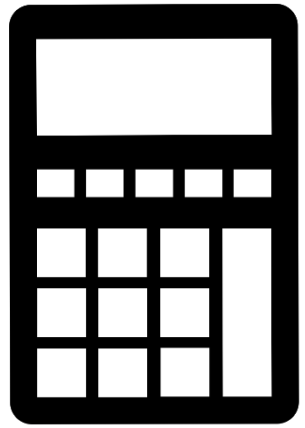
- Categorical
- Continuous
- Counts

## Nested Data




S-SPIRE

# Estimating Power



# Error Types

Type I Error (α): False positive

- Find an effect when it is truly not there
- Due to: Spurious association
- Solution: Repeat the study

Probability β = 0.20

**Type II Error β: False negative**

- **Do not find an effect when one truly exists**
- Due to: Insufficient power, high variability / measurement error
- Solution: Increase sample size

S-SPIRE

# Statistical Power

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

- Power = $1 - \beta$ (typically $1 - 0.2 = 0.8$)
- 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.

How many subjects do I need to find a statistical & **meaningful** effect size?

- Sample size calculation pitfalls:
  - Requires many assumptions
  - Should focus on the minimal clinically important difference (MCID)
  - If power calculation estimated effect size >> observed effect size, sample may be inadequate or observed effect may not be meaningful.

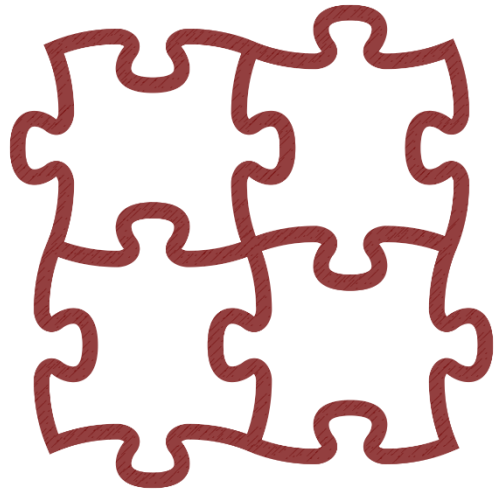
# Statistical Power Tools

## 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

# Components of 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 = m1 – m2 / pooled SD
  - Small d=0.2; Medium d=0.5; Large d=0.8
- Expected values 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?
- Inverse relationship with sample size
  - ↑ effect size, ↓ sample size
  - ↓ effect size, ↑ sample size

S-SPIRE

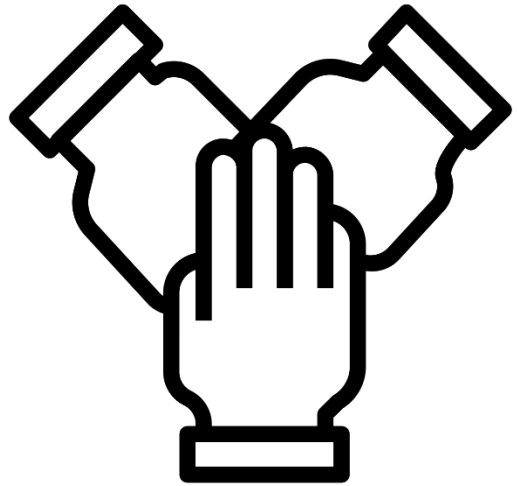
# Confidence Interval-Based Power

- How precisely can you estimate your measure of interest?
- Examples
  - Diagnostic tests: Sensitivity / Specificity
  - Care utilization rates
  - Treatment adherence rates
- Calculation components
  - N
  - Variability
  - α level
  - Expected outcomes

# 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

# 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**

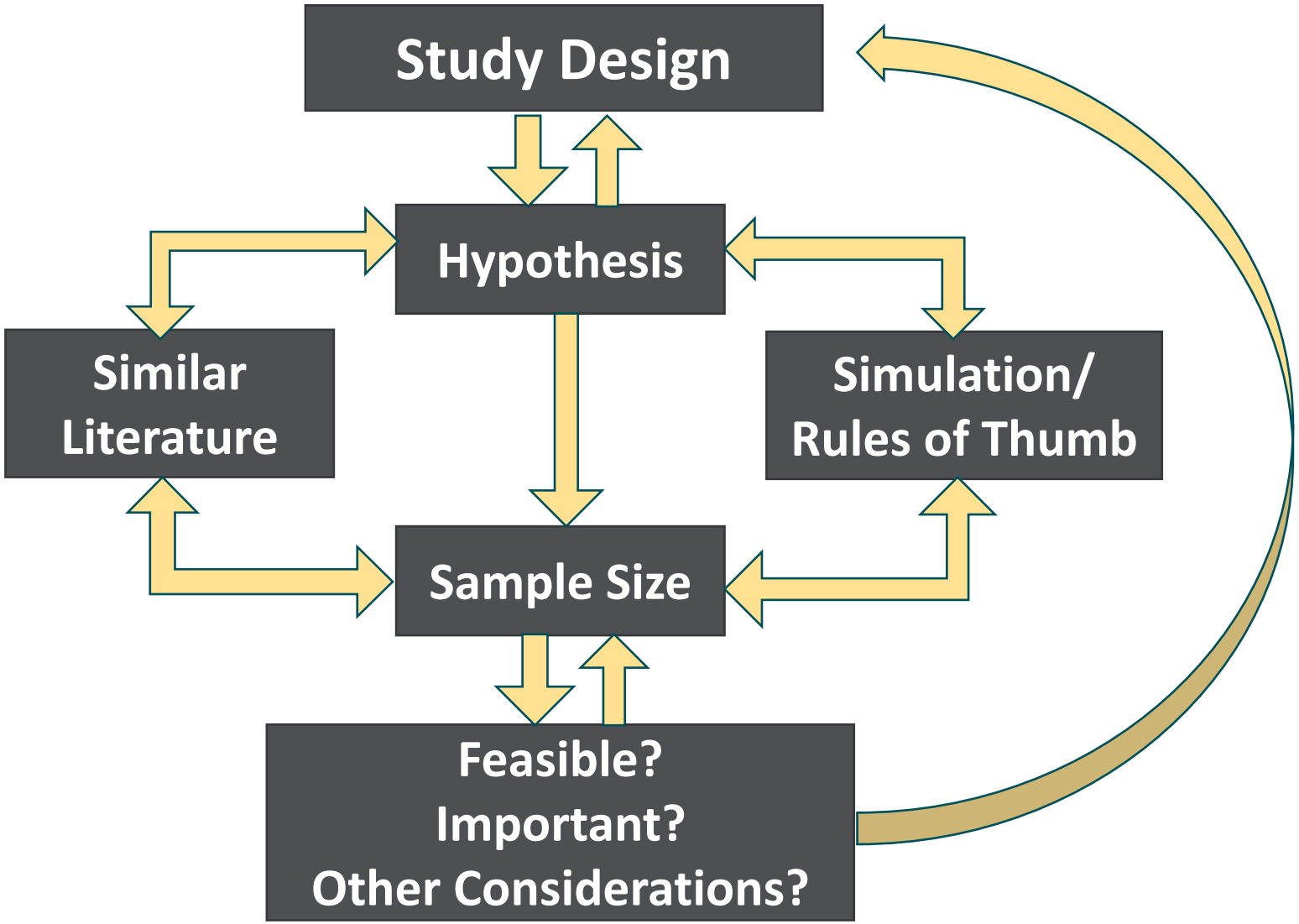
[Altman, 2002]

# 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

S-SPIRE

# Stages of Power Calculation




[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 minimal 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 w/ varying assumptions/conditions

*[Revicki, 2008]

# Authorship

International Committee of Medical Journal Editors (ICMJE) rules:

All authors must have…

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
- Must be discussed


# Consultation

Stanford MEDICINE | S-SPIRE CENTER
*Stanford-Surgery Policy Improvement Research & Education Center*

**http://med.stanford.edu/s-spire/research.html**

## S-SPIRE CENTER

- Faculty
- Research
- Monthly Work In Progress (SoM campus)


"Interested in research, but do not know how to begin..."

- **The 11 Steps to Conduct Outcomes Research**

## Practical Guides to Surgical Datasets

- **Military Health System Tricare Encounter Data**
- **National Trauma Data Bank (NTDB)**
- **National Surgical Quality Improvement Program (NSQIP) and Pediatric NSQIP**
- **Medicare Claims Data**
- **National Cancer Database (NCDB)**
- **Metabolic and Bariatric Surgery Accreditation and Quality Program (MBSAQIP)**
- **Society for Vascular Surgery**

## S-SPIRE Consultation Request Program

S-SPIRE

REQUEST A CONSULTATION

We are happy to help with consultation requests related to Health Services Research Projects. Please download and complete the form and email to s-spire_consult@stanford.edu and contact Ana Mezynski mezynski@stanford.edu to schedule an in-person 30-60 minute meeting

Download Form Here →

## Monthly Work In Progress (SOM CAMPUS)


Our monthly WIP session at Stanford School of Medicine features Stanford and guest faculty presentations of well-developed projects. This WIP takes place on the main campus and provides an opportunity to discuss high impact research, share a beer or soft drink with colleagues, and create synergy within the Stanford HSR/Surgery communities

See Schedule Here →

## Weekly Work In Progress (Arastradero)

Our weekly WIP sessions at 1070 Arastradero Rd feature trainees and faculty projects in every phase of development—from drafting specific aims pages, to parsing grant review committee comments, to abstracts/papers/methods in preparation. Anyone can attend and happy hour conditions apply here too

See Schedule Here →

S-SPIRE

# Timelines for Initial Consultation

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 consultations** will better inform grant development

# Summary

- Power calculations are complex, but S-SPIRE statisticians can help
- Effective statistical collaboration can be achieved
- Contact us early
  - power/sample calculations are iterative & take time
- Gather information prior to consult
  1. Study design
  2. Expected effect size
  3. Feasible sample size
  4. Similar literature
  5. Pilot data
- Come meet us at 1070 Arastradero!

S-SPIRE

# References


1. Farrokhyar F, Reddy D, Poolman RW, Bhandari M. Practical Tips for Surgical Research: Why perform a priori sample size calculation?. *Canadian Journal of Surgery*. 2013 Jun;56(3):207.
2. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of clinical epidemiology. 2008 Feb 1;61(2):102-9.
3. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. (2007). *Designing Clinical Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins.
4. Gordis, L. (2014). Epidemiology. Philadelphia: Elsevier/Saunders.
5. Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. JAMA. 2002 Jun 5;287(21):2817-20.
6. Pye V, Taylor N, Clay-Williams R, Braithwaite J. When is enough, enough? Understanding and solving your sample size problems in health services research. BMC research notes. 2016 Dec;9(1):90.


S-SPIRE