Planning a Study with Power and Sample Size Considerations in Mind

David Yanez

Professor, Biostatistics
OHSU/PSU School of Public Health

Biostatistics, Epidemiology,
Research & Design
(BERD) Seminar

May 29, 2019

Workshop Overview

- Welcome
- What this workshop is
  - An overview on how to approach scientific inquiries with consideration to
    - Design, outcomes, comparisons, effects, hypotheses, statistical tests, **information**, and **statistical power**
    - Sufficiently general, intuitive
  - Prerequisite: an introductory course in statistics

Motivating Example

Hypothetical study to evaluate a cure for Stage 4 pancreatic ductal adenocarcinoma (PDAC).

- **Population**: Stage 4 PDAC patients
  - Median survival for population is between 3-6 months
- **Design**: Double-blind, randomized controlled trial
- **Treatments**: Two (independent) groups
  - Gene therapy (Treatment)
  - Standard care (Control)

Motivating Example

- **Outcome (endpoint)**: Survival at six months (yes/no)
- **Scientific Summary**: Relative Risk (RR)
  - **RR** = P[survival | treatment] / P[survival | control]
- **Scientific Hypotheses**:
  - A. The *true* survival probabilities for the treatment groups *differ*
  - B. The *true* survival probabilities for the two groups *do not differ*

Motivating Example

- **Hypotheses**: In terms of the Scientific Summary
  - A. P[survival | treatment] $\neq$ P[survival | control] or **RR $\neq$ 1**
    - Also know as the *Alternative Hypothesis (H_1)*
  - B. P[survival | treatment] $=$ P[survival | control] or **RR $=$ 1**
    - Also know as the *Null Hypothesis (H_0)*
- **Method of Analysis**:
  - Use an exact statistical test to compare the two groups for the binary (survive versus did not survive) outcome
    - Fisher's Exact Test

Motivating Example

- **Decision Rule**: P-value
  - Definition: The probability of obtaining the observe test statistic result or a more extreme result *under the assumption that there is no difference in the rates of survival for the two treatment groups* (i.e., that the null hypothesis is true)
    - A p-value < 0.05 is compelling evidence against the two survival probabilities being equal (RR = 1); we would [reject the null hypothesis and] conclude the survival probabilities differ.
    - A p-value $\geq$ 0.05 is something we might expect to observe due to chance variation in the sample data; we would [not reject the null hypothesis and] conclude the survival probabilities do not differ.

Motivating Example

- **Data**:
  - Treated: 3 of 3 alive at 6 months
    - Estimated 6-month survival: 100% survival or P[survival | treatment] = 1.0
  - Control: 0 of 3 alive at 6 months
    - Estimated 6-month survival: 0% survival or P[survival | control] = 0.0
  - Relative Risk Estimate = 1.0/0 is infinite!
    - The risk of surviving six months for Stage 4 PDAC patients receiving the novel gene therapy treatment is infinite relative to the risk of surviving six months for patients receiving standard therapy.
  - The Fisher's Exact test two-sided P-value = 0.10
  - **Question**: How would you characterize the presence of an association between this treatment and Stage 4 PDAC?

Motivating Example

Comments:

- Our hypothetical gene therapy really was a cure for Stage 4 PDAC
- Assume the study design was perfectly executed
- The RR estimate showed overwhelming evidence of a treatment effect
- Unable to reject the null hypothesis and conclude a treatment effect

- **Question**: What are the problems, if any?

Motivating Example

Are there problems with

- the design?
  - Two (independent) group comparison
- the outcome?
  - A binary (dead/alive) characteristic
- the statistical test?
  - Fisher's Exact test
- the information?
  - Sample size N=6 (3 per group)?
- None of these study features separately were ruinous, but collectively, they doomed the inquiry

Motivating Example

What considerations might have been exercised in designing the study better?

- in the design?
  - one group “comparison”?
- different outcome(s)?
  - precise *quantitative* measure of cancer (CA19-9)
- a different statistical test?
  - chi-square test, permutation test?
- gather more information?
  - recruit > 6 patients?
- We should have had the ability to detect an effect…

Statistical Power

Formal:

- *Power* is the probability of rejecting a false null hypothesis.

$$\text{Power} = \Pr[\text{ Rejecting } H_0 \mid H_0 \text{ is false }]$$

Informal:

- Power is the probability that a statistical test can demonstrate there is a *difference* (e.g., *in survival rates between two treatment groups)* GIVEN THERE IS A DIFFERENCE.

Motivating Example

We were interested in

- comparing six-month survival, a binary outcome
- two indep. groups: treatment vs standard care
- had limited information: n=3 patients per group

These design constraints led to the use of

- Fisher's Exact test (to test the hypotheses)

Unfortunately, it was not possible to conclude there was an effect of treatment for FDA evidentiary standards for drug approval (i.e., p-value < 0.05)!

Planning a Research Study

- Example template
  - **RPG/R01/R03/R15/R21 Review**
    If you cannot access the hyperlinks below,
    visit http://grants.nih.gov/grants/peer/critiques/rpg.htm.
    - Overall Impact (summary)
      - Strengths and weaknesses
    - Review Criteria
      - Significance
      - Investigators
      - Innovation
      - **Approach**
      - Environment

Additional Score Descriptors

Impact	Score	Descriptor	Additional Guidance on Strengths/Weaknesses
High	1	Exceptional	Exceptionally strong with essentially no weaknesses
	2	Outstanding	Extremely strong with negligible weaknesses
	3	Excellent	Very strong with only some minor weaknesses
Medium	4	Very Good	Strong but with numerous minor weaknesses
	5	Good	Strong but with at least one moderate weakness
	6	Satisfactory	Some strengths but also some moderate weaknesses
Low	7	Fair	Some strengths but with at least one major weakness
	8	Marginal	A few strengths and a few major weaknesses
	9	Poor	Very few strengths and numerous major weaknesses
Non-numeric score options: NR = Not Recommended for Further Consideration, DF = Deferred, AB = Abstention, CF = Conflict, NP = Not Present, ND = Not Discussed			
Minor Weakness: An easily addressable weakness that does not substantially lessen impact **Moderate Weakness:** A weakness that lessens impact **Major Weakness:** A weakness that severely limits impact			

Planning a Research Study

- **Approach**

  - Formal evaluation of study's research aims/questions
  - Statistical in nature

    1. "Quantification" of research aims
       - Study to *improve lung function* in CF patients
         - How to measure this (e.g., FEV_1, FVC)?
    2. Study design
       - How to structure investigation of research aims
         - Experimental versus observational
    3. Comparison of groups
       - Identification of groups, prediction making?

Approach to Science

- Formal evaluation of study's research aims/questions
- Statistical in nature

  - Careful specification of 1-3 provide key information and constraints in designing proper metrics, hypotheses and tests of a study's research aims

    1. Scientific outcomes, data types
       - Quantitative (FEV_1, SBP, CA19-9)?
       - Categorical (e.g., poor/fair/excellent)
       - Binary (e.g., dead/alive, threshold)

Approach to Science

- Formal evaluation of study's research aims/questions
- Statistical in nature

  - Careful specification of 1-3 provide key information and constraints in designing proper metrics, hypotheses and tests of a study's research aims

    2. Example study designs
       - Evaluation of treatment for weight loss
         - *Randomize* subjects to specific treatments
         - *Observe* subjects on specific treatments
           - Adjustment necessary?

Approach to Science

- Formal evaluation of study's research aims/questions
- Statistical in nature

  - Careful specification of 1-3 provide key information and constraints in designing proper metrics, hypotheses and tests of a study's research aims

    3. Comparison of groups
       - Evaluation of treatment for weight loss
         - *Randomize subjects* to specific treatments
         - *Randomize order* of treatments to subjects

Approach to Science

1. Scientific outcomes (endpoints) we select determine

    - What statistical summaries are reported

        - Quantitative outcomes (e.g., FEV1, SBP, wgt.)
            - Means or percentiles, geometric means

        - Binary outcomes (diseased/no disease, threshold)
            - Percentages or proportions, rates, relative rates

        - Categorical (ordinal/Likert, nominal/disease status)
            - Percentages or poportions, frequencies

Approach to Science

1. Scientific outcomes (endpoints) we select determine

    - What statistical tests are executed

        - Quantitative outcomes
            - T-tests or regression for means or geometric means, sign or Mood's tests for percentiles

        - Binary outcomes
            - Chi-square tests or logistic regression for proportions and rates

        - Categorical outcomes
            - Chi-square tests, proportion-odds models, multinomial regression

Approach to Science

1. Scientific outcomes (endpoints) we select impact

    - Precision and sensitivity for our comparisons

        - e.g., IMT of carotid artery for arterial disease
            - What to measure?
                - EKG, Ultrasound, MRI?
                - Var(EKG) > Var(Ultrasound) > Var(MRI)

        - e.g., fasting plasma glucose for metabolic disorders
            - How to measure?
                - Means for treatment groups?
                - Risks (> 126 mg/dL) for treatment groups?

Approach to Science

2. Study designs: Experimental vs. observational

    - How to best assess/evaluate mechanisms, efficacy, safety of treatments, devices, exposures

        - Experimental studies
            - controlled randomized trials, cross-over studies
            - easier to make case for causation

        - Observational studies
            - retrospective, cross-sectional
            - can identify associations, challenge to infer causation

Approach to Science

3. Statistical classification of scientific questions

- Statistics is primarily used to
  - **Compare groups** / detect associations
    - majority of our questions
  - Make predictions
  - Cluster cases or characteristics
  - Quantification of distributions

- Statistical tasks may overlap, but the kind of questions determine the types of methods used and how we address those questions

Example

- Study to evaluate novel statin to reduce ischemic stroke in high risk patients

- What “outcome” to measure?

    - IMT of common carotid artery (surrogate endpoint)
    - Incident ischemic stroke (clinical endpoint)
    - Composite (combination) endpoint

Example

- Study to evaluate novel statin to reduce ischemic stroke in high risk patients

- When to measure?

  - IMT
    - annually for 5 years?
    - end of study only?
    - collect pre-randomization values?

  - Ischemic stroke
    - at end of study (e.g., 5 years)?
    - in real time (time to stroke)?

Example

- Study to evaluate novel statin to reduce ischemic stroke in high risk patients

- How to measure?

  - IMT
    - Ultrasound, MRI
    - raw measures, threshold (e.g., > 15% improvement)
    - post only, post – pre designs
    - repeated measurements over time (e.g., longitudinal)

  - Ischemic stroke
    - end of study (e.g., 5 years), time to stroke
      - What about censored subjects, missing values?

Example

- Study to evaluate novel statin to reduce ischemic stroke in high risk patients

- Comparison groups

  - IMT
    - Assigned only to one treatment group
    - Assigned to one treatment then later assigned to other

  - Ischemic stroke
    - Assigned only to one treatment group
    - Can treatment assignment be reasonably crossed over?

Approach to Science

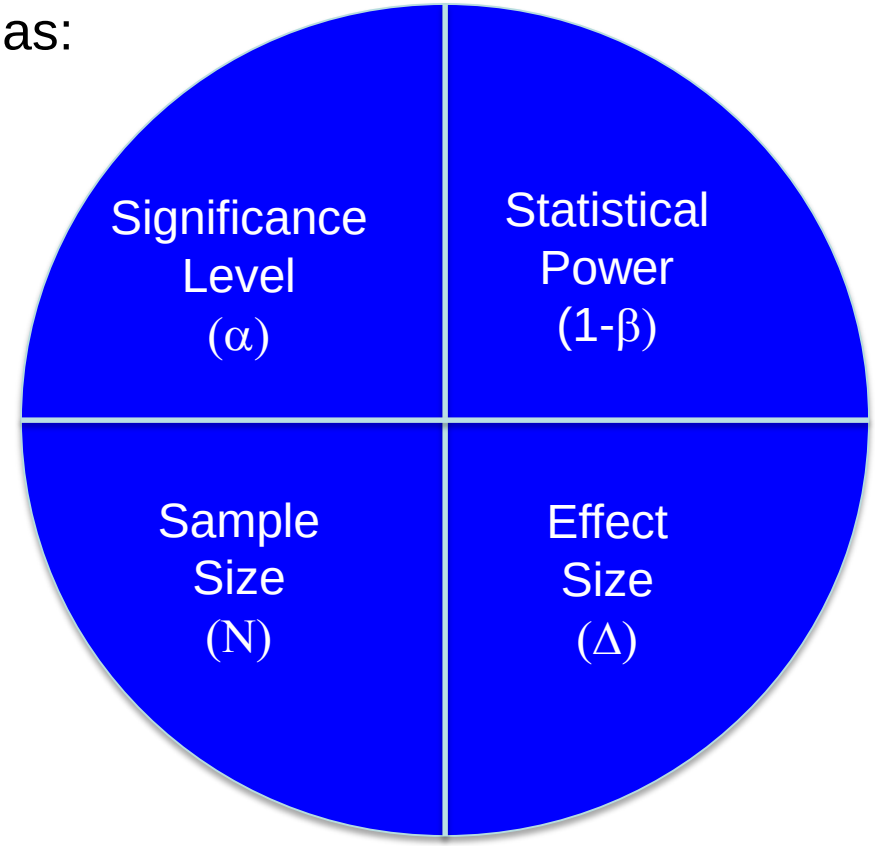
- Choices of outcomes, design considerations and comparisons are made to best address our scientific questions in the presence of constraints (e.g., design, cost, ethics, technology)
- Additional examples
  - Investigate effect of a treatment for a rare cancer, but there may be too few cancer events to perform a prospective longitudinal study (consider surrogate endpoints, other designs)
  - Investigate effect of a treatment for arteriosclerosis, but measuring IMT with MRI is not cost effective (consider less costly modality)
- Ultimately we are required to formulate our scientific questions into testable hypotheses
- At this point we are positioned to consider power and sample size issues

Power & Sample Size

- Statistical Power cannot be investigated without the specification of a hypothesis test
- Example requests
  - *I would like to characterize the temporal profiles of these two biomarkers for prostate cancer*
  - *I would like to quantify the distribution of HIV-2 in sub Sahara Africa*
- Both investigations may be impactful/meritorious, but without additional information that we can formally test, power considerations are challenging
- After we specify candidate outcomes, designs, comparisons and have formulated statistical hypotheses and tests, then we can consider the operating characteristics for power and sample size in study planning

Operating Characteristics

- Given hypotheses and a test statistic, the operating characteristics for sample size and power are summarized as:
  - Significance level (α)
  - Power ($1 - \beta$)
  - Sample size (N)
  - Effect size (Δ)



Operating Characteristics

- **Significance level (α)** – the design allowed probability of making a "Type I" or false positive error in our hypothesis test, Pr(reject H_0 | H_0 is true)
  - It is uniformly "fixed" at some small value (e.g., 0.05)
- **Power (1 – β)** – a design targeted probability of correctly rejecting the null hypothesis for a true "difference", Pr(reject H_0 | H_0 is false). It is 1 minus the probability of making a "Type II" or a false negative error
  - It is uniformly targeted to be 0.80 or higher.

Operating Characteristics

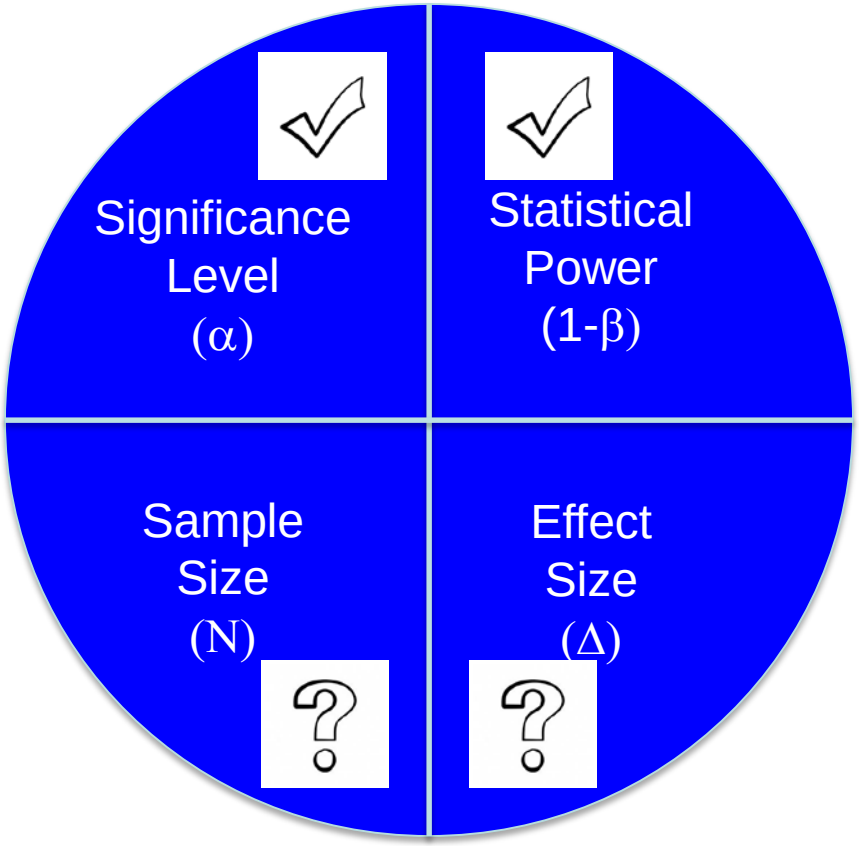
- **Sample size (N)** – a measure of the information in the study data. It is the operating characteristic tends to be most in the investigator's control

- **Effect size (Δ)** – the most scientifically meaningful quantity in the study, but the most challenging item for investigators to evaluate in planning a study

  - Examples of effect sizes
    - Quantitative endpoints:
      - Difference in treatment group means
      - Ratios of treatment group geometric means (logged values)
    - Binary endpoints:
      - Difference in treatment group probabilities (attributable risks)
      - Ratios of treatment group probabilities (relative risk)
      - Ratios of treatment group odds (odds ratios)
    - Time to event endpoints:
      - Ratios of treatment group hazards (hazard ratios)

Operating Characteristics

- If we can specify three of the four characteristics, we can often determine for the fourth
  - We typically don't know two, the sample size and effect size.
  - How might we proceed?



Operating Characteristics

Example exchanges

- Investigator: *I would like to conduct a study to evaluate whether there is an effect of this novel treatment on cancer outcome Y for study population Z… I really don't have a good idea what the reduction in disease incidence will be for patients receiving the treatment. How many patients should I enroll in my study to have 80 percent power for a two-sided, $\alpha = 0.05$ level test?*

- Statistician: *How many patients can you enroll and what is the disease incidence in the population?*

- Perhaps by learning what number of enrolled patients is actually feasible (e.g., annual number of cancer cases), and what the incidence is for untreated patients, the statistician could estimate a "minimum" incidence rate needed for the treatment group for the targeted level of power.

Operating Characteristics

Example exchanges

- Investigator: *I am investigating a treatment for CF pediatric patients, measuring FEV_1 as at six months. I can enroll 100 patients per treatment. I would like power to be 80 percent using a two-sample t-test. I do not know what the effect will be on this population, but it should be good. I don't have pilot data either.*

- Statistician: *Is it possible to obtain FEV_1 summary measures (means, SD's) on this population (e.g., published data)? Is it reasonable to look at these summaries for older CF or different pediatric populations?*

- Armed with everything but the effect size, the statistician could reasonably determine the minimum necessary difference in the means for the treatment groups (i.e., effect size), provided a decent estimate of the SD is available

Operating Characteristics

Example exchanges

- Investigator: *I am investigating a novel treatment on a novel biomarker endpoint. I would like to target power at 80 percent for a two-sample t-test. Could you provide sample size estimates assuming a small, medium and large effect size for a difference in treatment means? I have to use these effects because we have no data on this biomarker.*

- Statistician: *Is it possible to obtain "dimensionless" effect sizes using a standardized formula*

$$\Delta = \text{mean difference / SD}$$

- *where small, medium and large are often taken to be 0.2, 0.5, 0.8 standard deviations difference between means. It might be more prudent, however, to consider another approach to address the problem.*

Operating Characteristics

- Effect sizes: final thoughts
  - As investigators gain expertise in their particular science, the designs, comparisons, hypotheses and tests and the scientific summaries of interest (i.e., the effects) tend to become more familiar
  - The exercise in considering effect sizes can also take on “what if” scenarios
    - What is an effect size that must be detected
    - What is a biologically or clinically meaningful effect?
  - It is does get easier

Cautionary note

Post-hoc Considerations

- Some investigators (and even journal referees) often want to know the statistical power of a study *after* it is done. It is better to provide confidence intervals or other summary measures of precision from the sample results. Power and sample size considerations should be used as a tool for *planning* a study.

- It is not meant to be used as a post-mortem examination tool.

Final Thoughts

- Sample size calculations are only ESTIMATES, determined by a set of potentially variable assumptions
  - They may be crude metaphors of the models that will be used
  - The more crude they are, the more conservative they should be (i.e., resulting in the need for *more* information, larger sample sizes)
- It is recommended to provide *power curves* or tables, showing how the power and sample size estimates vary depending upon the different operating characteristics selected

Thank you