Power Analysis & Sample Size

Skeletal Biology and Pathophysiology
25SEP2020

Karen Steger-May
Division of Biostatistics

Acknowledgements

Deb Veis
Ken Schechtman

Use Chat to Participate

There will be some planned pauses for questions

Deb will interrupt when prudent with other questions

Why is sample size important?

- Sample size that is too big or too small
  - may be unethical; patients/animals…
    - undergo unnecessary risk
    - denied effective treatment
    - participate in a study that may not yield useful results
  - may waste time, money, and resources

- Calculations are required for grants, proposals, IRBs, journals

POWER vs. PRECISION

A comparative analysis compares outcomes/results/responses for different techniques/treatments/groups.

For a comparative analysis, sample size determination involves the power to test a null hypothesis.

POWER vs. PRECISION

Analysis goal is to estimate the magnitude of the effect, not the ability to reject the null hypothesis.

For estimation, sample size determination involves the precision of the estimate of the effect.

Determining Sample Size for Comparative Studies

A study must be “big enough” that a difference of scientific importance will also be statistically significant, but not so big as to detect differences so small that they are not even clinically meaningful.

Hypotheses are used to state the goals of a research study.

H_0 = null hypothesis that there is no difference

H_A = alternative hypothesis that there is a difference

- Statistical tests are used to distinguish between H_0 and H_A
- We can potentially make two kinds of errors...

1) find a difference when there is no true difference = Type I = **α**
2) fail to find a difference when a true difference exists = Type II = **β = 1 - power**

The power of a statistical test is the probability of finding a statistically significant difference when there is a true difference.

Statistical power depends upon:

1. How big the difference is
2. Variability
3. The significance level of the test
4. The sample size

Once any 3 are established the 4th is completely determined.

Sample size is the only item that researchers can control; and the only way to control Type I and II errors.

Steps to determine sample size:

1. identify 1-4 primary outcomes
2. specify the hypothesis
3. determine what statistical test will be used
4. decide the alpha and power level
5. determine feasible sample size(s)
6. estimate anticipated results
7. perform calculations

1: Identify 1-4 primary outcomes

- Primary vs. secondary outcomes
- Variables:
  - must be valid, reliable, and responsive
  - generic vs. disease-specific variables?
  - dichotomous variables require more sample size than continuous variables (don't categorize inherently continuous variables!)
- Endpoints:
  - comparisons at a single time point or change over time?
  - immediate effects or persistence of long-term effects?

2: Specify the hypothesis

- Testable prediction of the outcome
- Consider the study design
  - repeated measures?
  - within and/or between group comparisons?
  - all things being equal, paired analysis is more powerful
- Consider the number and types of groups
  - harder to find differences between very diverse groups
  - harder to find differences between less extreme groups
  - consider a control “normal” rather than different clinical levels

3: Determine what statistical test will be used

- one-tailed vs. two-tailed test
  - two-tailed tests interpreted if the effect meets the criterion in either direction
  - one-tailed, directional tests are more powerful, but very hard to defend
  - even though H_A is usually directional, the statistical test is usually not
- Consider assumptions…are data normally distributed or are transformations/nonparametric analyses needed?
- Are there other variables that may affect results?

3: *continued*

- Sophisticated analyses require larger sample sizes
  - Some studies are heavily dependent upon inferential statistics, others may not require statistical tests
  - Goal is the most parsimonious model with the least number of assumptions, particularly with small sample sizes
- Alternative techniques, such as confidence intervals (i.e., estimation), may be more valuable than inferential statistics

4: Decide the alpha and power level

- consider the consequences of Type I and Type II errors
- alpha is ‘usually’ 0.05
  - increase alpha if truly exploratory or pilot
    - want to avoid Type II error
    - reduces the sample size required
  - decrease alpha if definitive study or when multiple comparisons
    - want to avoid Type I error
    - increases the sample size required
- power should be at least 80% or 90%, sometimes even 95%

5: Determine feasible sample size(s)

- estimate recruitment using real data:
  - the number of willing eligible patients available during the recruitment period
  - the maximum sample size that you can manage (resources, staff, budget)
- estimate retention using previous studies:
  - the proportion of each group that will not provide “complete” data
  - for humans, ≥ 20% dropout rate is typical unless strong data otherwise
- consider generalizeability:
  - tight eligibility criteria may increase power but will decrease generalizeability

6: Estimate anticipated results

- Use the literature or your own research
  - with same or similar endpoints and patient/animal population/clinical setting that you will be studying
- Estimate:
  - means and standard deviations in each group at each time point
  - rates or proportions in each group at each time point
- Gather evidence that the proposed effect is reasonable
  - should be plausible (is a 60º change in ROM even possible?)
  - should be clinically meaningful (is a 1º change in ROM relevant?)

MCID = Minimum Clinically Important Difference

- consider the severity of the illness: a treatment that reduces mortality by 1% might be clinically important while a treatment that reduces transient complications by 20% may be of little interest
- consider treatment cost and side effects: a costly and high risk treatment would only be adopted if the treatment effect was substantial
- MCID must be defensible
  - choice is NOT statistical
  - you or someone on your team should be an expert
  - generic MCIDs and effect sizes should not be used for *a priori* computations...or used as a last resort

7: Perform calculations

At this stage, you are going to perform several sample size calculations to find the proper balance between statistical power and resources.

Once that balance is identified, you have the "final" sample size. The "final" sample size justification will often include only a subset of the calculations that were performed.

Prior Data Available (compare treatment with a control)

H_A: At 6-months post-op, patients who undergo the study treatment will have significantly less disability (ODI) as compared to patients who undergo the standard treatment.

- Prior data on 227 patients similar to those I will study suggest standard treatment patients will have ODI of 40 ± 16.
- Prior data on 24 pilot patients that underwent the study treatment suggest that mean ODI following treatment will be 28 ± 20.
- Prior data suggest a 15% dropout rate during a six month study

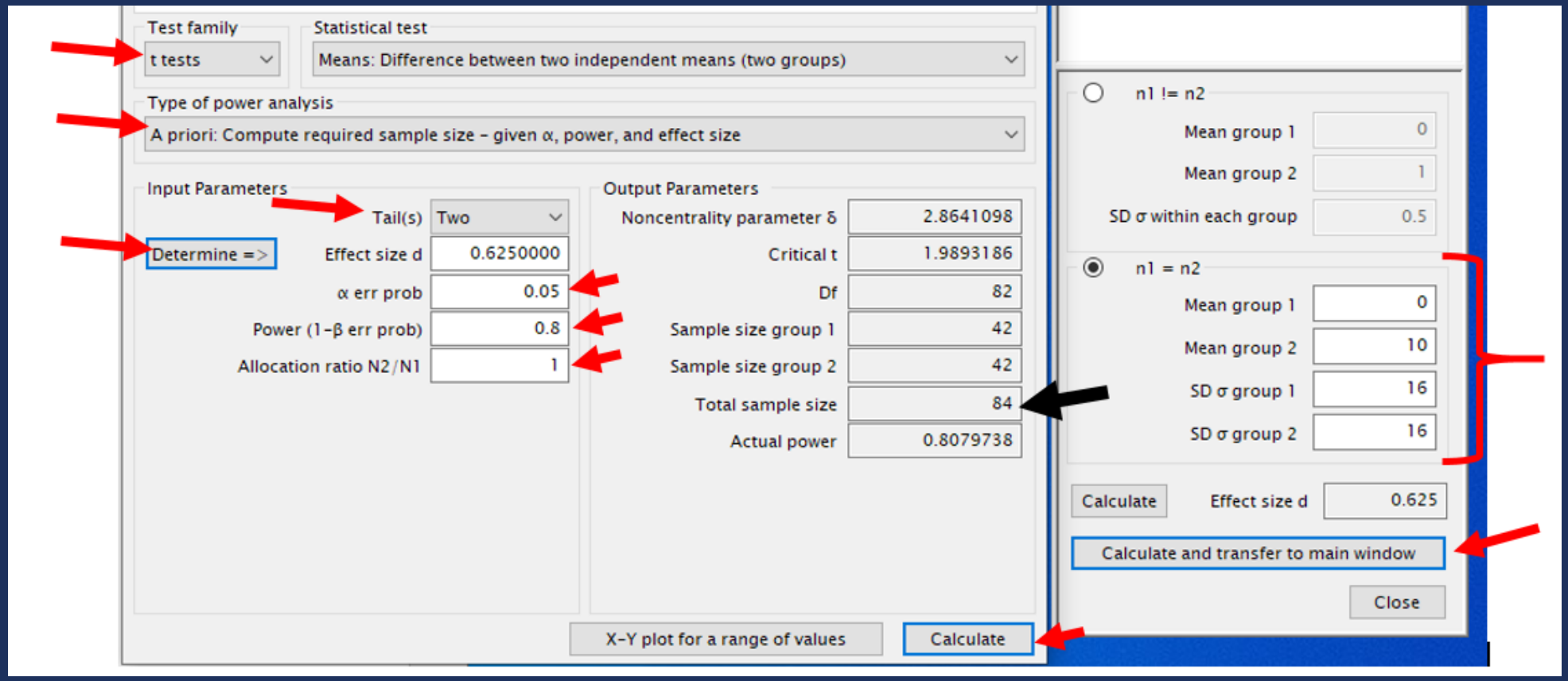
G*Power Demonstration

1. identify outcome(s) = ODI
2. specify hypothesis = At 6-months post-op, patients who undergo the study treatment will have significantly less disability (ODI) as compared to patients who undergo the standard treatment.
3. determine statistical test = two-tailed unpaired t-test, assume normality or transformed normal
4. decide alpha and power = α=0.05, power ~0.8-0.9
5. determine feasible sample size(s) = N_{total} ~150-220
6. estimate results = difference ≥10 (MCID), SD_{group} ~16-20
7. perform calculations

G*Power Demonstration

Assume:

1) α = 0.05, two-sided unpaired t-test
2) outcome is normally distributed
3) detect a minimum 10 point difference between groups = MCID
4) standard deviation (SD) 16-20
5) 1:1 allocation ratio of treatment and controls, N_{total} 150-220



G*Power Demonstration

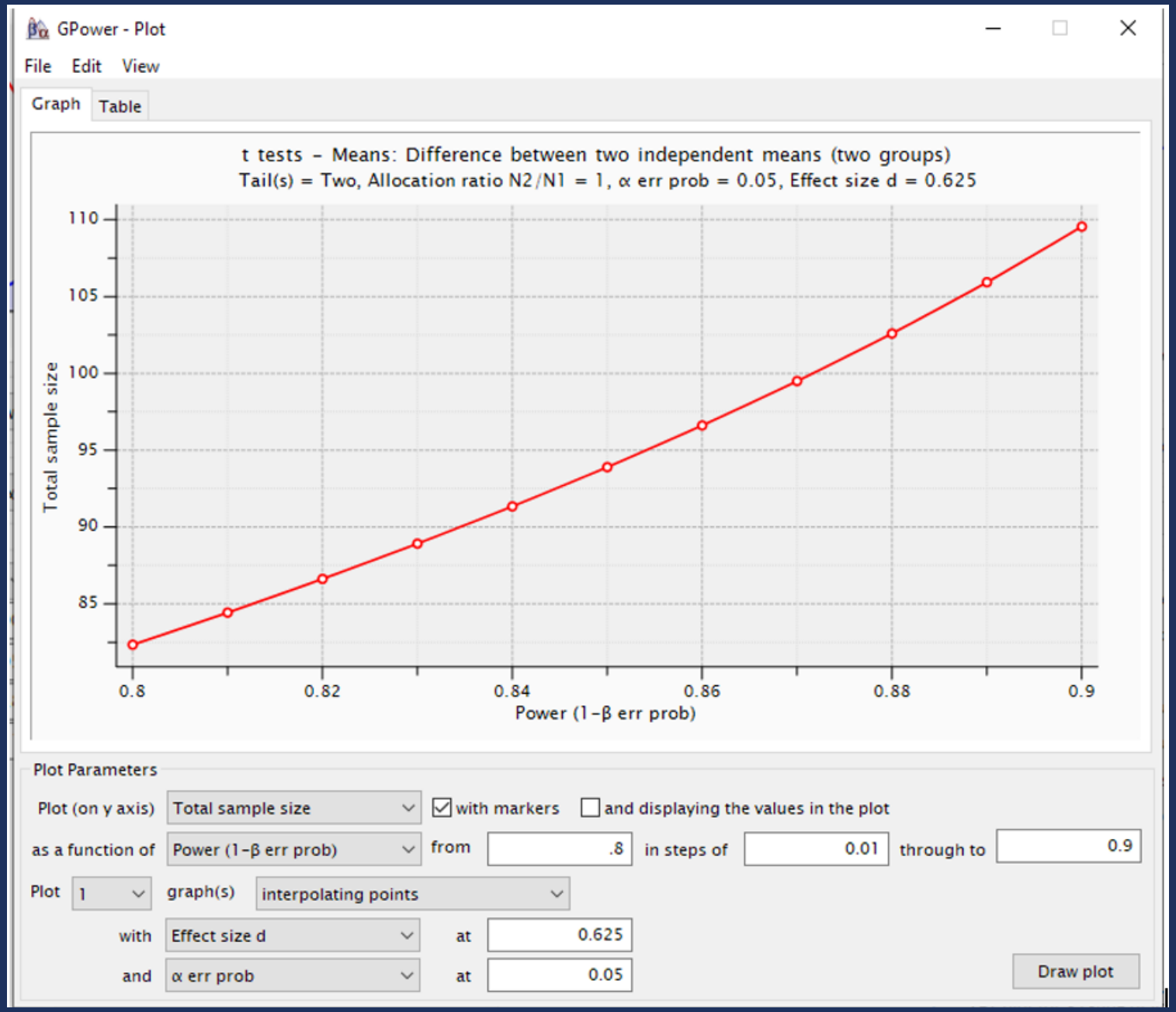
Assume:

1) $\alpha = 0.05$, two-sided unpaired t-test
2) outcome is normally distributed
3) detect a minimum 10 point difference between groups= MCID
4) standard deviation (SD) 16-20
5) 1:1 allocation ratio of treatment and controls, N_{total} 150-220

Model	β	Power	SD_1	SD_2	Effect Size (mean difference/SD)	N_{total}
1	0.2	0.8	16	16	0.62	84
2	0.1	0.9	16	16	0.62	110
3	0.2	0.8	16	20	0.55	106
4	0.1	0.9	16	20	0.55	140
5	0.2	0.8	20	20	0.50	128
6	**0.1**	**0.9**	**20**	**20**	**0.50**	**172**

Using conservative assumptions of a between-group difference of 10 ± 20, for 90% power I need 172 subjects (86 subjects per group) with complete data. Thus, I inflate the 172 by 20% = 206 subjects (103) per group.

Models 1-2: Change in N_{total} as a function of power 0.8 to 0.9



Reporting

- statistical test used for the power calculation
- one- or two-sided test
- variability estimate with justification
- the detectable effect with justification that it is plausible
- alpha level
- power
- consideration of dropouts/missing data
- justify recruitment feasibility

Reporting (compare a treatment with a control)

Power computations are based on a two-sided unpaired t-test at the 0.05 level of significance. They focus on the ODI, the primary outcome measure for the proposed research. Computations related to the comparison between control and treatment ODI are based on a preliminary study of 227 patients over the age of 65 who underwent a standard treatment for rod failure, and 24 patients that underwent the novel rod failure revision treatment *(insert ref for each)*. These patients are analogous to our control and treatment groups, respectively.

The preliminary data yielded a control value of 40 ± 16 and a treatment value of 28 ± 20. We conservatively assume that our patient groups will have a mean difference in ODI of at least 10 ± 20, which corresponds to the ODI MCID of 10-points *(insert ref for MCID)*. Based on these data, we will require a total sample size of 86 per group to achieve a power of 0.9.

If we use preliminary data *(insert ref)* to assume a 20% dropout rate during the study, we will need to enroll 103 patients per group to achieve a power of 0.9.

A sample size section should include a discussion of the availability of subjects, with the bottom line goal being to demonstrate that the number of available and willing eligible subjects can be expected to substantially exceed the number that are required for the study during the target recruitment period.

An evaluation of the patient population during the last four years indicate that, on average, 529 eligible patients are seen by this physician each year. Thus, even if only 40% of eligible subjects agree to participate in the study, far less than the 70% experience suggests will be the case, we can still recruit more than enough eligible subjects during the planned one year recruitment period.

Just in case scenario: An additional physician has agreed to join the study in the unlikely event that the above patient availability and recruitment estimates do not turn out to be the case in practice.

What if there are no prior data?

- Evidence why there are no prior data (e.g., very novel treatment, population never evaluated, etc.).
- Explore interplay between plausible differences, clinically significant differences, feasible sample sizes, and associated power levels.
- Report range of detectable differences given certain assumptions.

Examples that Focus on Estimation

Goal is to estimate the incidence of pseudotumours 10 years after THA. Assume the observed proportion of positive cases to be 0.4. With a sample size of 30, the 95% CI on the true proportion will range from 0.25 to 0.58, that is, within ~17% of the true incidence (precision ±17%).

A sample size of 50 is needed to estimate the correlation between serum cobalt levels and the volume of pseudotumours with a 95% CI of ±0.15 points, assuming an observed correlation of 0.70 (95% CI on the true correlation 0.52 to 0.82).

Analyses that Focus on Estimation

Estimate a mean/rate/correlation with a specified level of precision and confidence.

- Goal is precision, not statistical significance
- Margin of error is the level of precision that you require
  - it is the range in which the "true" population value is estimated to be
  - "confidence interval (CI)"
- Confidence is the probability that the margin of error contains the "true" value
- The larger the sample
  - the more certain you can be that the estimate reflects the "true" population
  - the narrower the CI

Determine the sample size required to estimate the statistic with the required margin of error and confidence level.

Key Points

- Sample size calculations must
  - ✓ be based on solid reasoning
  - ✓ include data on which assertions are based
  - ✓ consider feasibility
- A sample size justification must parallel
  - ✓ hypotheses
  - ✓ study design
  - ✓ statistical analysis plan

Sample size computations are rarely just plugging numbers into a formula.

Validated and Free Calculators

- **G*Power**

  https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html/

- **Power and Sample**

  http://powerandsamplesize.com/

- **Vanderbilt: PS Power and Sample Size Calculation**

  http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

Resources

- UCLA: Tutorials, examples using various software packages https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/
- Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. *J Bone Joint Surg Am* 1999; 81(10): 1454-1460.
- Zlowodzki M, Bhandari M. Outcome measures and implications for sample-size calculations. *J Bone Joint Surg Am* 2009; 91(Suppl 3): 35-40.

Know When to Get Help

Division of Biostatistics

Biostatistical Consulting Service

https://biostatistics.wustl.edu/consulting/

http://bit.ly/wubios

Power Analysis Workshop Oct. 2

- **identify 1 primary outcome**: Specify the type of variable (e.g., dichotomous, continuous) and the endpoint (e.g., single time point, change over time).
- **specify the hypothesis:** Include a brief summary of the study goal, the predicted outcome, and the variable/groups being compared.
- **determine what statistical test will be used:** Consider the assumptions of the test (e.g., normality) and directionality (one- vs. two-tailed test).
- **decide the alpha and power level:** Do you need alpha to be more liberal (e.g., 0.10) or conservative (e.g., 0.01)? Do you need power to be greater than 0.80?
- **determine feasible sample size(s):** Consider availability of the sample, observations that may not provide complete data, and generalizeability.
- **estimate anticipated results**: Consider variability and plausibility of estimates. Are your estimates from your own research of the literature?
- **perform calculations:** Perform several calculations to find the proper balance between statistical power and resources.
- **describe:** Write a 1-3 sentence summary of the power analysis that is suitable for a grant proposal.