Experimental Design and Statistical Power

LISA BRAMER

PNNL

2019 SRP Workshop

Experimental Design Considerations

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

- 3 R's of Experimental Design
  - Replication
  - Randomization
  - Control (Reduce variability and confounding variables)

- Replication is needed to establish statistical significance in any analysis
  - An increased number of replicates is necessary to guard against loss of statistical significance due to sample-to-sample variation or other technical problems
  - Balance between cost and quality data – If the data can't be used, then reduced price is not worth the effort

Tests of Significance

- Hypothesis = claim/belief that we wish to test
- 2 competing hypotheses
  - Null (H_0) → what we assume to be true
  - Alternative (H_A) → what we want to show
- "Innocent until proven guilty"
  - Assume H_0 is true until/unless we have enough evidence in the data in favor of H_A

Tests of Significance

- H_0 & H_A stated in terms of some population parameter, *p*
- Different types of hypotheses, e.g.

2-sided	1-sided	1-sided
H_0: $p = p_0$	H_0: $p \geq p_0$	H_0: $p \leq p_0$
H_A: $p \neq p_0$	H_A: $p < p_0$	H_A: $p > p_0$

- Note:
  - "=" sign always included in H_0
  - H_0 & H_A must contradict each other

Errors in Hypothesis Testing

- Rejecting H_0 in favor of H_A does not guarantee that H_A is true despite very strong evidence
- For any hypothesis test, there are 2 kinds of errors we can make

		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0	Correct decision	Type I error = α (false positive)
	H_A	Type II error (false negative)	Correct decision (power)

Errors in Hypothesis Testing

- By construction, hypothesis testing limits the rate of Type I errors (false positives) to a significance level, α
  - We choose ahead of time what significance level we will test at
  - Multiple tests on different attributes of the same data require an adjustment in order to preserve the significance level
    - E.g. testing the salmon for levels of multiple chemicals requires an adjustment
    - The more tests we do, the more likely we are to find a false positive
- Type II error rate (1 - power) is a function of sample size, significance level, & effect size
  - Trade-offs
    - In general, larger sample sizes allow us to detect smaller differences with more power
    - Standard deviation of the values also affects power

What is Statistical Power?

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

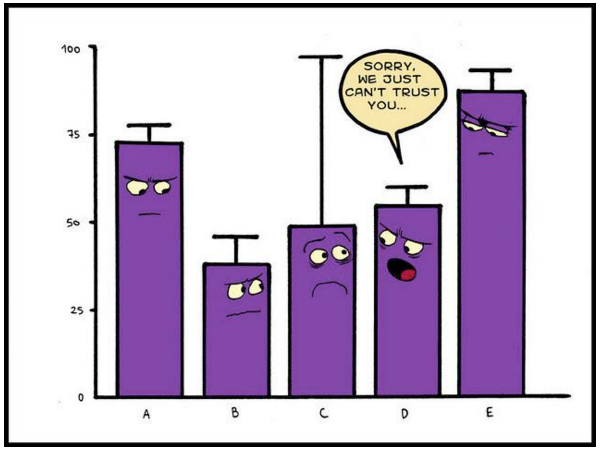
Statistical Power – probability of rejecting the null hypothesis, when the alternative hypothesis is true

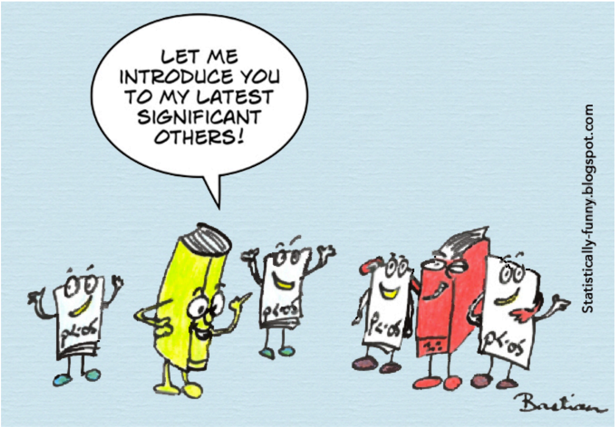
HYPOTHESIS TESTING OUTCOMES

Research	Reality: The Null Hypothesis Is True	Reality: The Alternative Hypothesis is True
The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

Why Worry About Power?

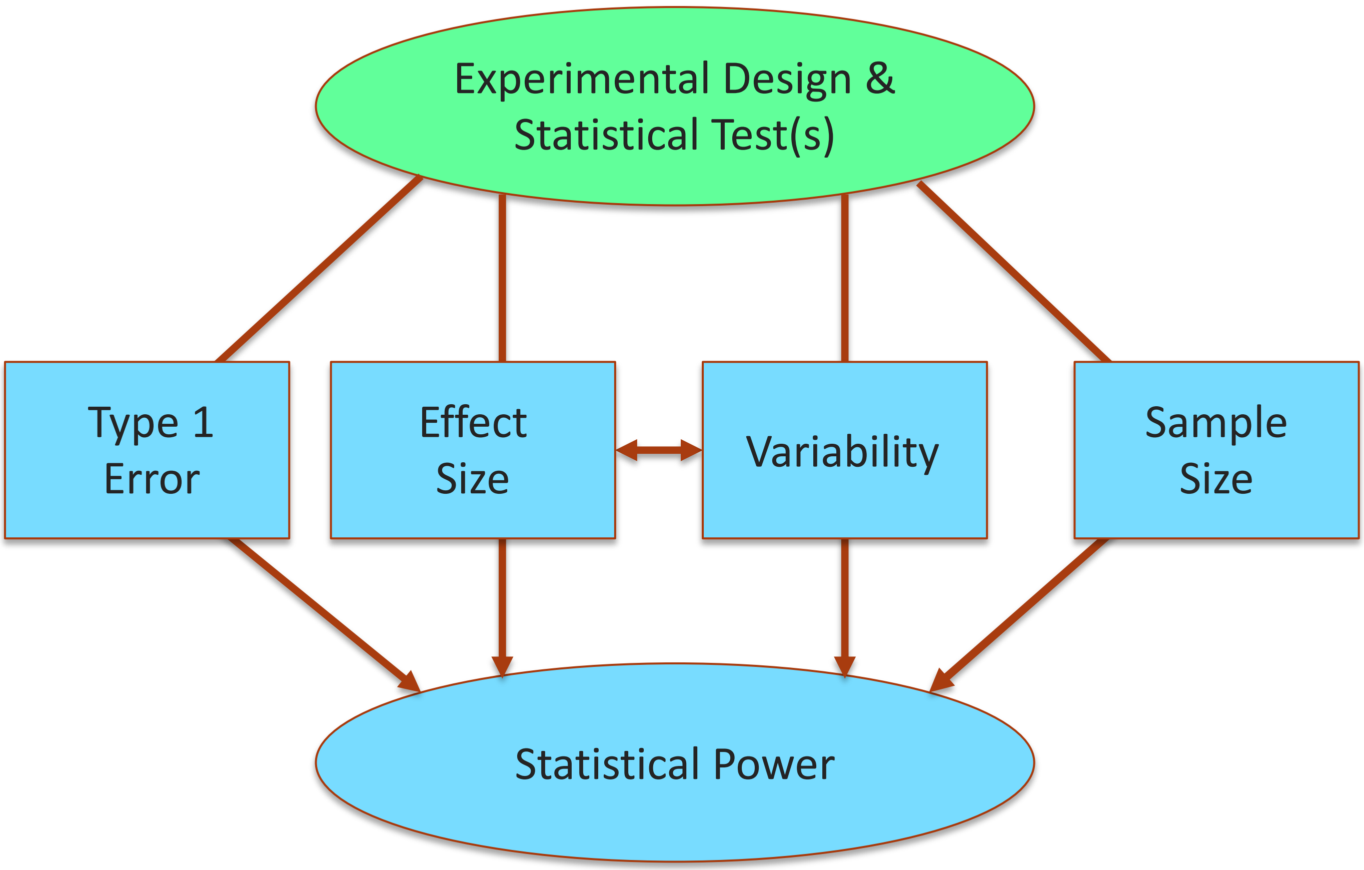
- Scientifically Meaningful Effect
  - Want enough samples to detect a scientifically meaningful effect
- Money
  - Undersized study wastes resources by not having capacity to produce useful results
  - Oversized study uses more resources than necessary
- Ethical issues when using live subjects (humans, etc.)
- Grant reviewers (and IRBs) are looking for sample size and power calculations





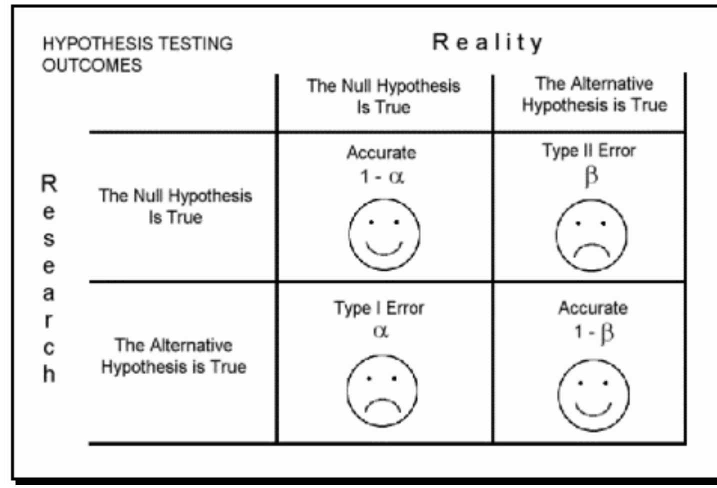
IT WAS GETTING HARDER AND HARDER TO FIND A TRULY MEANINGFUL RELATIONSHIP AT THE MEDICAL JOURNAL HAPPY HOUR.

Statistical Power



Type 1 Error

- Tradeoff between Type 1 and 2 errors
  - Inversely related
- Lower Type 1 error ➔ Lower power
  - Holding other factors constant
- Typical values: 0.05, 0.1
  - Domain dependent



HYPOTHESIS TESTING OUTCOMES		Reality: The Null Hypothesis Is True	Reality: The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
Research	The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

Effect Size/Variability

- Effect Size – desired detectable difference (if a difference exists)
- Variability – variability of the parameter being estimated
  - If evaluating difference in means ➔ sd of values from same group
- Ratio of Effect Size/Variability determines power
  - More variability ➔ less power
  - Smaller effect size ➔ less power
- Typical values?
  - Determined by data and/or domain knowledge

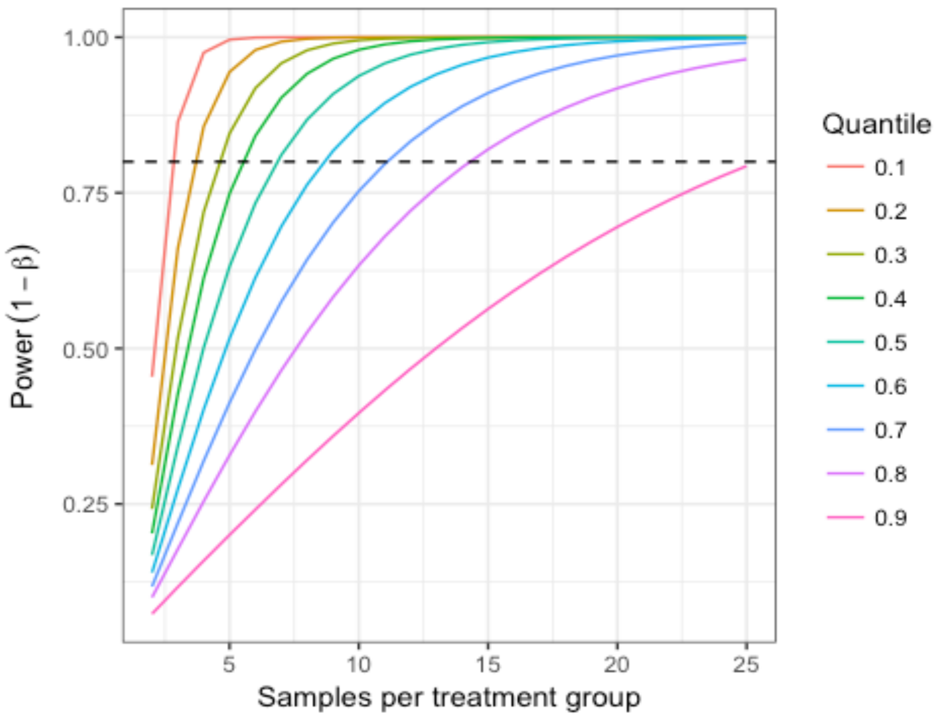
Sample Size and Power

- Usually the quantities we want to estimate
  - If I have X number of samples, what is my power?
  - How many samples do I need for 80% power?

- Most domains, we have one estimate of variability ➔ one power calculation
- Biology ‘omics data often have multiple response variables

Omics Data

- Many measurements ➔ many variability estimates ➔ many power calculations
  - 200 metabolites ➔ 200 power calculations



Given 80% Power

N	SD_Quantile
2	0.008
3	0.136
4	0.236
5	0.328
6	0.424
7	0.504
8	0.560
9	0.628
10	0.652
11	0.696
12	0.728
13	0.752
14	0.784
15	0.804
16	0.816
17	0.832
18	0.844
19	0.844
20	0.860

Other Considerations

- Will you have missing data (e.g. proteomics)?
  - Power calculations assume no missing data
- Are you testing more than two groups?
  - Multiple test correction
- Hypotheses not dealing with means (e.g. trend analysis over time)
  - Variability is not straightforward calculation
  - Example data is key

Tips

- Contact your favorite statistician before you plan replicates
- Groups vs Replicates
  - If determining how to allocate resources - More replicates per group, rather than adding more groups

Example

- Claim: The mean level of benzo(a)pyrene in tipi-smoked salmon is at least that of shed-smoked salmon

We will smoke samples of *n* salmon in a tipi to compare to the level of benzo(a)pyrene in shed-smoked salmon