Experimental Design Principles for Bioinformatics Analyses

Bruce A Craig
Department of Statistics
Purdue University

My Background

- Professor of Statistics (at Purdue 20+ years)
- Director, Purdue's Statistical Consulting Service (SCS)
  - Free service available to all associated with Purdue
  - Service balances research and teaching
    - Education and training of student consultants (and clients)
    - Improving the quality and productivity of research at Purdue
- Prior to becoming Director, had been involved in the design and analysis of spotted microarrays and QTL analysis in polyploids
- Now just try to keep up with the ever-changing technologies within my SCS role

Importance of experimentation

- “There are three principal means of acquiring knowledge… observation of nature, reflection, and experimentation. Observation collects facts; reflection combines them; experimentation verifies the result of that combination.” – Denis Diderot

- “The proper method for inquiring after the properties of things is to deduce them from experiments.” – Isaac Newton

Importance of design

- “You can't fix by analysis what you bungled by design.” - Light, Singer and Willett

- “To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.” - Sir Ronald Aylmer Fisher

Importance of design

- "All experiments are designed experiments, it is just that some are poorly designed and some are well-designed."

- “No experiment is ever a complete failure. It can always be used as a bad example.” – Paul Dickson

- “If your experiment needs statistics, you ought to have done a better experiment.” - Lord Ernest Rutherford

Experimental Design

- Is NOT simply a statistical issue…
- Requires a combination of
  - Scientific/biological insight
  - Scientific logic
  - Common sense
  - Planning
  - Statistics
  - Communication skills
    - https://www.youtube.com/watch?v=PbODigCZqL8

Not a one person task!!!

Experimental Design

- Cannot properly design experiment without also thinking about the analysis
  - Giving design its proper due saves plenty of frustration with analysis later on
  - Not just sample size determination
- Too often designs based on published studies
  - Replicate what has been successful makes sense
  - But is design the most cost effective?
  - Can we do better?

Experimental Design

- Basic design principles do not change with number of outcomes (e.g., genes, proteins), choice of technology platform, or size of the data
  - There are additional platform decisions (i.e., read length, number of reads, sequencing depth)
- "Keep it simple stupid" (KISS) principle often plays even more of a role in these bioinformatics studies due to
  - Costs involved
  - Complexity/size of data obtained

Basic Design Principles

- Randomization
  - "balancing out" effects of lurking/hidden variables
- Replication
  - Improving precision
- Blocking
  - Control the effects of nuisance factors that otherwise increase the noise in an experiment
  - Discuss options with service that will generate data
    - Multiplexing or incomplete block for RNA-seq experiments?

Treatment Structure Strategies

- One-factor-at-a-time (OFAT)
  - Favored when data cheap and abundant
  - Often very inefficient
  - Cannot investigate interaction
- Factorial treatment structure
  - Hidden replication advantage
  - Can assess interaction

Both can include blocking factors

Data = Treatment structure factors + Design structure factors + error

Factorial Treatment Structure

Factor Levels	B1	B2
A1	xx	xx
A2	xx	

Factor Levels	B1	B2
A1	x	x
A2	x	x

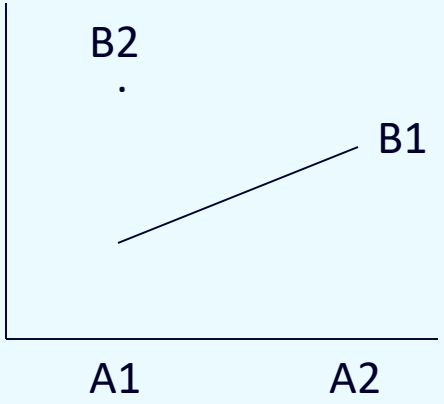
$n = 6$ Assume variance is σ^2 $n = 4$

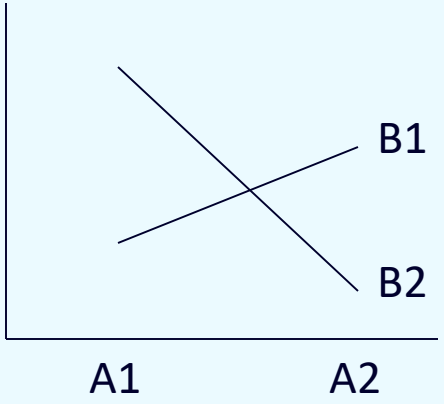
Effect of A $= \overline{\text{A2B1}} - \overline{\text{A1B1}}$

Var(Effect of A)$= {}^{2\sigma^2}/_{2} = \sigma^2$

Effect of A $= \overline{\text{A2}} - \overline{\text{A1}}$

Var(Effect of A)$= {}^{2\sigma^2}/_{2} = \sigma^2$





Nuisance Factors

- Imperative to consider all possible factors that could obscure or alter your results
- Often rely on sound judgement of center generating the bioinformatics data (i.e., temp, humidity)
  - Block on day or technician?
- Controlling these factors allow results to be as generalizable as possible
  - Sample split in half → Trt1 to one half, Trt2 to other
- Randomization provides protection against the unknown….helps avoid possible biases

Replication

- Biological versus technical replicates
  - Technical replicate: Measuring same biological sample multiple times
  - Technical replicates assess non-biological variation
  - May be beneficial if this variation expected to be large
  - Biological replicates typically improve precision more and allow conclusions to be more generalizable
  - Biological estimates usually more expensive

- Number of replicates depends on numerous factors
  - Cost and availability of resources
  - Desired precision / power

Replication

- Pooling
  - Used when sample material scarce
  - Larger sample → better precision of measurement
  - Danger in pooling
    - Measurement obtained from pooled sample can be different from the average of individual measurements
  - Better to avoid pooling if possible
    - More precision from multiple biological replicates

Calculating Power

- Numerous calculators available in software and online

- Be wary…you will always get numbers but whether they're meaningful depends on the quality of the inputted values
  - Trusting the quality of the inputs requires a basic understanding of the process

Why Power Analysis?

- Research is expensive…wouldn't want to conduct experiment with far…
  - too few experimental units (EUs)
    - Project won't find important differences that exist
    - Not worth the time and money
  - too many experimental units (EUs)
    - Project is unnecessarily too expensive
- Typical funding agency requirement
  - Demonstrates thinking about plan and organization

A Simple Experiment

- Study the effect of cold on a fat gene in rat
- Use a Completely Randomized Design (CRD):
  - 6 rats are randomly assigned to one of two different environments
    - Trt 1: Normal environment (20°C)…*n*=3
    - Trt 2: Cold environment (5°C)…*n*=3
- Investigator expects lower expression of gene when under Trt 2
- Is *n*=3 per trt enough to detect this difference?

Statistical Analysis

- Two competing hypotheses:

  H_0: $\log_2(\mu_1/\mu_2) = \log_2(\mu_1) - \log_2(\mu_2) = 0$

  H_1: $\log_2(\mu_1) - \log_2(\mu_2) > 0$ i.e. one-tailed test (for now)

- Basis for choosing between the two hypotheses
  - *P*-value quantifies degree of evidence against H_0
  - Compare *P*-value to significance level α, commonly α=0.05
    - $P \leq \alpha$ → reject H_o and conclude mean larger in Trt 1
    - $P > \alpha$ → fail to reject H_o , not enough evidence to conclude H_1
  - Report *P*-value and estimate / SE → not just significance
  - Small *P*-value only means H_0 can be rejected

Type I and Type II errors

		What the data indicate:	
		Fail to reject H_o: ($P>\alpha$)	Reject H_o: ($P\leq\alpha$)
True state	H_o: is true	No error	Type I error (Prob is α)
	H_1: is true	Type II error (Prob = β)	No error

So is *n* = 3 rats large enough?

- Rephrase: Do we have enough statistical power?
- Need to "know" several things
  - How large is the true mean difference ($\delta = \log_2(\mu_1/\mu_2)$)?
    1) What do you anticipate?
    2) What would be scientifically/practically important?
       - Suppose researchers believe that δ **=1 (fold change)**
  - How much variability (σ) exists between rats within a grp?
    - Some prior information potentially available from previously published studies or small pilot study. Variability due to rat and method
    - May also have to guess
      - Suppose researchers believe that σ **= 0.7**
  - Effect size (δ/σ) often used when scale arbitrary
- Power analysis involves "educated guessing"

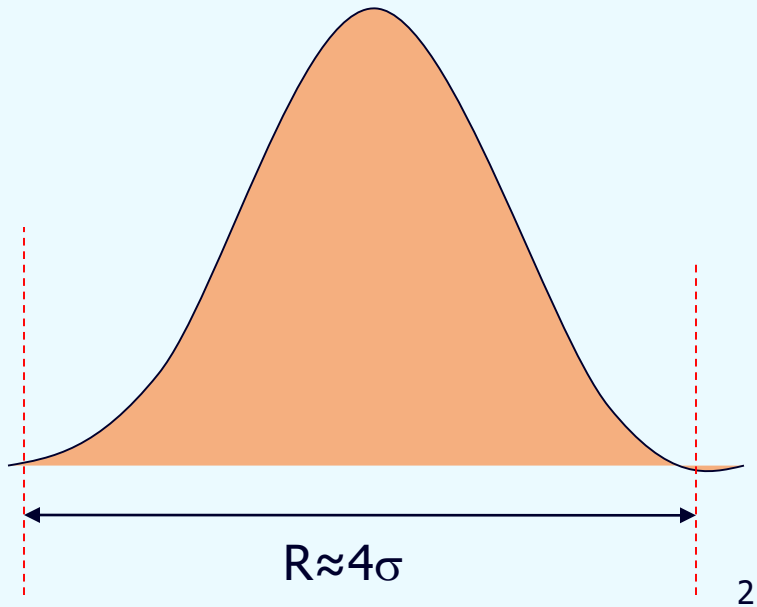
One way to elicit values for σ

- Use an empirical rule:

  Consider range of responses to be equal to 4σ or 6σ

- Question: What would be the likely range (max-min) of log expression levels for rats within the same trt?
  - Suppose the answer was 2.8
    - R = 2.8 → σ = 0.7

Can often find similar published studies with estimates of σ. Always round up to be a little conservative.



Two competing hypotheses:

- Under H_0 - $\bar{y}_2 - \bar{y}_1 \sim N\left(0, \frac{2\sigma^2}{n}\right)$
- Under H_1 - $\bar{y}_2 - \bar{y}_1 \sim N\left(\delta, \frac{2\sigma^2}{n}\right)$

Assuming *y* are the $\log_2$ expression levels. Difference in distributions is the mean.

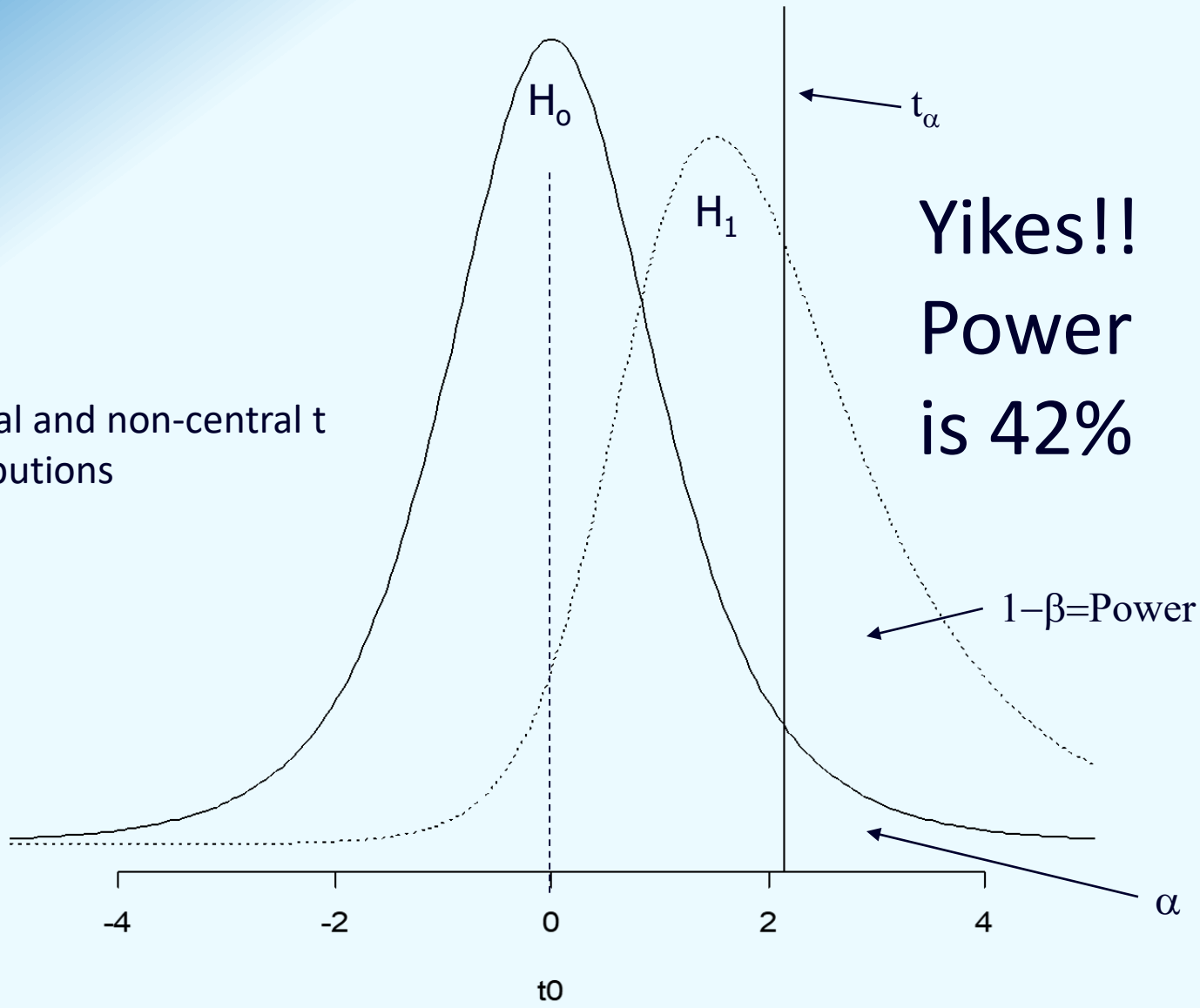
- Conduct one-tailed *t*-test for a certain α

Reject H_o: if $t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{2s^2/n}} > t_\alpha$

Need to study distribution of t_0 under H_0 and H_1

Distributions of t_0

Central and non-central t distributions



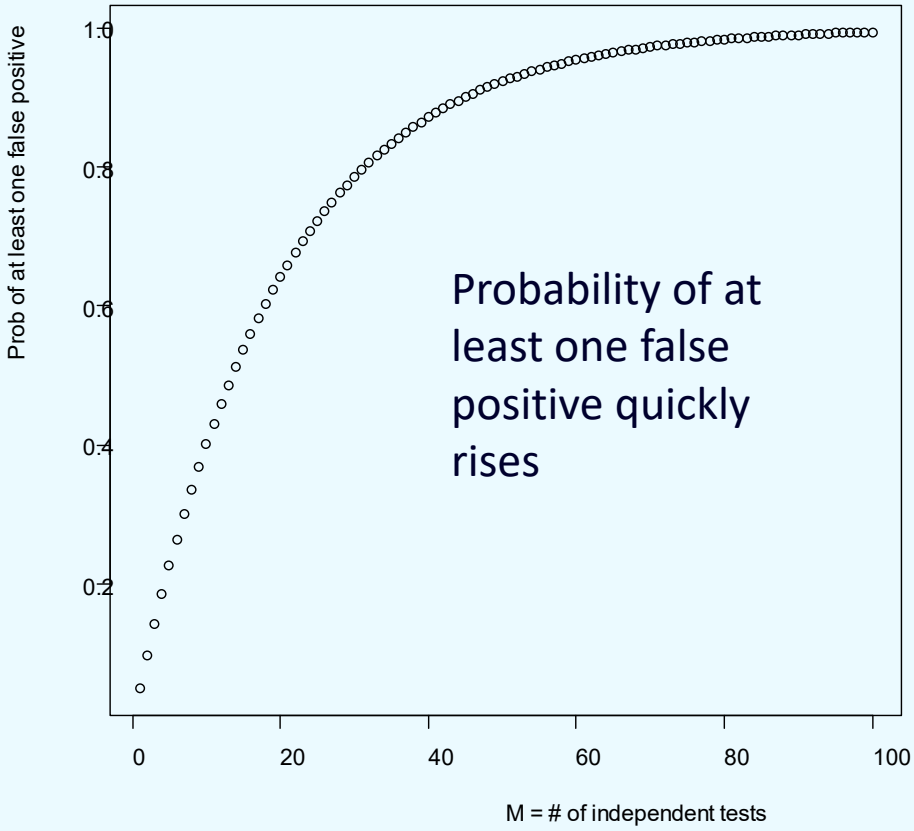
Yikes!! Power is 42%

Power Calculators

- Calculators available in software such as Minitab, SAS, JMP, and R
  - Be wary of calculators (such as PWR in R) that asks just for an effect size
  - Effect size essentially a signal versus noise ratio
  - Noise may be more than just biological variation
- Many calculators also take into account issue of multiple comparisons

Multiple Comparisons

- Testing changes in expression for **thousands** of features across **several** treatments



Multiple Comparisons

- Trade-off between control of false positive and false negative (power) rates
- Two common types of control on false positives
  - **Familywise error rate** : P(at least one) ≤ α

    Bonferroni (compare P-value to α/M)
  - **False discovery rate**: Expected proportion of rejected hypotheses rejected incorrectly

False Discovery Rate

- Proposed by Benjamini and Hochberg (1995)

	Not Rejected	Rejected	Total
True Null	U	V	m_0
False Null	T	S	$m-m_0$
	m-R	R	M

- FWER controls V
- FDR = E(V/R | R>0)Pr(R>0)

Microarray Myths and Truths

appeared: The Scientist, Vol. 16, No.17 (2002), p.22

- **Myths:**
  - "That complex classification algorithms such as neural networks perform better than simpler methods for class prediction
  - That **multiple testing issues can be ignored** without filling the literature with spurious results
  - That **prepackaged analysis tools are a good substitute for collaboration** with statistical scientists in complex problems."

- **Truths**
  - "The greatest challenge is organizing and training for a more **multidisciplinary approach** to systems biology. The greatest specific challenge is **good practice in design and analysis** of microarray-based experiments.

  - Comparing expression in two RNA samples tells you only about those samples and may relate more to sample handling and assay artifacts than to biology. Robust knowledge requires **multiple samples that reflect biological variation**.

  - Biologists need both good analysis tools and good statistical collaborators. Both are in short supply."

Summary

- Paying attention to design basics pays great dividends in both time and money
  - Issues do not go away with more complicated analyses or bigger data
- Seek expert help early

- Questions: bacraig@purdue.edu
  stat-help@stat.purdue.edu