Day 5: Experimental Design

Principles and Practice for Biological Research

Statistics for Biology Course

Core Philosophy

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*
>
> — *R.A. Fisher*

Good design > Clever analysis

Today’s Learning Objectives

By the end of today, you will be able to:

1. Understand the principles of good experimental design
2. Recognize and avoid common design flaws
3. Choose appropriate designs for different research questions
4. Calculate sample sizes and understand power
5. Implement randomization and blocking correctly
6. Handle practical constraints in biological experiments

Part I: Principles of Experimental Design

The Three Rs of Experimental Design

Replication

Multiple independent observations

Quantifies variability

Increases precision

Randomization

Controls for known & unknown confounders

Validates statistical inference

Eliminates bias

Reduction

Reduce error variance

Use blocking, controls

Increase power

Control Groups: What Do They Control For?

Purpose: Establish a baseline for comparison

Types of controls:

- Negative control: No treatment (what happens naturally?)
- Positive control: Known effective treatment (is system working?)
- Vehicle control: Treatment delivery method (e.g., DMSO, saline)
- Sham control: Procedure without active component (e.g., surgery without lesion)

Key principle: Controls should differ from treatment group in only one way — the factor being tested

Experimental Units vs. Observational Units

Experimental Unit
The smallest unit to which a treatment is independently applied

Observational Unit
The unit on which measurements are taken

Why does this matter?

- Your experimental unit determines your true sample size
- Multiple observations from the same experimental unit are not independent
- Getting this wrong = pseudo-replication

Example: Cell Culture Experiment

Scenario: Testing effect of drug on cell viability

Wrong:

- Apply drug to 1 flask
- Count cells in 10 wells
- Claim $n = 10$

Problem: Flask is the experimental unit, not wells!

True $n = 1$

Right:

- Apply drug to 10 independent flasks
- Count cells in 1 well per flask
- Claim $n = 10$

Solution: Each flask is an independent replicate

True $n = 10$

Example: Mouse Cage Effects

Scenario: Testing effect of diet on mouse weight

Wrong:

- Cage 1: 5 mice on diet A
- Cage 2: 5 mice on diet B
- Claim $n = 5$ per group

Problem: Cage is the experimental unit! Diet confounded with cage.

Right:

- 10 cages total
- Randomize diet to cages
- 5 cages per diet
- Average mice within cage

Solution: Cage is experimental unit, $n = 5$ per group

Pseudo-replication: The Biggest Problem in Biology

Definition
Treating non-independent observations as independent replicates

Consequences:

- Inflated sample size (too many “degrees of freedom”)
- Artificially narrow confidence intervals
- Increased false positive rate
- Misleading conclusions

How to avoid:

1. Identify your experimental unit before starting
2. If taking multiple measurements, average within experimental unit
3. Use appropriate statistical models (e.g., mixed models) if needed

Confounding

Definition
When the effect of the treatment cannot be separated from the effect of another variable

Example: Testing new teaching method

- Class A: New method, morning slot, experienced teacher
- Class B: Old method, afternoon slot, new teacher

Teaching method is confounded with time and teacher!

Solution: Randomization

- Randomly assign teaching method to classes
- Controls for known AND unknown confounders

Activity 3: Pseudo-replication Detective

Time for hands-on practice!

You will be given several study descriptions.

Your task:

1. Identify the experimental unit
2. Spot pseudo-replication
3. Calculate correct degrees of freedom
4. Suggest how to fix the design

Duration: 20 minutes

Part II: Sample Size and Power

Why Sample Size Matters

Underpowered studies:

- Waste resources (time, money, animals)
- More likely to miss real effects (false negatives)
- Published “significant” findings more likely to be false positives
- Contribute to irreproducibility crisis

Overpowered studies:

- Waste resources
- Detect trivial effects as “significant”
- Ethical concerns (unnecessary animal use)

Do power analysis BEFORE starting your experiment!

The Four Quantities

In any statistical test, four quantities are related:

1. **Sample size** (n): How many observations?
2. **Effect size** (δ): How big is the difference/relationship?
3. **Significance level** (α): Usually 0.05
4. **Power** ($1 - \beta$): Probability of detecting a real effect

Key Principle
If you know any three, you can calculate the fourth

Most common: Estimate effect size, set α and power (0.8 or 0.9), solve for n

Effect Size: Where Do We Get It?

Sources for effect size estimates:

1. Pilot study: Small preliminary experiment
   - Caution: Estimates are imprecise with small n
   - Use conservative (smaller) effect size for main study
2. Literature: Previous similar studies
   - Publication bias: published effects often overestimated
   - Look for meta-analyses
3. Biological importance: What size effect matters?
   - Not just statistical significance
   - Clinical/biological relevance
4. Feasibility: What can you realistically detect?
   - Given practical constraints on n

Power Analysis: Standard Conventions

Typical choices:

Parameter	Standard Value
Significance level (α)	0.05
Power ($1 - \beta$)	0.80 or 0.90

What does 80% power mean?

- If there is a real effect of the specified size…
- …you have 80% chance of detecting it ($p < 0.05$)
- …and 20% chance of missing it (false negative)

Higher power (0.90) when:

Example: RNA-seq Experiment

Question: How many biological replicates do I need?

Given information:

- Want to detect 2-fold change in expression
- Expected variability: CV = 0.3 (from pilot data)
- $\alpha = 0.05$, power = 0.80

Using power analysis:

- Need approximately $n = 6$ per group
- For 90% power: $n = 8$ per group

Note: This is biological replicates, not technical replicates!

When You Can’t Get Enough Samples

Practical constraints are real

If power analysis says you need $n = 100$ but you can only get $n = 20$:

1. Acknowledge the limitation
    - Report achieved power in paper
    - Interpret negative results cautiously
2. Consider alternative designs
    - Within-subjects instead of between-subjects
    - More efficient blocking
3. Focus on effect size estimates
    - Report confidence intervals
    - Contribute to future meta-analyses
4. Don’t do the study?
    - If hopelessly underpowered
    - Especially for animal studies

Activity 1: Power Analysis Workshop

Hands-on practice with power calculations

You will work through several scenarios:

1. Calculate required sample size from pilot data
2. Explore trade-offs between power, effect size, and *n*
3. Simulate data to verify power calculations
4. Real biology scenarios (RNA-seq, mouse studies)

Duration: 30 minutes

Part III: Common Experimental Designs

Overview of Common Designs

1. **Completely Randomized Design (CRD)**
    - Simplest design
    - Random assignment of treatments to units
2. **Randomized Block Design (RBD)**
    - Group similar units into blocks
    - Randomize within blocks
3. **Factorial Designs**
    - Test multiple factors simultaneously
    - Detect interactions
4. **Split-plot / Nested Designs**
    - Factors applied at different scales
5. **Repeated Measures / Crossover**
    - Subjects as their own controls

Completely Randomized Design (CRD)

When to use:

- Experimental units are homogeneous
- No obvious grouping of units

How it works:

1. List all experimental units
2. Randomly assign treatments to units
3. Apply treatments and measure outcomes

Advantages:

- Simple to design and analyze
- Flexible: any number of treatments, any group sizes
- Maximum degrees of freedom for error

CRD Example: Drug Testing on Cells

Scenario: Testing 3 drug concentrations on cell viability

Design:

- 15 cell culture flasks (experimental units)
- 3 treatments: Low, Medium, High drug concentration
- $n = 5$ flasks per treatment
- Randomly assign treatments to flasks

Analysis:

- One-way ANOVA
- Compare mean viability across treatments
- Post-hoc tests if significant

Key: All flasks treated similarly except for drug concentration

Randomized Block Design (RBD)

When to use:

- Experimental units are heterogeneous
- Can group units by some characteristic
- Want to reduce error variance

How it works:

1. Group units into blocks of similar units
2. Each block contains one unit per treatment
3. Randomize treatments within each block

Advantages:

- Removes variation due to blocks
- Increases power to detect treatment effects

What Makes a Good Block?

Good Block
Units within a block should be more similar to each other than to units in other blocks

Common blocking factors in biology:

- Litter (mice from same mother)
- Batch / Day (experiments done together)
- Location (position in incubator, field plot)
- Time (morning vs afternoon)
- Technician (who did the work)
- Sequencing lane (genomics)

Key principle: Block what you can, randomize what you cannot

RBD Example: Mouse Litter Effects

Scenario: Testing drug effect on tumor growth in mice

Problem:

- Mice from same litter are genetically similar
- Litter-to-litter variation is large
- This is nuisance variation

Solution:

- Use litters as blocks
- 2 treatments: Drug vs Control
- 8 litters (blocks)
- 2 mice per litter

Design:

- Within each litter, randomly assign one mouse to Drug, one to Control
- This removes litter-to-litter variation from error term

Analysis:

- Two-way ANOVA with blocking
- Or: paired t-test (difference within each litter)

Factorial Designs

Purpose: Test multiple factors simultaneously

Advantages:

- More efficient than separate experiments
- Can detect interactions between factors
- Generalizable results across factor combinations

Common factorial designs:

- 2×2: Two factors, each with 2 levels (4 treatment combinations)
- 2×3: Two factors, 2 and 3 levels (6 combinations)
- $2 \times 2 \times 2$: Three factors, each with 2 levels (8 combinations)

Key concepts:

Factorial Example: Diet × Temperature

Question: How do diet and temperature affect fly development?

Factors:

- Diet: Standard vs High-sugar (2 levels)
- Temperature: 20°C vs 25°C (2 levels)

2×2 **design: 4 treatment combinations**

1. Standard diet, 20°C
2. Standard diet, 25°C
3. High-sugar diet, 20°C
4. High-sugar diet, 25°C

Possible outcomes:

Interactions: Why They Matter

No interaction:

Effect of Factor A is the same at all levels of Factor B

Lines are parallel

Example: High-sugar diet slows development by 2 days at both temperatures

Interaction present:

Effect of Factor A differs across levels of Factor B

Lines cross or converge

Example: High-sugar diet slows development at 20°C but speeds it up at 25°C

Why it matters: Can't make simple statements about Factor A without considering Factor B!

Split-Plot / Nested Designs

When to use:

- Some factors applied at larger scale than others
- Practical constraints on randomization

Example: Agricultural field trial

- Whole-plot factor: Irrigation method (requires large plots)
- Sub-plot factor: Fertilizer type (can vary within plots)

Example: Cell culture with sub-sampling

- Whole-plot: Treatment applied to flask
- Sub-plot: Multiple wells sampled from each flask

Critical: Analysis must account for hierarchical structure!

Repeated Measures / Crossover Designs

Key idea: Each subject receives multiple treatments (at different times)

Advantages:

- Each subject is own control
- Removes between-subject variation
- Requires fewer subjects

Requirements:

- Washout period: Time for first treatment to clear
- No carryover effects: First treatment doesn't affect response to second
- Randomize order: Which treatment first?

Example: Before-after treatment design

Activity 4: Design Comparison

Comparing different experimental designs

Given the same research question:

1. Design experiments using CRD, RBD, and factorial approaches
2. Simulate data from each design
3. Compare power and efficiency
4. Identify which design is best for specific scenarios

Duration: 30 minutes

Part IV: Randomization & Blocking Strategies

Why Randomization Matters

Randomization controls for:

1. Known confounders (things you're aware of)
2. Unknown confounders (things you haven't thought of)
3. Selection bias (conscious or unconscious)

How it works:

- Random assignment ensures treatment groups are comparable on average
- Any differences between groups are due to chance, not systematic bias
- Validates use of probability theory for statistical inference

Critical Point
Randomization is what makes your statistical test valid!

How to Randomize Properly

Not randomization:

- Alternating treatments (A, B, A, B, A, B...)
- Treating "similar looking" animals similarly
- Processing treatment group first, then controls
- "I just mixed them up"

Proper randomization:

- Use a random number generator
- Assign treatments using randomization table/software
- Document your randomization scheme
- Do it before starting the experiment

In R:

What to Randomize

Randomize everything you can:

1. Treatment assignment (obviously!)
2. Physical positions
   - Plate positions in incubator
   - Animal cage positions on rack
   - Field plot locations
3. Order of operations
   - Order of sample processing
   - Order of measurements
   - Order of imaging
4. Technical details
   - Which technician does which samples
   - Which batch includes which samples
   - Which day for which treatments

Stratified and Block Randomization

Stratified randomization:

- Ensure balance on important covariates
- Example: Equal numbers of males/females in each treatment group
- Randomize within strata

Block randomization:

- Ensure balance across blocks
- Each block contains all treatments
- Example: Each litter has one animal per treatment

When to use:

- When complete randomization might yield unbalanced groups
- When you have identifiable sources of variation to control

Practical Randomization Examples

Example 1: Incubator plate positions

- Problem: Temperature gradients in incubator
- Solution: Randomize which treatment goes in which position
- Also: Rotate positions during experiment if possible

Example 2: Sample processing order

- Problem: Degradation over time, technician fatigue
- Solution: Randomize order of sample processing
- Don't process all treatment samples first, then all controls

Example 3: Animal cage positions

- Problem: Position effects (light, noise, access)
- Solution: Randomize cage positions on rack

Common Randomization Mistakes

1. **Pseudo-randomization**
    - Alternating treatments instead of truly random
    - Creates systematic patterns
2. **Forgetting to randomize measurement order**
    - Even if treatment assignment was random
    - Time-related confounding can occur
3. **Not accounting for spatial gradients**
    - Position matters (incubators, racks, fields)
    - Need to randomize or block
4. **Post-hoc “randomization”**
    - Deciding treatment assignment after seeing the units
    - Not truly random!

Temporal Blocking

Problem: Can't do all experiments at once

Day effects:

- Different days = different conditions
- Equipment calibration drift
- Different reagent batches
- Technician effects

Solution: Use day as a blocking factor

- Include all treatments on each day (each day is a block)
- Randomize order within each day
- Include day as a factor in analysis

Activity 2: Randomization Simulator

Hands-on randomization practice

You will:

1. Generate proper randomization schemes in R
2. Compare systematic vs random assignment
3. See how pseudo-randomization fails
4. Practice block randomization

Duration: 25 minutes

Part V: Avoiding Common Pitfalls

Major Pitfalls in Experimental Design

1. Pseudo-replication
    - Most common problem
    - Wrong experimental unit
2. Confounding
    - Treatment mixed with other factors
3. Batch effects
    - Systematic technical variation
4. Selection bias
    - Non-random exclusion of data
5. Insufficient power
    - Too few samples to detect effect
6. Multiple testing
    - Testing many outcomes without correction

Pitfall 1: Pseudo-replication (Review)

Example 1: 10 wells from 1 plate $\neq n = 10$

- Plate is experimental unit
- Wells are technical replicates
- True $n = 1$

Example 2: Multiple measurements from same animal

- Animal is experimental unit
- Repeated measurements are correlated
- Can't treat as independent

Solution: Always know your experimental unit!

- What is the smallest unit that independently receives treatment?
- Average within experimental unit if multiple observations

Pitfall 2: Confounding

Example 1: Spatial confounding

- All treated animals in Room A, controls in Room B
- Can't separate treatment effect from room effect

Example 2: Temporal confounding

- Process treatment samples first (morning)
- Process control samples later (afternoon)
- Treatment confounded with time of day

Solution: Randomization!

- Randomize treatment to rooms
- Randomize processing order
- Include potential confounders as covariates in analysis

Pitfall 3: Batch Effects

Definition: Systematic variation between groups of samples processed together

Common in genomics:

- Different library prep days
- Different sequencing runs
- Different technicians
- Different reagent lots

Problems:

- If treatments align with batches, confounding!
- Large batch effects can obscure biological signal

Solutions:

Example: RNA-seq Batch Effect Disaster

Poor design:

- Batch 1 (January): All disease samples
- Batch 2 (March): All control samples

Disease is completely confounded with batch!

Result:

- Can’t tell if differences are due to disease or batch
- Study is uninterpretable
- Wasted time and money

Better design:

- Batch 1: Half disease, half control samples

Pitfall 4: Selection Bias

Definition: Excluding data or subjects in a non-random way

Examples:

- Excluding “sick looking” animals after treatment starts
- Removing “outliers” that don’t fit your hypothesis
- Only reporting outcomes that show significant effects

Why it’s a problem:

- Biases results
- Invalidates statistical tests (based on random sampling)
- Often unconscious

Solution:

Pitfall 5: Insufficient Power

Consequences for science:

- Waste of resources (especially animals)
- Likely to miss real effects (false negatives)
- If significant result found, likely to be:
  - False positive, or
  - Overestimated effect size
- Contributes to irreproducibility

Survey findings:

- Median power in neuroscience: 20-30%
- Most studies need 2-3× more samples

Solution: Always do prospective power analysis!

Pitfall 6: Multiple Testing

The problem:

- Test 20 outcomes at $\alpha = 0.05$
- Expect 1 false positive even if no real effects
- If you only report the “significant” ones...

Example:

- Measure 50 genes, 3 show $p < 0.05$
- Report these 3 as “significantly different”
- Problem: Expected 2.5 false positives!

Solutions:

1. Pre-specify primary outcome
2. Correct for multiple comparisons (Bonferroni, FDR)

Activity 5: Batch Effect Simulation

See batch effects in action

You will:

1. Generate data with batch effects
2. See how confounding with treatment ruins inference
3. Demonstrate benefits of balanced design
4. Show batch correction in analysis

Duration: 25 minutes

Part VI: Special Considerations in Biology

Animal Experiments: The 3Rs

Ethical framework:

1. **Replace:** Use alternatives to animals when possible
    - Cell culture, computer models, human studies
2. **Reduce:** Use minimum number of animals
    - But: Must have adequate power!
    - Efficient designs (blocking, within-subjects)
    - Share controls across experiments when appropriate
3. **Refine:** Minimize suffering
    - Humane endpoints
    - Analgesia, environmental enrichment

Good experimental design is an ethical imperative

Animal Experiments: Key Considerations

Cage effects:

- Animals in same cage are more similar
- Cage should often be experimental unit
- Or randomize across cages and block by cage

Litter effects:

- Siblings are genetically similar
- Use litter as blocking factor
- Don’t use only siblings for one treatment

Sex considerations:

- NIH now requires both sexes in studies
- Consider sex as a factor (factorial design)

Cell Culture Experiments

Common sources of variation:

1. **Plate effects:**
    - Different 96-well plates behave differently
    - Include all treatments on each plate (block by plate)
2. **Well position effects:**
    - Edge wells often different (evaporation)
    - Randomize treatments across well positions
    - Or exclude edge wells
3. **Passage number:**
    - Cells change with passages
    - Use similar passage numbers across experiment
    - Record and consider as covariate
4. **Confluence:**
    - Cell density affects behavior
    - Standardize seeding density

Technical vs Biological Replicates

Technical Replicate
Multiple measurements of the same biological sample

Biological Replicate
Independent biological samples/individuals

Examples:

- Technical: Run same RNA sample on 3 qPCR plates
- Biological: Extract RNA from 3 different mice

Why it matters:

- Technical replicates measure measurement error
- Biological replicates measure biological variation
- You need biological replicates for inference about populations!

Field Studies: Special Challenges

Limited control:

- Often can’t randomize (treatments already assigned)
- Observational rather than experimental
- Use matching, covariates, statistical controls

Spatial pseudo-replication:

- Nearby locations are more similar
- Can’t treat as independent
- Account for spatial correlation in analysis

Temporal considerations:

- Weather, seasonal effects
- Observer effects (learning over time)

Genomics Studies: Batch Effects Are HUGE

Sources of batch effects:

- Library preparation day/person
- Sequencing run/lane
- Time of sample processing
- Reagent lot
- Extraction method

These effects can be larger than biology!

Best practices:

1. Design: Balance samples across batches
2. Documentation: Record all batch information
3. Analysis: Include batch as covariate

Part VII: Practical Workflow

Before Starting Your Experiment

1. **Clear research question**
   - What exactly are you testing?
   - What is your hypothesis?
2. **Primary outcome defined**
   - What is the main thing you're measuring?
   - This prevents p-hacking
3. **Power analysis done**
   - How many samples do you need?
   - Is the study feasible?
4. **Randomization plan written down**
   - How will you assign treatments?
   - How will you control for confounders?
5. **Analysis plan pre-specified**
   - What statistical tests will you use?
   - Consider pre-registration

During Your Experiment

6. **Follow your randomization plan**
   - No deviations without good reason
   - Document any deviations
7. **Keep detailed records**
   - Lab notebook
   - Electronic data with metadata
   - Anything unusual that happens
8. **Blind when possible**
   - Blinding reduces bias
   - At minimum: Blind outcome assessment
9. **Handle missing data appropriately**
   - Document why data are missing
   - Don’t selectively exclude data

After Your Experiment

10. **Check for deviations from plan**
    - Did everything go as planned?
    - Document problems
11. **Report any excluded data**
    - How many samples/subjects excluded?
    - Why were they excluded?
    - Show this matches pre-specified criteria
12. **Follow pre-specified analysis**
    - Resist urge to try multiple tests
    - If you deviate, label as exploratory

Transparency is key to reproducible science!

Experimental Design Checklist

Before you collect any data, can you answer YES to all these?

1. Have you identified your experimental unit?
2. Have you done a power analysis?
3. Have you written down your randomization scheme?
4. Are treatments balanced across potential confounders?
5. Have you pre-specified your primary outcome?
6. Have you planned your statistical analysis?
7. Have you considered ethical issues (if using animals)?
8. Do you have appropriate controls?
9. Have you discussed design with a statistician?

If not, stop and fix your design first!

Activity 6: Sample Size Calculator

Interactive power/sample size tool

You will use an interactive tool for:

- t-tests
- ANOVA
- Regression

Input effect size, get required sample size

Visualize power curves

Duration: 20 minutes

Summary & Key Takeaways

Key Messages

1. **Design trumps analysis**
    - No statistics can fix a poorly designed study
2. **Randomization is crucial**
    - Controls for known AND unknown confounders
3. **Know your experimental unit**
    - Determines true sample size and df
4. **Power matters**
    - Underpowered studies waste resources
5. **Pre-specify your plan**
    - Prevents p-hacking and HARKing
6. **Consult early**
    - Talk to statistician BEFORE collecting data

Common Mistakes to Avoid

1. Pseudo-replication (wrong experimental unit)
2. Confounding (lack of randomization)
3. Batch effects (poor planning)
4. Selection bias (post-hoc exclusions)
5. Insufficient power (no power analysis)
6. Multiple testing (measuring everything, reporting some)
7. Starting without a plan

An ounce of prevention is worth a pound of cure!

Connection to Course Structure

Building on previous days:

- Day 1 (Philosophy): Pre-registration prevents p-hacking
- Day 2 (Probability): Power calculations use probability distributions
- Day 3 (Inference): Proper design ensures valid p-values and CIs
- Day 4 (Linear models): Design determines model structure

Today's foundation:

- Design principles inform all downstream analysis
- Good design makes analysis straightforward
- Poor design makes analysis impossible

Resources for Further Learning

Books:

- Ruxton & Colegrave: *Experimental Design for the Life Sciences*
- Quinn & Keough: *Experimental Design and Data Analysis for Biologists*
- Montgomery: *Design and Analysis of Experiments*

Online resources:

- NC3Rs Experimental Design Assistant (for animal studies)
- G*Power software (power analysis)
- ARRIVE guidelines (reporting animal research)

Statistical consultation:

- Your institution's statistics department
- Before you collect data!

Wrap-up Discussion

Questions? Concerns? Insights?

Discussion topics:

- What surprised you today?
- What will you do differently in your next experiment?
- What challenges do you face in implementing these principles?

Duration: 20 minutes

Thank you!

Design well, analyze confidently