Stat 5200, Module 3: Cost Optimization in Experimental Design

Jesse Wheeler

Overview

Objectives and overview

- This supplement covers optimal resource allocation in nested experimental designs.
- Module outcomes include:
  - Understanding how variance components affect the precision of treatment means.
  - Formulating a cost function for two-stage/nested designs.
  - Deriving the optimal sample size (Measurement Units per Experimental Unit) to maximize statistical power on a fixed budget.
  - Combining optimal allocation with power analysis to determine total sample size.

Variance in Nested Designs

Deriving the Variance of a Treatment Mean

- To understand where the variance formula comes from, we start with the linear statistical model for a nested design (subsampling):

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$$

- Where:
  - μ is the overall grand mean (fixed).
  - α_i is the effect of the i-th treatment (fixed).
  - $\beta_{ij} \sim N(0, \sigma^2_{EU})$ is the random error for the j-th Experimental Unit (EU).
  - $\epsilon_{ijk} \sim N(0, \sigma^2_{MU})$ is the random error for the k-th Measurement Unit (MU) nested within the EU.

Deriving the Variance of a Treatment Mean II

- The treatment mean ($\bar{y}_{i..}$) is the average of all observations in group i, which has b EUs and n MUs per EU:

$$\bar{y}_{i..} = \frac{1}{bn} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$$

- Substituting the model into this average yields:

$$\bar{y}_{i..} = \mu + \alpha_i + \left(\frac{1}{b} \sum_{j=1}^{b} \beta_{ij} \right) + \left(\frac{1}{bn} \sum_{j=1}^{b} \sum_{k=1}^{n} \epsilon_{ijk} \right)$$

Deriving the Variance of a Treatment Mean III

- Next, we apply the variance operator, $Var(\bar{y}_{i..})$.
- Because μ and α_i are constants, their variance is 0.
- Because the errors (β_{ij} and ϵ_{ijk}) are independent, the variance of their sum is the sum of their variances:

$$Var(\bar{y}_{i..}) = Var\left(\frac{1}{b}\sum_{j=1}^{b}\beta_{ij}\right) + Var\left(\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\epsilon_{ijk}\right)$$

Deriving the Variance of a Treatment Mean IV

- Using the fundamental property of variance $Var(cX) = c^2 Var(X)$:
- **EU Component:**

$$Var\left(\frac{1}{b}\sum_{j=1}^{b}\beta_{ij}\right) = \frac{1}{b^2}\sum_{j=1}^{b}Var(\beta_{ij}) = \frac{1}{b^2}(b \cdot \sigma^2_{EU}) = \frac{\sigma^2_{EU}}{b}$$

- **MU Component:**

$$Var\left(\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\epsilon_{ijk}\right) = \frac{1}{(bn)^2}\sum_{j=1}^{b}\sum_{k=1}^{n}Var(\epsilon_{ijk}) = \frac{1}{(bn)^2}(bn \cdot \sigma^2_{MU}$$

Deriving the Variance of a Treatment Mean V

- Combining these components provides the final formula for the variance of the treatment mean:

$$Var(\bar{y}_{i..}) = \frac{\sigma^2_{EU}}{b} + \frac{\sigma^2_{MU}}{b \cdot n}$$

Precision of Treatment Means

- In a nested design, our goal is often to maximize the precision (minimize the variance) of the estimated treatment effects.
- Using the formula we just derived, let V be the variance of the treatment mean:

$$V = \frac{\sigma^2_{EU}}{b} + \frac{\sigma^2_{MU}}{b \cdot n}$$

Precision of Treatment Means II

- *EX:* If EUs are pens and MUs are pigs, this is:

$$V = \frac{\sigma^2_{pen}}{b} + \frac{\sigma^2_{pig}}{b \cdot n}$$

- If our only constraint is the total number of MUs ($K = b \times n$), then rewriting the variance gives:

$$V = \frac{n \cdot \sigma^2_{pen}}{K} + \frac{\sigma^2_{pig}}{K}$$

- Because K and the variance components are fixed, the only way to minimize V is to minimize n.
- Thus, if EUs are "free", the optimal design is $n = 1$ MU per EU (which reduces the nested design to a Completely Randomized Design).

The Cost Function

Budget Constraints

- In reality, Experimental Units (like pens, hospital rooms, or field plots) are rarely free.
- We must optimize our study based on a fixed financial budget (C).
- Let C_{EU} be the fixed cost to run one EU (e.g., rent and maintenance for a pen).
- Let C_{MU} be the marginal cost to add one MU to an existing EU (e.g., purchase and feed for a pig).
- The total budget for one treatment group is the sum of these costs:

$$C = b \cdot C_{EU} + (b \cdot n) \cdot C_{MU}$$

Optimum Allocation

Minimizing Variance for a Fixed Cost

- We want to find the optimal number of MUs per EU (n) that minimizes variance V, subject to the cost constraint C.
- First, isolate b (the number of EUs) in the cost equation:

$$b = \frac{C}{C_{EU} + n \cdot C_{MU}}$$

- Substitute this into the variance equation:

$$V = \left(\sigma_{EU}^2 + \frac{\sigma_{MU}^2}{n}\right) \cdot \frac{C_{EU} + n \cdot C_{MU}}{C}$$

Minimizing Variance for a Fixed Cost II

- Expanding the terms yields:

$$V = \frac{1}{C}\left[\sigma_{EU}^2 C_{EU} + n\sigma_{EU}^2 C_{MU} + \frac{\sigma_{MU}^2 C_{EU}}{n} + \sigma_{MU}^2 C_{MU}\right]$$

- To minimize variance, we take the derivative of V with respect to n and set it to zero. The constant terms drop out:

$$\frac{dV}{dn} = \frac{1}{C}\left[\sigma_{EU}^2 C_{MU} - \frac{\sigma_{MU}^2 C_{EU}}{n^2}\right] = 0$$

Minimizing Variance for a Fixed Cost III

- Solving for n (number of MUs per EU) gives the optimum allocation formula, based on cost:

$$n = \sqrt{\frac{\sigma_{MU}^2 \cdot C_{EU}}{\sigma_{EU}^2 \cdot C_{MU}}} \tag{1}$$

Practical Considerations: Estimating Variance

- A common question is: *How do we know σ^2_{EU} and σ^2_{MU} before the experiment is run?*
- To use the optimum allocation formula, you need *a priori* estimates of these variance components. Common sources include:
  - **Pilot Studies:** Running a small-scale nested design specifically to estimate variance components via REML (e.g., using `lme4`).
  - **Historical Data:** Re-analyzing data from previous experiments conducted in the same lab, field, or facility.
  - **Literature Review:** Searching for published papers that report variance components or mean squares for similar nested structures.

Combining with Power Analysis

From Allocation to Total Sample Size

- Optimum allocation gives us the ideal **shape** of our design (the ratio n), but it does not tell us the **scale** (the total number of Experimental Units, b) needed to actually detect an effect.
- To find b, we combine the optimal n with a traditional power analysis.
- For comparing two treatments, the approximate number of EUs required *per treatment group* is:

$$b \approx \frac{2(Z_{1-\alpha/2} + Z_{power})^2}{\Delta^2} \left(\sigma^2_{EU} + \frac{\sigma^2_{MU}}{n} \right)$$

- Where:

From Allocation to Total Sample Size II

- Δ is the minimum detectable difference between the two treatment means.
- $Z_{1-\alpha/2}$ and Z_{power} are the standard normal quantiles for the chosen significance level (e.g., 0.05) and desired statistical power (e.g., 80%).

Applied Example in R

Step 1: Find Optimal Allocation

- Suppose we fit a mixed model on previous pig data and extracted the variance components:
- $\sigma^2_{pen} = 214.5$, $\sigma^2_{pig} = 796.3$.
- A single pen costs \$500 to rent, and a pig costs \$100 to purchase and feed.

Step 1: Find Optimal Allocation II

```r
# Variance components
var_pen <- 214.5
var_pig <- 796.3

# Cost parameters
cost_pen <- 500
cost_pig <- 100

# Calculate optimal MU per EU
n_opt <- sqrt((var_pig * cost_pen) / (var_pen * cost_pig))
n_opt <- round(n_opt)
n_opt # Optimal pigs per pen
```

```
[1] 4
```

Step 2: Power Analysis for Total Units

- We know we need 4 pigs per pen. Now, how many pens do we need to detect a weight difference of $\Delta = 25$ lbs with 80% power at $\alpha = 0.05$?

Step 2: Power Analysis for Total Units II

```r
# Power analysis parameters
delta <- 25
alpha <- 0.05
power <- 0.80

# Z-scores
z_alpha <- qnorm(1 - alpha/2)
z_power <- qnorm(power)

# Calculate b (pens per treatment)
var_effective <- var_pen + (var_pig / n_opt)
b <- 2 * (z_alpha + z_power)^2 * var_effective / delta^2

ceiling(b) # Round up to whole pens
```

```
[1] 11
```

Further Reading

Recommended Resources

- For students interested in exploring cost optimization and sample size allocation further, the following texts are highly recommended:

- **Design of Experiments:**
  - *Design of Experiments: Statistical Principles of Research Design and Analysis* by Robert O. Kuehl. (Excellent coverage of variance components and nested designs in biological/agricultural contexts).
  - *Design and Analysis of Experiments* by Douglas C. Montgomery. (A standard text covering random effects, mixed models, and optimal resource allocation).

Recommended Resources II

- *A First Course in Design and Analysis of Experiments* by Gary W. Oehlert. (Provides highly accessible breakdowns of expected mean squares and optimal sizes).

Recommended Resources III

- **Sampling Theory (The origin of the formula):**
  - *Sampling Techniques* by William G. Cochran. (Chapter 10 on "Subsampling with Units of Equal Size" provides the foundational mathematical derivation of this cost optimization formula, originally developed for survey sampling).
  - *Sampling: Design and Analysis* by Sharon L. Lohr. (A modern, accessible take on two-stage cluster sampling and optimal cost allocation).

References and Acknowledgements

- Compiled on November 20, 2026 using R version 4.6.2.
- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.
- We acknowledge students and instructors for previous versions of this course / slides.