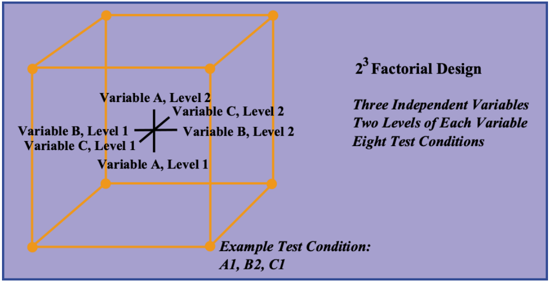
Randomized Controlled Trials (RCTs): Multi-Arm and Factorial Designs MATH 348, Vassar College, Spring 2024



https://en.wikipedia.org/wiki/Factorial_experiment

Prof. Lee Kennedy-Shaffer
March 18, 2024

Outline

1. Setting
2. Multi-Arm RCTs
3. Factorial RCTs
   - Background and Example Design
   - Analysis Example
   - Considerations for Factorial RCTs
4. Reading Considerations

Example: Selecting *Netflix* Artwork

See Krishnan, Gopal. "Selecting the best artwork for videos through A/B testing." 3 May 2016 *Netflix Technology Blog*. Accessed 13 March 2023, https://netflixtechblog.com/selecting-the-best-artwork-for-videos-through-a-b-testing-f61

Question of Interest

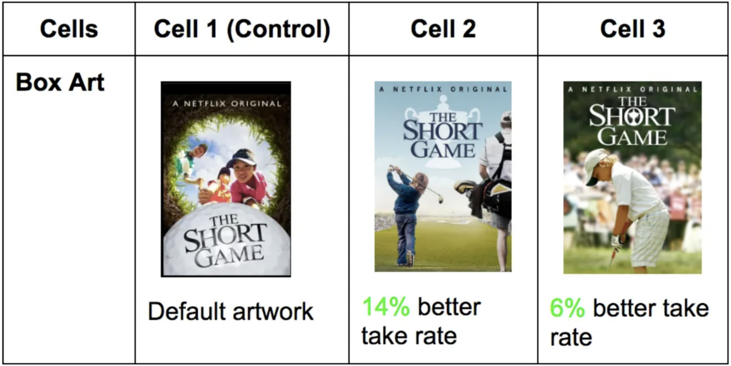
Which artwork design will lead to the highest number of user clicks?

Sidebar: Something to Consider

How does proprietary research for a company/organization differ from scientific research? How are they similar?

Multiple Interventions

In some cases, we're interested in testing **multiple interventions** (i.e., new artwork designs) against a control condition (i.e., the current/default artwork).



Cells	Cell 1 (Control)	Cell 2	Cell 3
Box Art	Default artwork	14% better take rate	6% better take rate

A NETFLIX ORIGINAL
THE SHORT GAME
A NETFLIX ORIGINAL
THE SHORT GAME
A NETFLIX ORIGINAL
THE SHORT GAME

Outline

1. Setting
2. Multi-Arm RCTs
3. Factorial RCTs
   - Background and Example Design
   - Analysis Example
   - Considerations for Factorial RCTs
4. Reading Considerations

Options

- Conduct two (or three) trials comparing the different options
- Conduct one trial that randomizes users into all three arms

Multi-Arm Trial

Definition and Estimands

A **multi-arm trial** is run the same as any other RCT, but participants are randomized into more than two treatment arms. Comparisons of all the interventions can then be made against the control (or against each other).

Multi-Arm Trial

Definition and Estimands

A **multi-arm trial** is run the same as any other RCT, but participants are randomized into more than two treatment arms. Comparisons of all the interventions can then be made against the control (or against each other).

Advantages:

Risk: Multiple Testing

We are using the same control group for multiple comparisons. That means we increase our chance of finding a significant p-value (**multiple testing problem**). Plus the p-values/CIs are correlated but not perfectly correlated.

Risk: Multiple Testing

We are using the same control group for multiple comparisons. That means we increase our chance of finding a significant p-value (**multiple testing problem**). Plus the p-values/CIs are correlated but not perfectly correlated.

Question

If both new artworks have the same true effect, but we know we failed to reject the null for the first new design, does that give us any information for the test of the second design?

Solution: Corrected Tests

Bonferroni Correction

Divide the significance level α by the number of tests we're conducting and use that as our new significance level.

Solution: Corrected Tests

Bonferroni Correction

Divide the significance level α by the number of tests we're conducting and use that as our new significance level.

Properties:

- Preserves TIE Rate
- *Conservative correction*

Power and Sample Size Comparison

Example:

- Overall $\alpha = 0.05$
- Desired power 80% to detect 10% increase in click rate
- Control click rate of 5%

Power and Sample Size Comparison

Example:

- Overall $\alpha = 0.05$
- Desired power 80% to detect 10% increase in click rate
- Control click rate of 5%

- Two trials of two arms each: $n = 141$ per arm
- Three arms (uncorrected): $n = 141$ per arm
- Three arms (Bonferroni): $n = 170$ per arm

Power and Sample Size Comparison

Example:

- Overall $\alpha = 0.05$
- Desired power 80% to detect 10% increase in click rate
- Control click rate of 5%

- Two trials of two arms each: $n = 141$ per arm; total 564
- Three arms (uncorrected): $n = 141$ per arm; total 423
- Three arms (Bonferroni): $n = 170$ per arm; total 510

Unbalanced Designs

Since we're using the control group twice, it can be useful to increase the size of the control group relative to the other two. Can optimize to find the best ratio.

Outline

1. Setting

2. Multi-Arm RCTs

3. Factorial RCTs
   - Background and Example Design
   - Analysis Example
   - Considerations for Factorial RCTs

4. Reading Considerations

Many Factors

In some cases, we may be testing multiple types of interventions at once. Each type (known as a factor) may have two or more different levels.

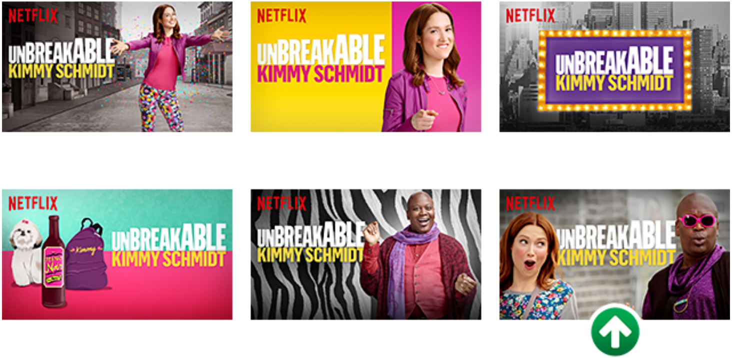
Many Factors

In some cases, we may be testing multiple types of interventions at once. Each type (known as a factor) may have two or more different levels.

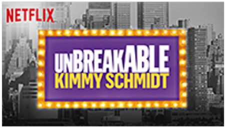
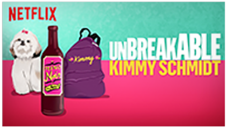
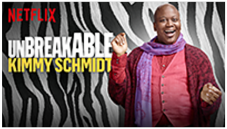
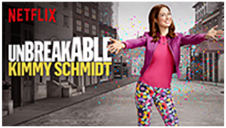
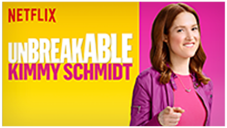
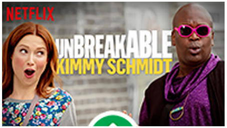
Factors

If we have two factors with 3 and 2 levels, respectively, how many treatment arms do we need to test all possible combinations?

Example



Arranging by (Possible) Factors

		X_2: Features	
X_1: Ellie Kemper?	0: BW City	1: Color	2: Titus Burgess
0: No			
1: Yes			

Design Choices

- Complete vs. Incomplete
- Balanced vs. Unbalanced

Estimands

- Main Effects
- Interaction Effects
- Higher-Order Interaction Effects

Hypothetical Results

Proportion of viewers who click on the title:

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$
Kemper No	0.05	0.07	0.06
Kemper Yes	0.06	0.08	0.09

Hypothetical Results

Proportion of viewers who click on the title:

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$	Mean
Kemper No	0.05	0.07	0.06	
Kemper Yes	0.06	0.08	0.09	
Mean				
Response to X_1				

Hypothetical Results

Proportion of viewers who click on the title:

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$	Mean
Kemper No	0.05	0.07	0.06	0.06
Kemper Yes	0.06	0.08	0.09	0.077
Mean	0.055	0.075	0.075	0.068
Response to X_1	$+0.01$	$+0.01$	$+0.03$	$+0.017$

Questions

- What are the main effects?
- What interaction effects exist?

Regression Analysis

Regression Model

$$
\begin{aligned}
E[Y|X_1, X_2] = \beta_0 &+ \beta_1 I(X_1 = 1) + \beta_2 I(X_2 = 1) + \beta_3 I(X_2 = 2) \\
&+ \beta_4 I(X_1 = 1) I(X_2 = 1) + \beta_5 I(X_1 = 1) I(X_2 = 2)
\end{aligned}
$$

More Complicated Factorial Designs

You can add more factors or more levels of each factor. The number of cells grows very quickly, however.

More Complicated Factorial Designs

You can add more factors or more levels of each factor. The number of cells grows very quickly, however.

One Solution

Higher-order interactions are often assumed to be negligible (or ignored by design). Then different levels of a second factor can still count as replication for the first factor, increasing the **effective sample size** for testing the levels of the first factor.

Specifying Primary Analysis and Power Calculation

Often, a specific comparison is chosen to be the **primary analysis**, and others are treated as **secondary analyses**. Depending on the scientific question of interest, this could be one specific cell to another, or focusing on one factor.

Advantages, Disadvantages, Tradeoffs

Outline

1. Setting
2. Multi-Arm RCTs
3. Factorial RCTs
   - Background and Example Design
   - Analysis Example
   - Considerations for Factorial RCTs
4. Reading Considerations

Consider While Reading

- What are the advantages of a multi-arm/factorial RCT in this setting?
- What are the disadvantages/trade-offs?
- What is the primary estimand of interest?
- Are interactions considered?
- Is there a discussion of the sample size/power calculation? What estimand is the focus of that calculation and is multiple testing addressed?
- How do the investigators describe the results?