A/B Testing and Beyond

Designed Experiments for Data Scientists

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Week 7

Wednesday October 18th, 2017

Outline

- $2^{k\text{-}p}$ Fractional Factorial Experiments
  - Recap
  - The Chehalem Example
- Multi-armed Bandit Experiments
  - The multi-armed bandit problem
  - A few solutions to the problem

RECAP

Recap

- Analyzing Factorial Experiments
  - Continuous Responses
  - Binary Responses
- Two-Level Factorial Experiments
  - 2^k Factorial Experiments
  - 2^{k-p} Fractional Factorial Experiments

TWO-LEVEL FACTORIAL EXP'S

Two-Level Factorial Experiments

- When investigating k factors, two-level factorial experiments are the smallest possible factorial experiments

- Such experiments are typically used for factor screening

- Pareto Principle: only *a vital few* factors are important relative to the *trivial many*

- The purpose of a screening experiment is to identify this small number of influential factors

TWO-LEVEL FACTORIAL EXP'S

Two-Level Factorial Experiments

Here we discuss two particular types of two-level factorial experiments for investigating k factors:

- 2^k factorial experiments
  - These investigate each of the unique 2^k conditions
- 2^{k-p} fractional factorial experiments
  - These investigate just a *fraction* of the unique 2^k conditions

2^k FACTORIAL EXPERIMENTS

Designing 2^k Factorial Experiments

Step 1: Choose k factors that are expected to influence the response in some way

Step 2: Choose two levels for each factor to experiment with

- It's important to choose levels that provide the largest opportunity for an influential factor to be noticed
- Levels should be chosen that are quite different from one another; even a very influential factor may not appear to be influential if the factor levels are too similar.

2^k FACTORIAL EXPERIMENTS

Designing 2^k Factorial Experiments

Step 3: The experimental conditions are defined to be the unique combination of these factors' levels

- There will be 2^k of them

Step 4: Assign experimental units to each of the 2^k conditions

- For ease of notation, we assume that the experiment is balanced and n units are assigned to each condition
- The sample size n can be determined by power analyses based on two-sample tests that account for the multiple comparison problem

2^k FACTORIAL EXPERIMENTS

Designing 2^k Factorial Experiments

- Using the ±1 coding, each experimental condition can be identified by a unique combination of plus and minus ones
- The design of the experiment can be displayed in what is known as a design matrix
- Using the data collected from such a study we fit a linear or logistic regression model
- Let’s review some design matrices

2^k FACTORIAL EXPERIMENTS

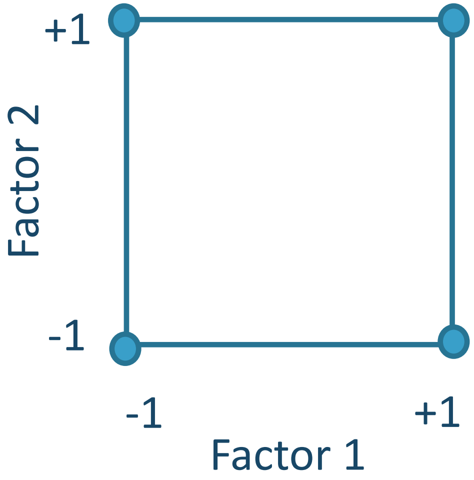
A 2^1 Factorial Experiment (i.e., an A/B test)

Condition	Factor 1
1	-1
2	+1

2^k FACTORIAL EXPERIMENTS

A 2^2 Factorial Experiment

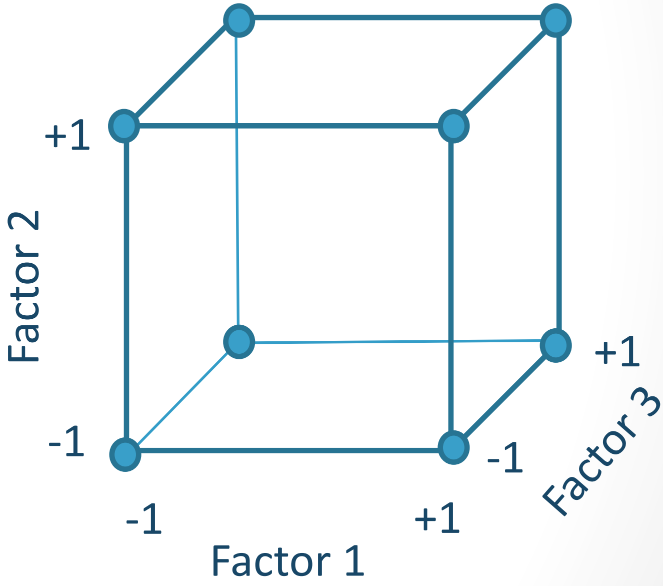
Condition	Factor 1	Factor 2
1	-1	-1
2	+1	-1
3	-1	+1
4	+1	+1



2^k FACTORIAL EXPERIMENTS

A 2^3 Factorial Experiment

Condition	Factor 1	Factor 2	Factor 3
1	-1	-1	-1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	-1
5	-1	-1	+1
6	+1	-1	+1
7	-1	+1	+1
8	+1	+1	+1



2^k FACTORIAL EXPERIMENTS

Analyzing 2^k Factorial Experiments

- Using the data collected from such a study we fit a linear or logistic regression model

- Apart from this difference, the models are similar in that

  - they are based on exactly the same linear predictor

  - we can evaluate the significance of main and interaction effects by performing tests concerning individual or multiple β's (although the specific tests that are used differ in the two settings)

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- 2^k factorial experiments are a useful special case of a general factorial experiment
  - They minimize the number of levels being investigated, and hence reduces the overall number of experimental conditions
- BUT they still investigate **all possible** combinations of the factor levels – which can be a lot!
  - With $k = 8$ factors the 2^k factorial experiment has 256 conditions

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- Alternatively we could use a 2^{k-p} fractional factorial experiment which also investigates k factors, but with just a fraction of the conditions
- Rather than performing 2^k conditions, we perform 2^{k-p} specially selected conditions which still allow us to estimate main effects and potentially important interaction effects
- **With these experiments we can investigate a relatively large number of factors with a relatively small number of conditions**
- However, we sacrifice the ability to separately estimate *all* main and interaction effects

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

Motivation: the linear predictor for a 2^k factorial experiment consists of

- $\binom{k}{1} = k$ main effect terms
- $\binom{k}{2}$ two-factor interaction terms
- $\vdots$
- $\binom{k}{k} = 1$ k-factor interaction term

That's a total of $\sum_{i=1}^{k} \binom{k}{i} = 2^k - 1$ terms

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- Of these $2^k - 1$ terms, only $k + \binom{k}{2}$ of them are main effects and two factor interactions – the remaining correspond to higher order interaction terms

- If $k = 8$, there are 8 main effects, 28 two-factor interactions and 219 higher order interactions, many of which are likely to be insignificant

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- Principle of effect sparsity: in the presence of several factors, variation in the response is likely to be driven by a **small number of main effects and low-order interactions**

- Thus, it is typically a waste of resources to estimate these higher order interaction terms

- It is a better use of these resources is to estimate the main effects and low-order interactions of a larger number of factors

- So how do we do this?

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

First let's discuss p:

- Investigating k factors in a full factorial experiments takes 2^k conditions
- If we'd like to investigate k factors in half as many conditions, we use a 2^{k-1} experiment
- If we'd like to investigate k factors with just a quarter of the conditions, we use a 2^{k-2} experiment
- In general, if we'd like to investigate k factors in $(1/2)^p$ as many conditions, we use a 2^{k-p} experiment

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- If a full factorial approach requires 2^k conditions and we only want 2^{k-p}, we need to choose **which** 2^{k-p} conditions to experiment with

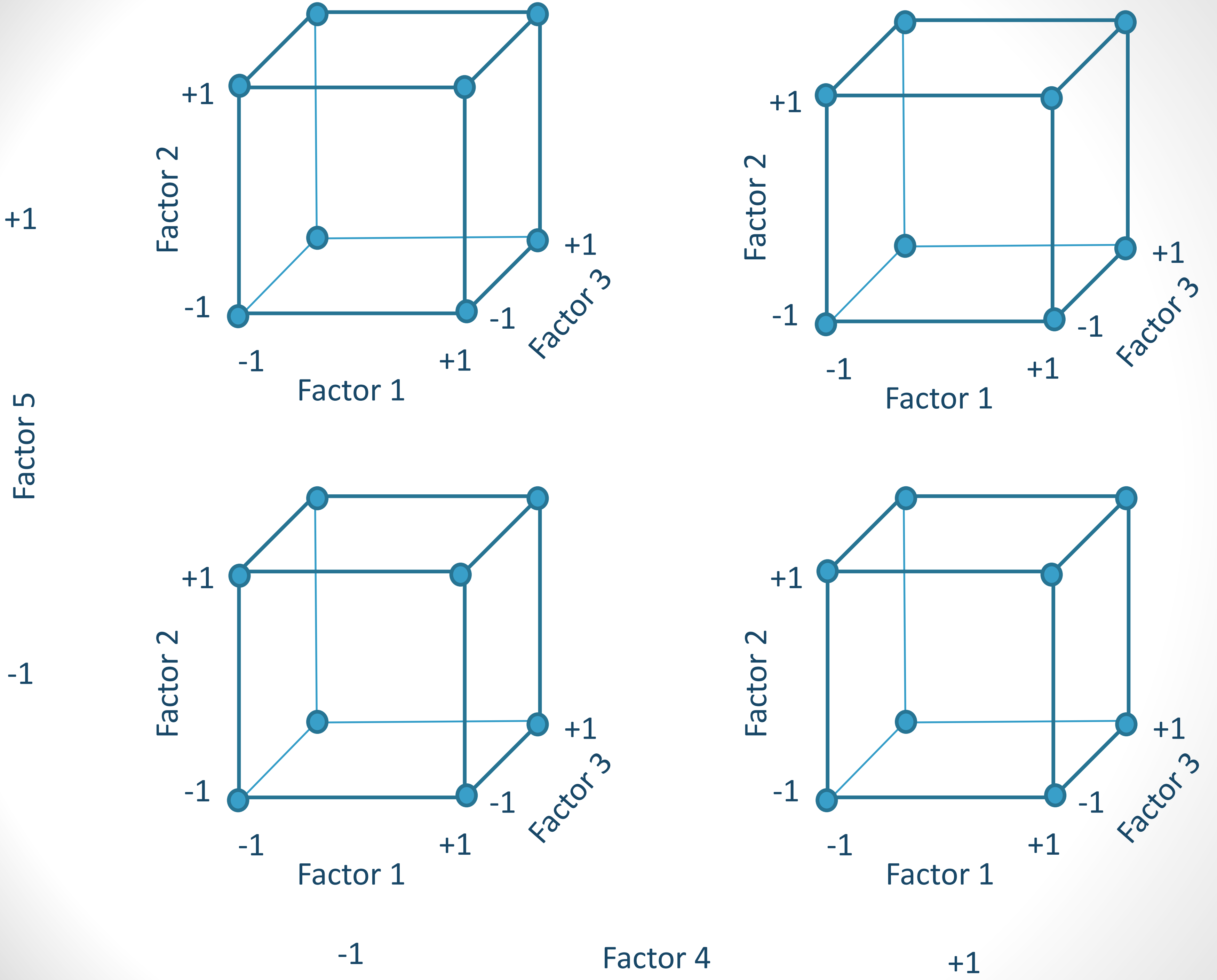
- For instance, if $k = 5$ and $p = 2$, then the goal is to investigate 5 factors in 2^3 = 8 conditions (where normally 32 conditions would be required with the full factorial approach).

- The question is, among these 32 conditions, which 8 do we choose to for the $2^{5\text{-}2}$ fractional design?

Condition	A	B	C	D	E
1	-1	-1	-1	-1	-1
2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	-1
4	+1	+1	-1	-1	-1
5	-1	-1	+1	-1	-1
6	+1	-1	+1	-1	-1
7	-1	+1	+1	-1	-1
8	+1	+1	+1	-1	-1
9	-1	-1	-1	+1	-1
10	+1	-1	-1	+1	-1
11	-1	+1	-1	+1	-1
12	+1	+1	-1	+1	-1
13	-1	-1	+1	+1	-1
14	+1	-1	+1	+1	-1
15	-1	+1	+1	+1	-1
16	+1	+1	+1	+1	-1

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Condition	A	B	C	D	E
17	-1	-1	-1	-1	+1
18	+1	-1	-1	-1	+1
19	-1	+1	-1	-1	+1
20	+1	+1	-1	-1	+1
21	-1	-1	+1	-1	+1
22	+1	-1	+1	-1	+1
23	-1	+1	+1	-1	+1
24	+1	+1	+1	-1	+1
25	-1	-1	-1	+1	+1
26	+1	-1	-1	+1	+1
27	-1	+1	-1	+1	+1
28	+1	+1	-1	+1	+1
29	-1	-1	+1	+1	+1
30	+1	-1	+1	+1	+1
31	-1	+1	+1	+1	+1
32	+1	+1	+1	+1	+1


+1
Factor 5
-1
+1
Factor 2
-1
-1
Factor 1
+1
-1
Factor 3
+1
+1
Factor 2
-1
-1
Factor 1
+1
-1
Factor 3
+1
+1
Factor 2
-1
-1
Factor 1
+1
-1
Factor 3
+1
+1
Factor 2
-1
-1
Factor 1
+1
-1
Factor 3
+1
-1
Factor 4
+1

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- To answer this, we consider an extended version of the design matrix associated with a full 2^3 factorial experiment

Condition	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	+1	+1	+1	-1
2	+1	-1	-1	-1	-1	+1	+1
3	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	-1	+1	-1	-1	-1
5	-1	-1	+1	+1	-1	-1	+1
6	+1	-1	+1	-1	+1	-1	-1
7	-1	+1	+1	-1	-1	+1	-1
8	+1	+1	+1	+1	+1	+1	+1

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- When it comes to fitting a regression model, each column in this matrix is used to estimate the corresponding effect a particular effect
- For instance:
  - the AB column is used to estimate β_{AB}, the interaction effect between factors A and B
  - the ABC column is used to estimate β_{ABC}, the interaction effect between factors A, B and C
- Now recall the effect sparsity principle: if an interaction is likely to be negligible, why not use its column to dictate the levels of an extra factor?

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

For example:

- Let's use the ± 1's in the ABC column as a prescription for when to run D at its low and high levels
- Let's use the ± 1's in the BC column as a prescription for when to run E at its low and high levels

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

Condition	A	B	C	AB	AC	E=BC	D=ABC
1	-1	-1	-1	+1	+1	+1	-1
2	+1	-1	-1	-1	-1	+1	+1
3	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	-1	+1	-1	-1	-1
5	-1	-1	+1	+1	-1	-1	+1
6	+1	-1	+1	-1	+1	-1	-1
7	-1	+1	+1	-1	-1	+1	-1
8	+1	+1	+1	+1	+1	+1	+1

- Here we say that D and ABC are aliased and E and BC are aliased
- ‘D=ABC’ and ‘E=BC’ are called the design generators

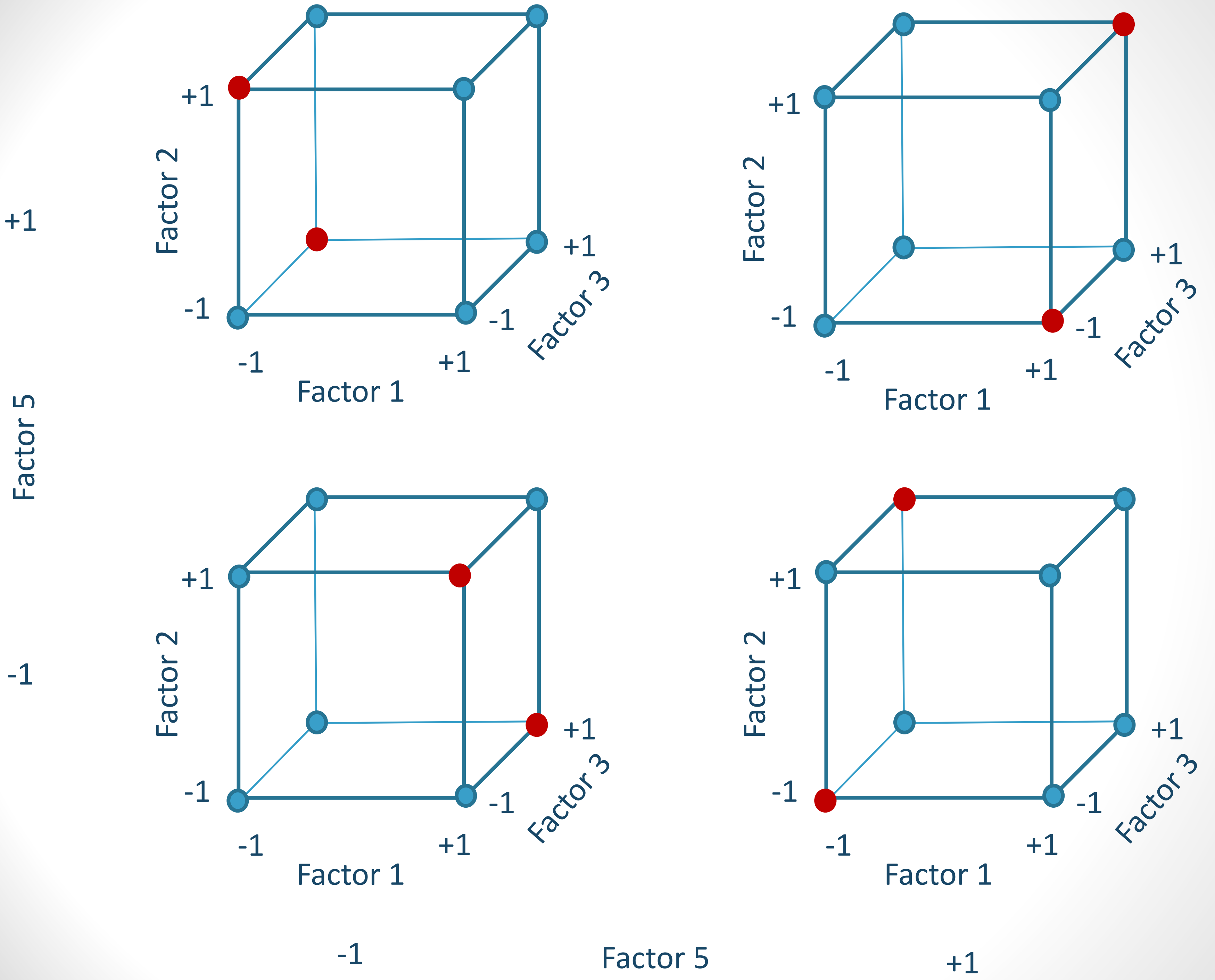
2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

The design matrix indicating the levels of factors A, B, C, D, and E in each condition is shown below

Condition	A	B	C	E	D
1	-1	-1	-1	+1	-1
2	+1	-1	-1	+1	+1
3	-1	+1	-1	-1	+1
4	+1	+1	-1	-1	-1
5	-1	-1	+1	-1	+1
6	+1	-1	+1	-1	-1
7	-1	+1	+1	+1	-1
8	+1	+1	+1	+1	+1

These are a subset of the 32 conditions required in a full 2^5 factorial experiment


+1
Factor 5
-1
+1
Factor 2
-1
-1
+1
Factor 1
+1
-1
Factor 3
+1
Factor 2
-1
-1
+1
Factor 1
+1
-1
Factor 3
+1
Factor 2
-1
-1
+1
Factor 1
+1
-1
Factor 3
+1
Factor 2
-1
-1
+1
Factor 1
+1
-1
Factor 3
-1
Factor 5
+1

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- When two terms are aliased, their effects become confounded

- For instance the D=ABC column estimates the ABC interaction **and** the main effect of D

- So the coefficient β_{ABC} quantifies the joint effects of the ABC interaction and the main effect of factor D

- Thus, we cannot separately estimate the main effect of D from the ABC interaction effect

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- Thus confounding results from aliasing a new main effect with an existing interaction
- As such, it is important to think carefully about **which** interaction to choose as an alias
- It is best to avoid aliasing a new factor with an interaction that is likely to be significant (since separately estimating significant effects is desirable)
- So high order interaction terms (that are unlikely to be significant) are good choices for aliases

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- This notion is quantified by the resolution of the fractional factorial design

- A design is of resolution R if main effects are aliased with interaction effects involving at least $R - 1$ factors

- In the design we've been discussing main effects are aliased with two- and three-factor interactions

- Thus it is a resolution III experiment denoted by

$$2_{III}^{5-2}$$

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- In general, higher resolution designs are to be preferred over lower resolution designs.
- For instance, resolution IV and V designs are to be preferred over a resolution III designs
- In these cases main effects will not be confounded with two-factor interactions
- Since two-factor interactions are typically important, it is best if their effects are not confounded with main effects

2^{k-p} FRACTIONAL FACTORIAL

Designing 2^{k-p} Fractional Factorial Experiments

- The resolution of a fractional factorial experiment is determined by two things:
  1. The degree of fractionation desired (i.e., the size of p relative to k)
  2. The design generators chosen for aliasing
- The degree of fractionation is typically determined by resource constraints – how many conditions can you manage?
- Given the degree of fractionation (p) we typically choose design generators to maximize resolution

2^{k-p} FRACTIONAL FACTORIAL

Analyzing 2^{k-p} Fractional Factorial Experiments

- The analysis of these fractional factorial experiments is based on regression models
  - Linear regression (if Y is continuous)
  - Logistic regression (if Y is binary)
- In fact, the analysis is not very different from what we saw in the credit card example
  - We perform individual and simultaneous hypothesis tests to compare full and reduced models
  - This allows us to evaluate the significance of various main and interaction effects

2^{k-p} FRACTIONAL FACTORIAL

Analyzing 2^{k-p} Fractional Factorial Experiments

The wrinkle:

- Here the effects estimated in these models are confounded with other effects
- So we can't be 100% certain that a given effect is due to say a main effect, or perhaps the interaction it is aliased with
- But, if the resolution is high, we hope that important effects are aliased with high-order interactions (that are likely negligible)
- This provides confidence that significant effects are not due to the high-order interactions

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

Chehalem is a winery in Newberg Oregon that regularly uses experiments to develop and refine wine recipes. Montgomery (2017) discusses a 2_{IV}^{8-4} fractional factorial experiment that was used to investigate $k = 8$ factors with just 16 conditions.

The goal of the experiment was to evaluate and quantify the influence of several factors on the quality of the wine. The response variable here is a tasting score provided subjectively by $n = 5$ taste-testers.

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

Factor	Low (-)	High (+)
Pinot Noir clone (A)	Pommard	Wadenswil
Oak type (B)	Allier	Troncais
Age of barrel (C)	Old	New
Yeast/Skin contact (D)	Champagne	Montrachet
Stems (E)	None	All
Barrel toast (F)	Light	Medium
Whole cluster (G)	None	10%
Fermentation Temperature (H)	Low (75°F max)	High (92°F max)

Condition	A	B	C	D	E	F	G	H	Avg. Rating
1	-1	-1	-1	-1	-1	-1	-1	-1	9.6
2	+1	-1	-1	-1	-1	+1	+1	+1	10.8
3	-1	+1	-1	-1	+1	-1	+1	+1	12.6
4	+1	+1	-1	-1	+1	+1	-1	-1	9.2
5	-1	-1	+1	-1	+1	+1	+1	-1	9.0
6	+1	-1	+1	-1	+1	-1	-1	+1	15.0
7	-1	+1	+1	-1	-1	+1	-1	+1	5.0
8	+1	+1	+1	-1	-1	-1	+1	-1	15.2
9	-1	-1	-1	+1	+1	+1	-1	+1	2.2
10	+1	-1	-1	+1	+1	-1	+1	-1	7.0
11	-1	+1	-1	+1	-1	+1	+1	-1	8.8
12	+1	+1	-1	+1	-1	-1	-1	+1	2.8
13	-1	-1	+1	+1	-1	-1	+1	+1	4.6
14	+1	-1	+1	+1	-1	+1	-1	-1	2.4
15	-1	+1	+1	+1	+1	-1	-1	-1	9.2
16	+1	+1	+1	+1	+1	+1	+1	+1	12.6

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- Because the response variable is continuous we use linear regression for this analysis
- Because only 2^4=16 conditions were used, we can only fit a model with 16 regression coefficients
- In the context of a full 2^4 factorial experiment this corresponds to a model with
  - 4 main effects
  - 6 two-factor interactions
  - 4 three-factor interactions
  - 1 four-factor interaction

```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 8.5000   0.2658     31.985   < 2e-16 ***
A           0.8750   0.2658      3.293   0.001619 **
B           0.9250   0.2658      3.481   0.000906 ***
C           0.6250   0.2658      2.352   0.021772 *
D          -2.3000   0.2658     -8.655   2.27e-12 ***
A:B        -0.3500   0.2658     -1.317   0.192532
A:C         1.3000   0.2658      4.892   7.07e-06 ***
B:C         0.4500   0.2658      1.693   0.095261 .
A:D        -0.8750   0.2658     -3.293   0.001619 **
B:D         1.2250   0.2658      4.610   1.98e-05 ***
C:D         0.3750   0.2658      1.411   0.163063
A:B:C       1.5750   0.2658      5.927   1.35e-07 ***
A:B:D      -0.3000   0.2658     -1.129   0.263168
A:C:D      -1.0000   0.2658     -3.763   0.000367 ***
B:C:D       1.1000   0.2658      4.139   0.000104 ***
A:B:C:D     0.4750   0.2658      1.787   0.078613 .
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 2.377 on 64 degrees of freedom
Multiple R-squared: 0.7873,Adjusted R-squared: 0.7374
F-statistic: 15.79 on 15 and 64 DF, p-value: 4.547e-16
```

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- Notice this output does not involve the factors E, F, G or H – it only directly references factors A, B, C and D

- However, because of the confounding associated with the aliasing in this experiment
  - BCD interaction estimate corresponds to E's main effect
  - ACD interaction estimate corresponds to F's main effect
  - ABC interaction estimate corresponds to G's main effect
  - ABD interaction estimate corresponds to H's main effect

```
Coefficients:
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 8.5000    0.2658     31.985    < 2e-16 ***
A            0.8750    0.2658      3.293    0.001619 **
B            0.9250    0.2658      3.481    0.000906 ***
C            0.6250    0.2658      2.352    0.021772 *
D           -2.3000    0.2658     -8.655    2.27e-12 ***
E            1.1000    0.2658      4.139    0.000104 ***
F           -1.0000    0.2658     -3.763    0.000367 ***
G            1.5750    0.2658      5.927    1.35e-07 ***
H           -0.3000    0.2658     -1.129    0.263168
A:B         -0.3500    0.2658     -1.317    0.192532
A:C          1.3000    0.2658      4.892    7.07e-06 ***
A:D         -0.8750    0.2658     -3.293    0.001619 **
A:E          0.4750    0.2658      1.787    0.078613 .
A:F          0.3750    0.2658      1.411    0.163063
A:G          0.4500    0.2658      1.693    0.095261 .
A:H          1.2250    0.2658      4.610    1.98e-05 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 2.377 on 64 degrees of freedom
Multiple R-squared: 0.7873,Adjusted R-squared: 0.7374
F-statistic: 15.79 on 15 and 64 DF, p-value: 4.547e-16
```

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- All of the main effects – except H (fermentation temperature) – are significant
- Factors D, E, F, G (yeast/skin contact, stems, barrel toast, whole cluster) are most influential
- AC, AH and AD interactions are also significant
- Because of aliasing and confounding, it is equivalent to conclude that the DF, FG and EG interactions are significant
- Because factors D, E, F and G are most influential, it is likely that the DF, FG and EG interactions are responsible for the significant effect

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- Note that partial *F*-test of

$$H_0: \beta_H = \beta_{AB} = \beta_{AE} = \beta_{AF} = \beta_{AG} = 0$$

which compares the full model above to the one that is reduced by H_0 has an associated p-value of

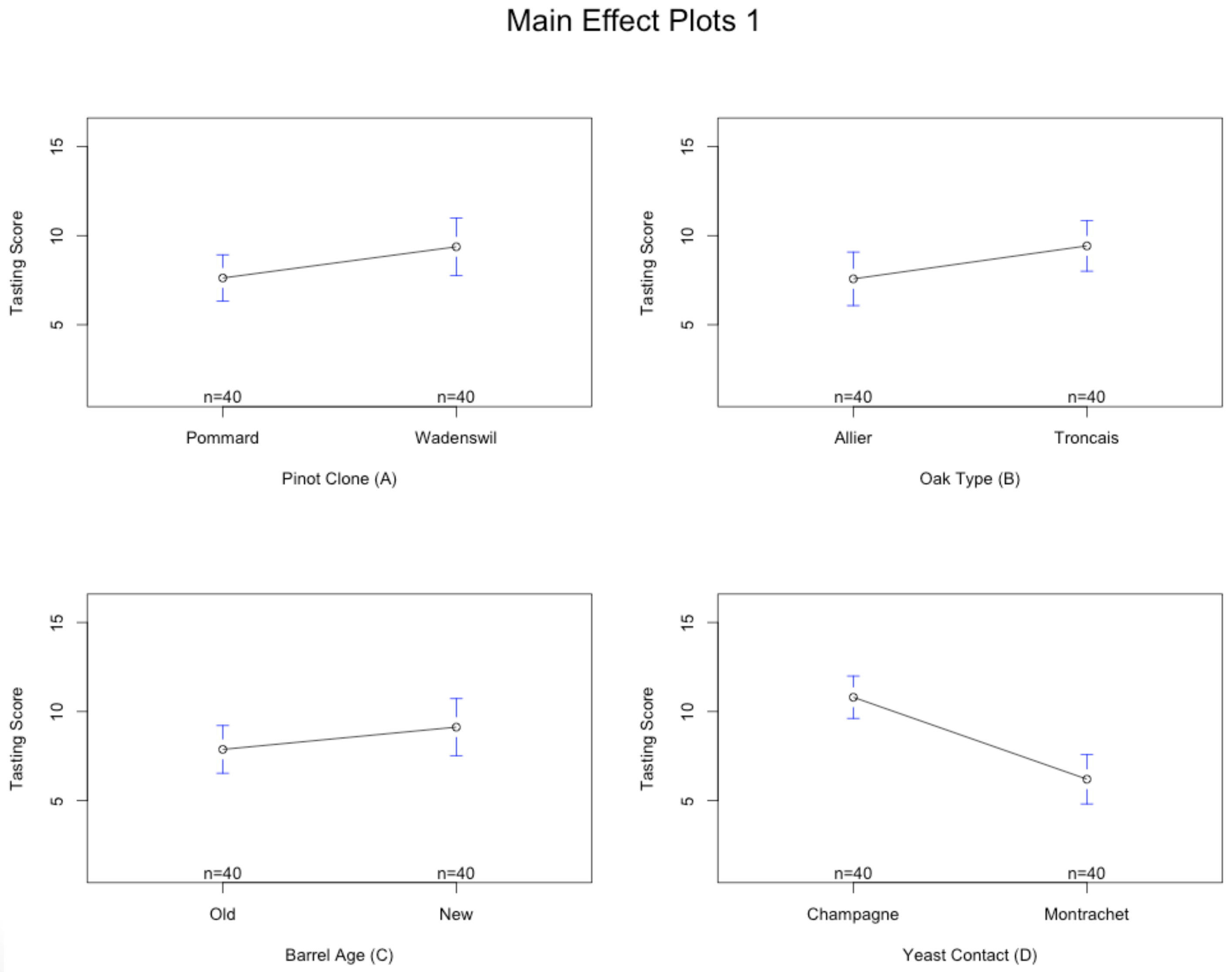
$$P(T \geq 2.2124) = 0.06375$$

where $T \sim F(5, 64)$

- Thus we do not reject H_0 and we conclude that all factors other than H are significantly influential, and the DF, FG and EG interactions are statistically significant

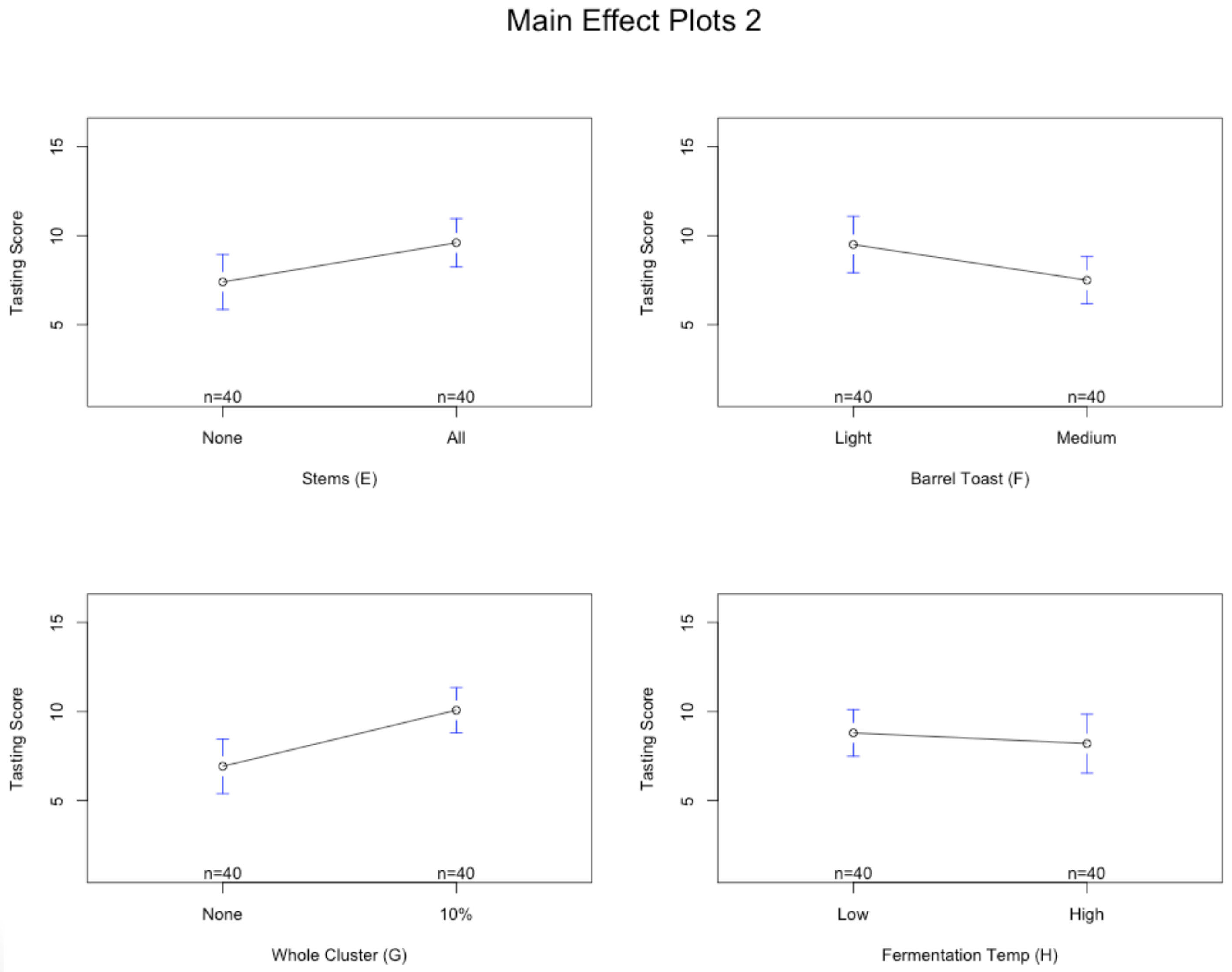
2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment



2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment



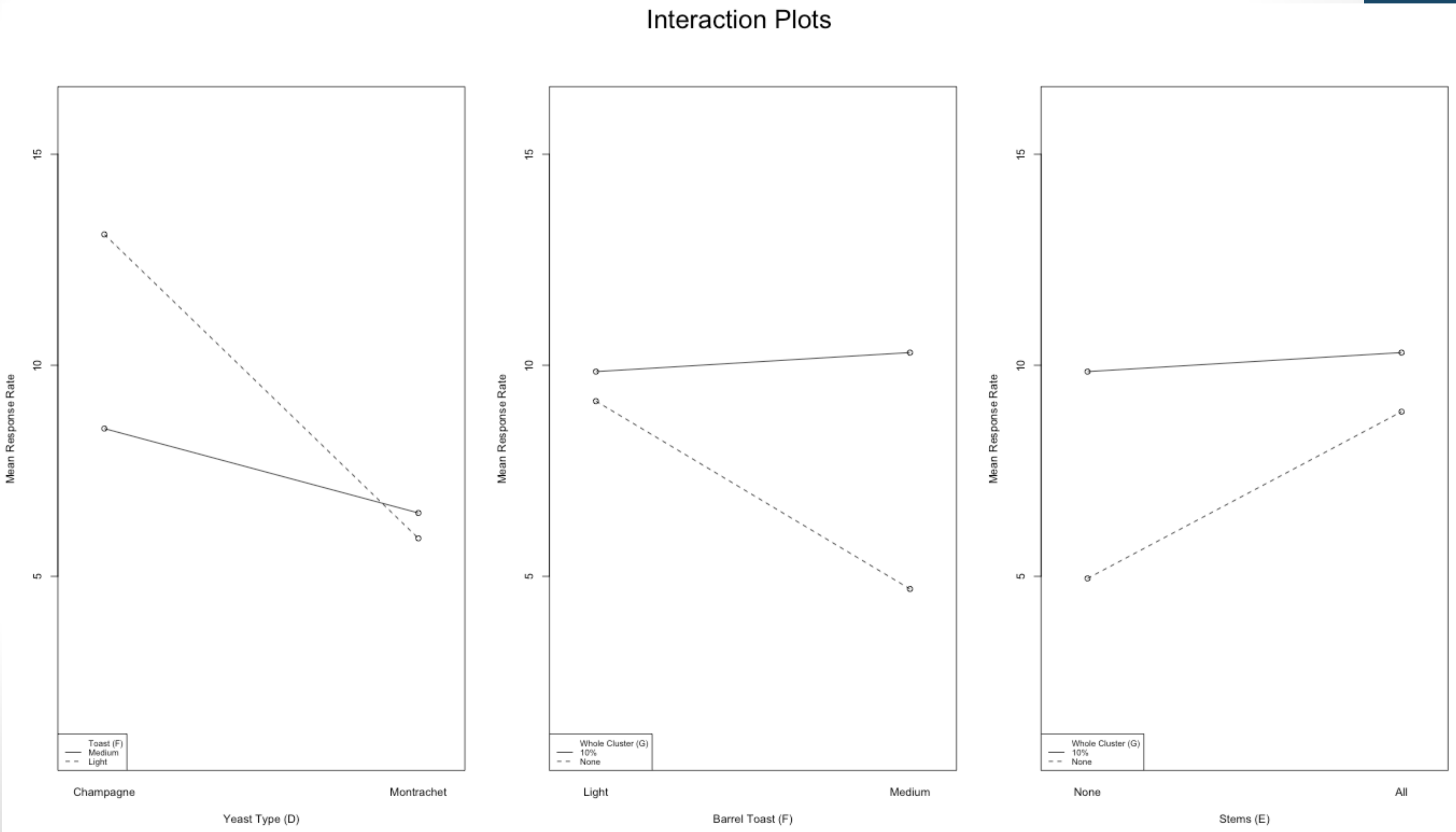
2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- The main effect plots suggest that:
  - yeast type (D) and the amount of whole clusters (G) used during fermentation are most important, with no whole clusters and Montrachet yeast producing a better tasting Pinot Noir
  - medium barrel toast (F) and no stems (E) also seem to correspond to a better tasting wine

2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment



2^{k-p} FRACTIONAL FACTORIAL

Example: The Chehalem Experiment

- The interaction effect plots suggest that:
  - If yeast type is Montrachet, the level of barrel toasting doesn't matter much, but if yeast type is Champagne, a medium barrel toast is best.
  - If barrel toast is chosen to be medium, then not including any whole-clusters is best
  - If using none of the stems, then it is also best not to include any whole-clusters

MULTI-ARMED BANDIT EXPERIMENTS

Multi-armed Bandit Experiments

- The comparison of $m \geq 2$ conditions, where the goal is to find the optimal condition, may be thought of as a multi-armed bandit problem

- A slot machine is colloquially referred to as a one-armed bandit

Multi-armed Bandit Experiments

- A row of slot machines is referred to as a multi-armed bandit

Multi-armed Bandit Experiments

- The goal, when faced with several slot machines, is to decide which one has the highest expected reward

- In other words, find the slot machine that is going to make you the most money, and repeatedly play that one

- Using notation from this course, let θ_j represent the expected reward from slot machine (a.k.a. "arm") j for $j = 1,2, \ldots, m$

- The goal is to find the best θ_j

Multi-armed Bandit Experiments

- **This is exactly what we've been doing!**
- The approach we have been taking is what some might call the "classical" approach:
- Collect a certain amount of data in accordance with Type I and Type II error constraints, and once that data has been observed, conduct a hypothesis test
- Typical multi-armed bandit solutions differ from classical experiments largely with respect to the exploration-exploitation trade off

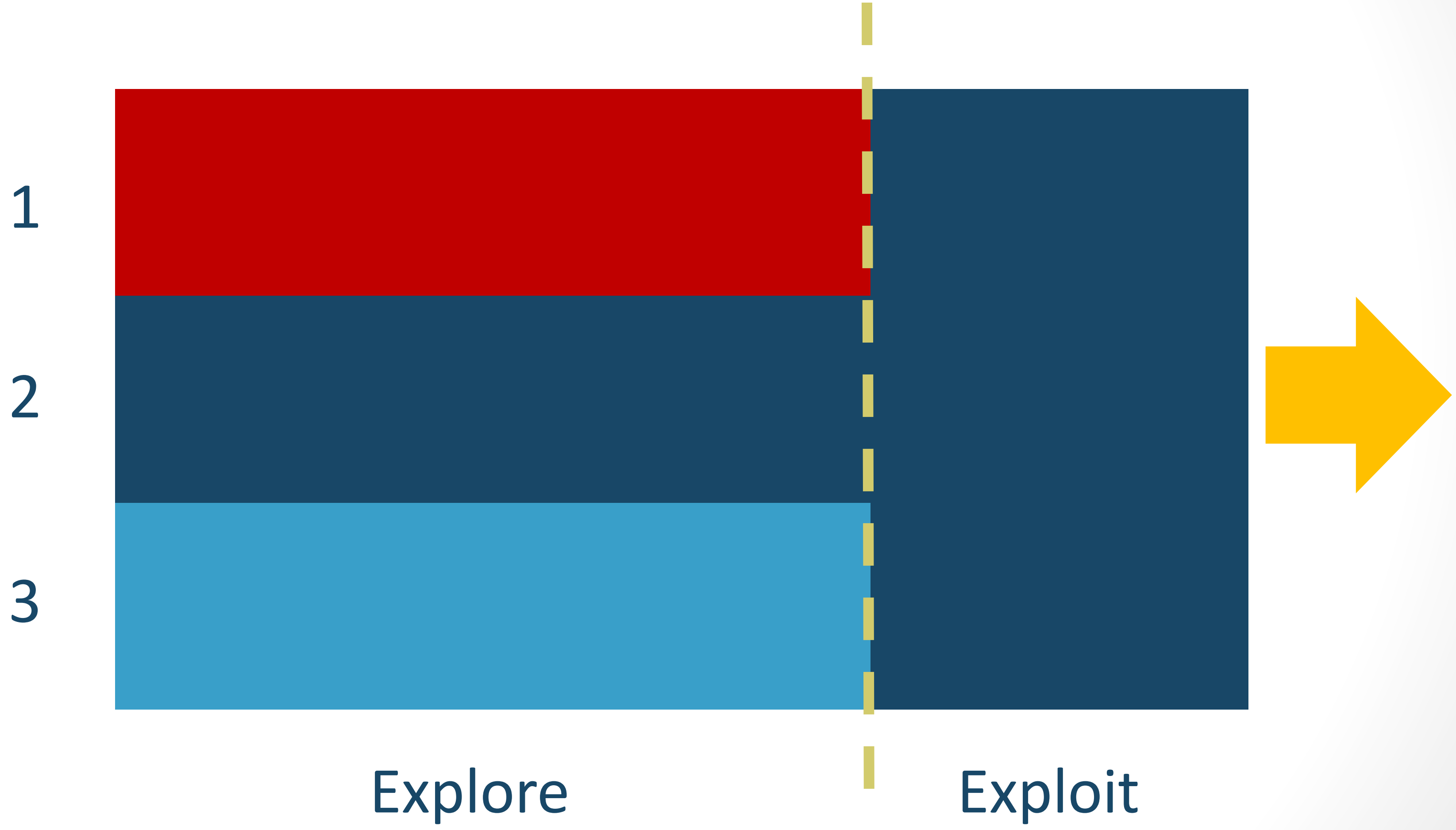
Multi-armed Bandit Experiments

- **Idea:** we need to explore all of the conditions, and when the optimal one is found, we want to exploit it

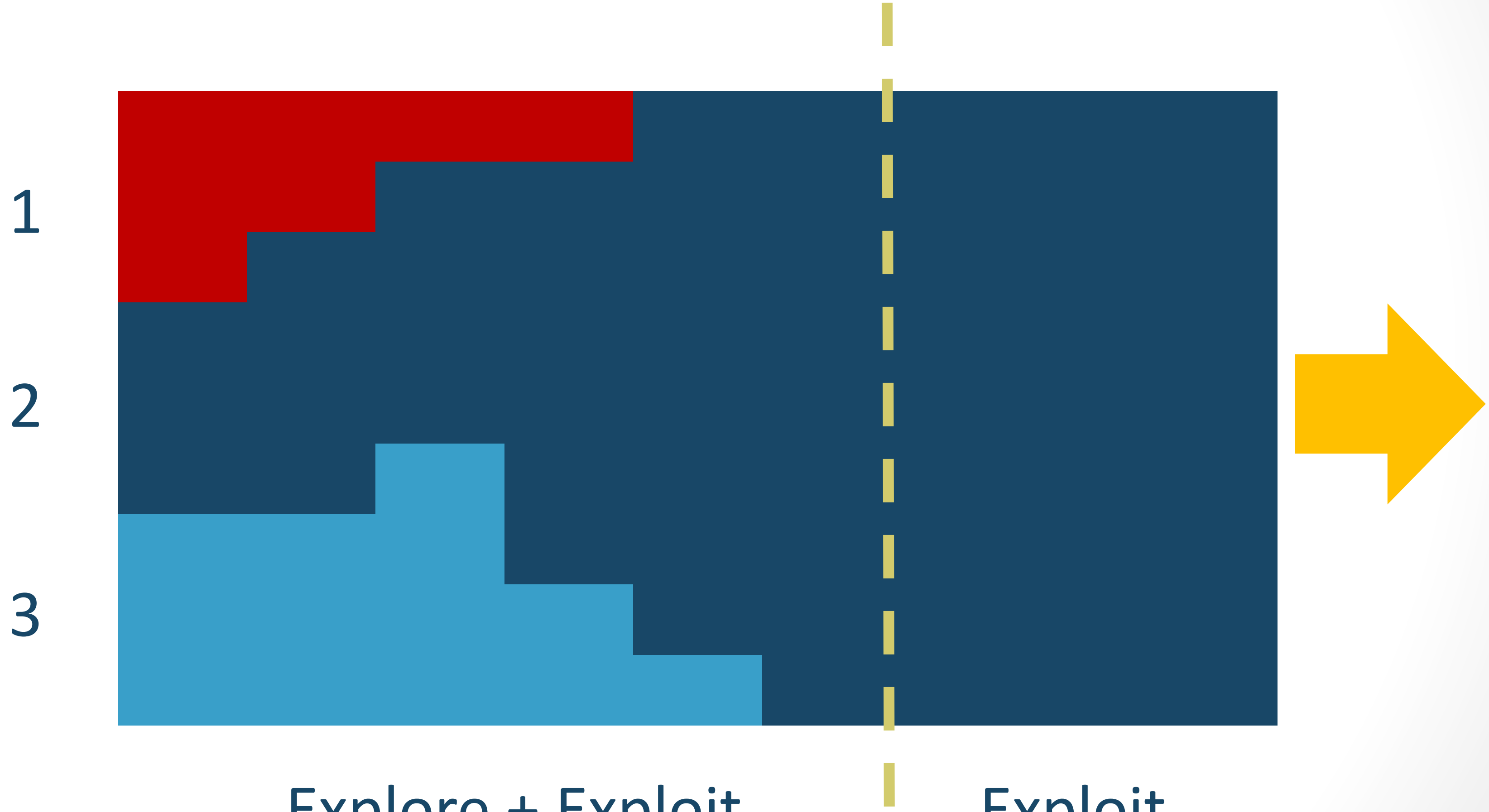
- Classical experiments represent a 100% exploration phase after which a 100% exploitation phase begins

- Multi-armed bandit experiments are typified by a combination of both exploration and exploitation as the experiment is conducted

Multi-armed Bandit Experiments



Multi-armed Bandit Experiments



Multi-armed Bandit Experiments

- Thus the classical approach to the problem requires that we spend the entire experimental period ”exploring” each condition at the same rate

- A multi-armed bandit experiment, on the other hand, does not provide equal allocation of units to conditions for the duration of the experiment

- At regular intervals the proportion of units allocated to each condition is updated to reflect the performance of each condition observed thus far

Multi-armed Bandit Experiments

- After an initial period of equal allocation, high performing conditions receive more units than lower performing conditions

- What does this updating rule look like specifically?

- In other words: How does this adaptive allocation of units to conditions work in practice?
  - Equal Allocation
  - Greedy Approach
  - Epsilon-Greedy Approach
  - Randomized Probability Matching

Multi-armed Bandit Experiments

Equal Allocation

- Each condition is allocated experimental units in equal proportions for the duration of the experiment

- This corresponds to the the classical experimental approach

- This represents 100% exploration during the experiment

- Detractors would say that this is inefficient and that the optimal condition can be found more quickly with an adaptive allocation strategy

Multi-armed Bandit Experiments

Greedy Approach

- Every experimental unit at a given point in time is assigned to the ‘best’ condition as determined by the data observed up to that point in time
- This approach is sometimes called “play-the-winner” and it represents 100% exploitation
- This approach may do a poor job at maximizing rewards as it does not adequately explore other conditions

Multi-armed Bandit Experiments

Epsilon-Greedy Approach

- This is a hybrid strategy that forces both exploration and exploitation

- Here allocation is performed via
  - the Greedy approach with probability $1 - \epsilon$
  - equal allocation with probability ϵ

- Thus a binary number is randomly generated using a Bernoulli distribution with probability of success/failure = $\epsilon/1 - \epsilon)$
  - If 1, perform greedy allocation
  - If 0, perform equal allocation

Multi-armed Bandit Experiments

Epsilon-Greedy Approach

- Notice that:
  - $\epsilon = 0$ corresponds to greedy allocation
  - $\epsilon = 1$ corresponds to equal allocation
- The choice of ϵ determines the desired balance of exploration and exploitation and is determined by the user
- Detractors would say that a drawback of the epsilon-greedy approach is that it will continue to explore even once an optimal condition has been found

Multi-armed Bandit Experiments

Randomized Probability Matching

Randomized probability matching (RPM) is a Bayesian alternative to the adaptive allocation strategies just discussed

- At successive time points, for each θ_j, the probability that θ_j is optimal is calculated

- This probability calculation is based on the joint posterior distribution of $(\theta_1, \theta_2, \ldots, \theta_m)$

- These m probabilities are then used as allocation weights in the next round of allocation

Multi-armed Bandit Experiments

Randomized Probability Matching

- For example, consider a standard A/B (two-condition) test

- Suppose that based on the rewards observed thus far:
  - condition A has a 0.73 probability of being the superior condition, and
  - condition B has a 0.27 probability of being the superior condition

- Then, during the next allocation round, condition A would receive 73% of the experimental units and condition B would receive 27% of them

Multi-armed Bandit Experiments

Randomized Probability Matching

- The experiment continues until the probability of optimality dominates for one of the conditions, while the other probabilities of superiority tend to zero

- If the optimality probability does not dominate for one condition in particular, it suggests that multiple conditions are equally optimal

- In the case of two conditions, this result would be manifested as the two optimality probabilities stabilizing at roughly 1/2 = 0.5

Multi-armed Bandit Experiments

Randomized Probability Matching

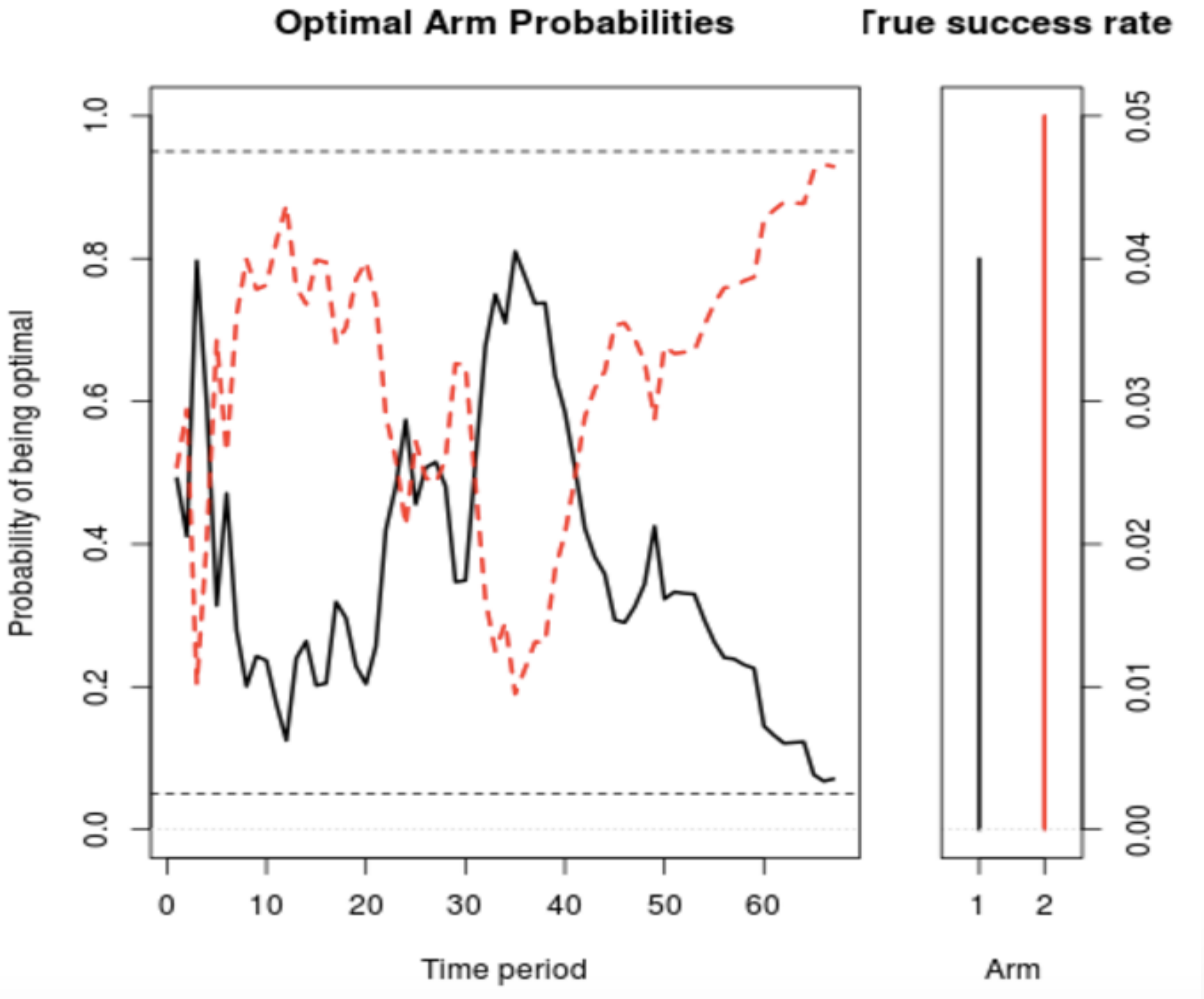


Image from: https://support.google.com/analytics/answer/2844870?hl=en

Multi-armed Bandit Experiments

Randomized Probability Matching

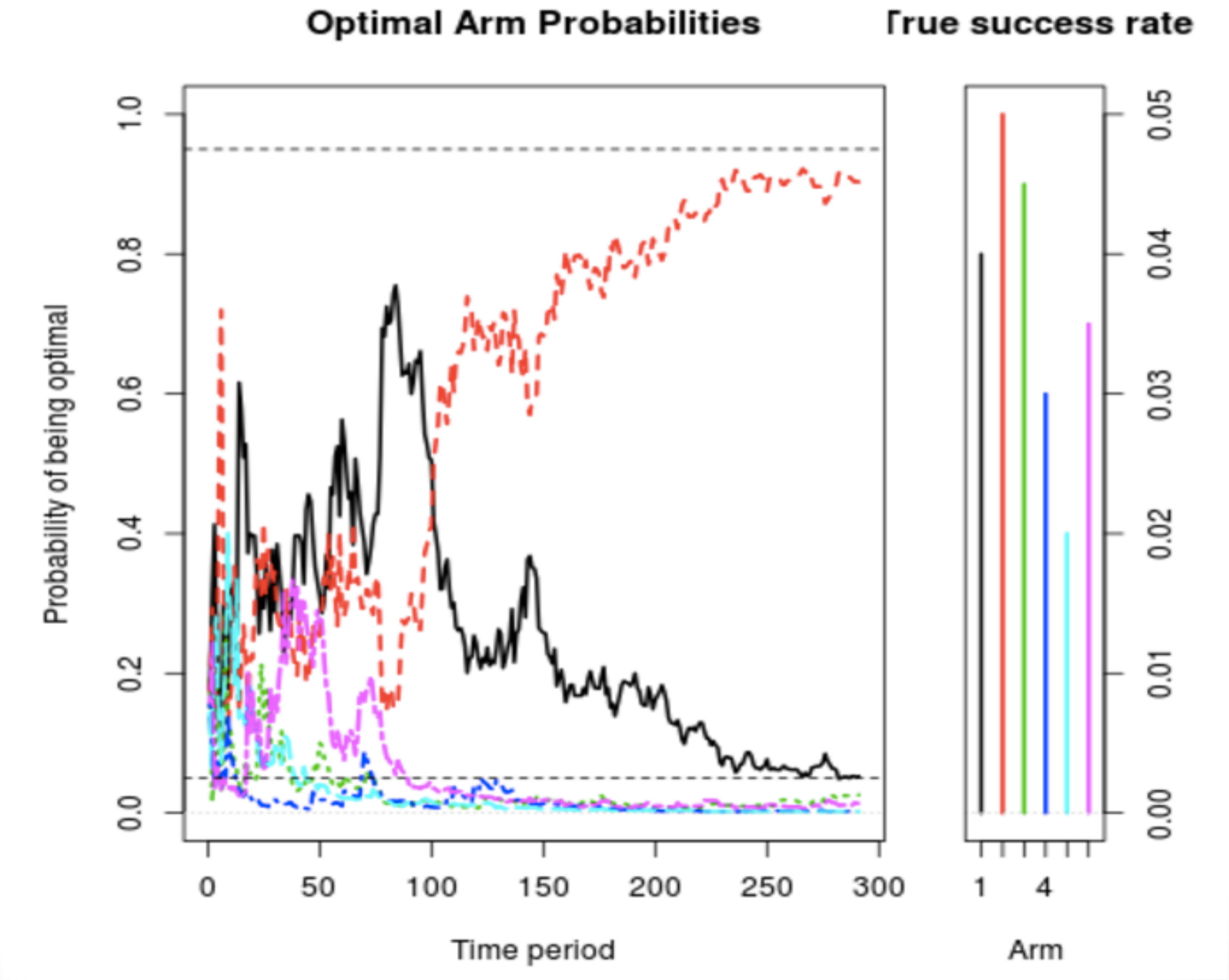


Image from: https://support.google.com/analytics/answer/2844870?hl=en

Multi-armed Bandit Experiments

Randomized Probability Matching

- RPM has been shown to do a better job at balancing exploration and exploitation than other allocation strategies, and is able to more quickly find the optimal condition
- However, it relies on understanding Bayesian statistics and being able to sample from a posterior distribution obtained from Markov chain Monte Carlo (MCMC) simulation
- That said, it is available as the default design and analysis strategy in the Google Analytics experimentation platform

Multi-armed Bandit Experiments

Advantages

- Multi-armed bandit experiments purport to find the optimal condition more quickly than the classical experimental approach
- And due to the exploitation within the experiment itself, fewer units are assigned to suboptimal conditions
- As such, these approaches reduce the opportunity cost associated with experimenting with a risky or inferior condition

Multi-armed Bandit Experiments

Disadvantages

- **BUT** this approach ignores the impact of Type I errors
- By adopting the multi-armed bandit approach you, in a sense, must believe that the consequences associated with this type of error are of no practical importance
- It also requires quick feedback and so it doesn’t work well when response observations are not obtained instantaneously
  - i.e., email advertising

A BRIEF SUMMARY (OF EVERYTHING)

A Brief Summary (of Everything)

- Designed Experiments are a useful tool for data scientists – in fact they are becoming a required tool

- They facilitate the identification, quantification, and understanding of causal relationships between a response and one or more factors

- Although A/B testing is common in the world of data science, there are many more informative, efficient, and useful approaches that may be applied – an effective data scientist will be familiar with these experimentation alternatives

THANK YOU!!