Experimental Design

STAT 430/STAT 830

Spring 2021 (1215)[1]

Cameron Roopnarine[2] Nathaniel Stevens[3]

January 11, 2022

[1]Online Course
[2]LaTeXer
[3]Instructor

Contents

1 INTRODUCTION **3**
- 1.1 Notation and Nomenclature . . . 3
- 1.2 Experiments versus Observational Studies . . . 5
- 1.3 QPDAC: A Strategy for Answering Questions with Data . . . 6
- 1.4 Fundamental Principles of Experimental Design . . . 8

2 EXPERIMENTS WITH TWO CONDITIONS **10**
- 2.1 Comparing Means in Two Conditions . . . 12
  - 2.1.1 The Two-Sample t-Test . . . 13
  - 2.1.2 When Assumptions are Invalid . . . 14
  - 2.1.3 Example: Instagram Ad Frequency . . . 15
- 2.2 Comparing Proportions in Two Conditions . . . 16
  - 2.2.1 Z-tests for Proportions . . . 16
  - 2.2.2 Example: Optimizing Optimizely . . . 16
- 2.3 Power Analysis and Sample Size Calculations . . . 17
- 2.4 Permutation and Randomization Tests . . . 20

3 EXPERIMENTS WITH MORE THAN TWO CONDITIONS **23**
- 3.1 Comparing Means in Multiple Conditions . . . 24
  - 3.1.1 The F-test for Overall Significance in a Linear Regression . . . 24
  - 3.1.2 Example: Candy Crush Boosters . . . 26
- 3.2 Comparing Proportions in Multiple Conditions . . . 26
  - 3.2.1 The Chi-squared Test of Independence . . . 27
  - 3.2.2 Example: Nike SB Video Ads . . . 28
- 3.3 The Problem of Multiple Comparisons . . . 29
  - 3.3.1 Family-Wise Error Rate . . . 30
  - 3.3.2 False Discovery Rate . . . 37
  - 3.3.3 Sample Size Determination . . . 39

4 BLOCKING **42**
- 4.1 Randomized Complete Block Designs . . . 43
  - 4.1.1 RCBD to Compare Means . . . 45
  - 4.1.2 Example: Promotions at The Gap . . . 46
  - 4.1.3 RCBD to Compare Proportions . . . 47
  - 4.1.4 Example: Enterprise Banner Ads . . . 47
- 4.2 Balanced Incomplete Block Designs . . . 48
  - 4.2.1 General Comments on the Design of a BIBD . . . 49
  - 4.2.2 General Comments on the Analysis of a BIBD . . . 50
- 4.3 Latin Square Designs . . . 51
  - 4.3.1 Latin Squares to Compare Means . . . 53
  - 4.3.2 Example: Netflix Latency . . . 55
  - 4.3.3 Latin Squares to Compare Proportions . . . 56
  - 4.3.4 Example: Uber Weekend Promos . . . 56

5 EXPERIMENTS WITH MULTIPLE DESIGN FACTORS 58
5.1 The Factorial Approach . . . 59
5.2 Designing a Factorial Experiment . . . 60
5.3 Analyzing a Factorial Experiment . . . 61
5.3.1 Continuous Response — The Instagram Example . . . 62
5.3.2 Binary Response — The TinyCo Example . . . 65
5.4 Two-Level Factorial Experiments . . . 69

6 2^K FACTORIAL EXPERIMENTS 70
6.1 Designing 2^K Factorial Experiments . . . 70
6.2 Analyzing 2^K Factorial Experiments . . . 71
6.2.1 An Intuition-Based Analysis . . . 72
6.2.2 A Regression-Based Analysis . . . 74
6.2.3 The Credit Card Example . . . 77

7 2^{K-p} FRACTIONAL FACTORIAL EXPERIMENTS 80
7.1 Designing 2^{K-p} Fractional Factorial Experiments . . . 83
7.1.1 Aliasing . . . 83
7.1.2 The Defining Relation . . . 85
7.1.3 Resolution . . . 87
7.1.4 Minimum Aberration . . . 88
7.2 Analyzing 2^{K-p} Fractional Factorial Experiments . . . 89
7.2.1 The Chehalem Example . . . 90

8 RESPONSE SURFACE METHODOLOGY 94
8.1 Overview of Response Optimization . . . 94
8.2 Method of Steepest Ascent/Descent . . . 96
8.2.1 The Path of Steepest Ascent/Descent . . . 98
8.2.2 Checking for Curvature . . . 99
8.2.3 The Netflix Example . . . 101
8.3 Response Surface Experiments . . . 104
8.3.1 The Central Composite Design . . . 106
8.3.2 The Lyft Example . . . 107
8.4 RSM with Qualitative Factors . . . 109

Chapter 1

INTRODUCTION

Week 1

1.1 Notation and Nomenclature

EXAMPLE 1.1.1: Experiment 1 — List View vs. Tile View

Suppose that Nike, the athletic apparel company, is experimenting with their mobile shopping interface, and they are interested in determining whether changing the user interface from *list view* to *tile view* will increase the proportion of customers that proceed to checkout.

EXAMPLE 1.1.2: Experiment 2 — Ad Themes

Suppose that Nixon, the watch and accessories brand, is experimenting with four different video ads that are to be shown on Instagram. The first has a surfing theme, the second has a rock climbing theme, the third has a camping theme, and the fourth has an urban professional theme. Interest lies in determining which of the four themes, on average, is watched the longest.

DEFINITION 1.1.3: Metric of interest

The **metric of interest** (MOI) is the statistic we wish the experiment investigates.

REMARK 1.1.4

Typically, we want to optimize for the metric of interest; that is, we would like to either maximize or minimize it.

EXAMPLE 1.1.5: Metric of Interest

- Key performance indicators (KPIs): a statistic that quantifies something about a business.
  - Click-through rates (CTRs).
  - Bounce rate.
  - Average time on page.
  - 95^{th} percentile page load time.
- *Nike Example*: checkout rate (COR).
- *Nixon Example*: average viewing duration (AVD).

DEFINITION 1.1.6: Response variable

The **response variable**, denoted y, is the variable of primary interest.

REMARK 1.1.7

The response variable is what needs to be measured in order for the MOI to be calculated.

EXAMPLE 1.1.8: Response Variable

- *Nike Example*: binary indicator indicating whether a customer checked out.
- *Nixon Example*: the continuous measurement of viewing duration for each user.

DEFINITION 1.1.9: Factor

The **factor**, denoted x, is the variable(s) of secondary interest.

Also known as: **covariates**, **explanatory variates**, **predictors**, **features**, **independent variables**.

REMARK 1.1.10

We usually think the factors influence the response (dependent) variable.

EXAMPLE 1.1.11: Factor

- *Nike Example*: the factor is the *visual layout*.
- *Nixon Example*: the factor is the *ad theme*.

DEFINITION 1.1.12: Experimental conditions

The **experimental conditions** are the unique combinations of levels of one or more factors.

Also known as: **treatments**, **variants**, **buckets**.

DEFINITION 1.1.13: Levels

The **levels** are the values that a factor takes on in an experiment.

EXAMPLE 1.1.14: Levels

- *Nike Example*: $\{\text{tile view}, \text{list view}\}$.
- *Nixon Example*: $\{\text{surfing}, \text{rock climbing}, \text{camping}, \text{business}\}$.

DEFINITION 1.1.15: Experimental units

The **experimental units** are what is assigned to the experimental conditions, and on which we measure the response variable.

EXAMPLE 1.1.16: Experimental Units

- *Nike Example*: Nike mobile customers.
- *Nixon Example*: Instagram users.

REMARK 1.1.17

Often, in online experiments, the unit is a user/customer (i.e., person), but it does not have to be.

EXAMPLE 1.1.18

Uber matching algorithm experiment.

1.2 Experiments versus Observational Studies

DEFINITION 1.2.1: Experiment

An **experiment** is a collection of conditions defined by *purposeful changes* to one or more factors. Here, we intervene in the data collection.

- The goal is to identify and quantify the differences in response variable values across conditions.
- In determining whether a factor significantly influences a response, like whether a video ad's theme significantly influences its AVD, it is necessary to understand how experimental units' response when exposed to each of the corresponding conditions.
- However, it would be nice if we could observe how the *same* units behave in each of the experimental conditions, but we can't. We only observe their response in a single condition.
- **Counterfactual**: the hypothetical and unobservable value of a unit's response in a condition to which they were not assigned. We may think of this as an "alternate reality."

  EXAMPLE 1.2.2

  Nixon Example: the "camping" response variable for units assigned to the "surfing" condition.

- Because counterfactual outcomes cannot be observed, we require a **proxy**. Instead, we randomly assign *different units* to *different experimental conditions*, and we compare their responses.
- Ideally, the only difference between the units in each condition is the fact that they are in different conditions.
  - We want the units to be as homogenous as possible, this will help facilitate **causal inference** (establishing causal connections between variables).
  - This is typically guaranteed by *randomization*.
- The key here is that we purposefully control the factors to observe the resulting effect on the response. This facilitates causal conclusions.
- In an **observational study**, on the other hand, there is no measure of control in the data collection process. Instead, we collect the data passively and the relationship between the response and factor(s) is observed organically.
- This hinders our ability to establish causal connections between the factor(s) and the response variables. However, sometimes we have no choice.

  EXAMPLE 1.2.3: Unethical Experiments

  - *Unethical Experiment 1*: In evaluating whether smoking lung cancer, it would be unethical to have a '*smoking*' condition in which we force the subjects to smoke.

- *Unethical Experiment 2*: In dynamic pricing experiments, it would be unethical to show different users different prices for the same products. For example, surge pricing in Uber/Lyft.
- *Unethical Experiment 3*: In social contagion experiments, it would be unethical to show some network users consistently negative content and others consistently positive content. But Facebook did this anyway.
- *Unethical Experiment 4*: Mozilla conducted an investigation in which the company was interested in determining whether Firefox users that installed an ad blocker were more engaged with the browser. However, it would have been unethical to force users to install an ad blocker, and so they were forced to perform an observational study with *propensity score matching* instead.

	Advantages	*Disadvantages*
Experiment	Causal inference is clean.	Experiments might be unethical, risky, or costly.
Observational Study	No additional cost, risk, or ethical concerns.	Causal inference is muddy.

1.3 QPDAC: A Strategy for Answering Questions with Data

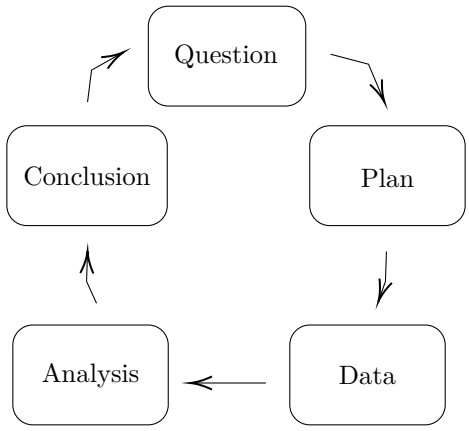


Figure 1.1: QPDAC Cycle

Question: Develop a clear statement of the question that needs to be answered.

- It is important that this is clear and concise and widely communicated, so all stakeholders are on the same page.
- The question should be quantifiable/measurable and typically stated in terms of the metric of interest.

EXAMPLE 1.3.1

- *Nike Example*: "which visual layout, tile view or list view, corresponds to the highest checkout rate?"
- *Nixon Example*: "which ad theme, camping, surfing, rock climbing, business, corresponds to the highest average viewing duration?"

Plan: In this stage, we design the experiment, and all pre-experimental questions should be answered.

- Choose the response variable. This should be dictated by the **Question** and the metric of interest.
- Choose the factor(s): brainstorm all factors that might influence the response and make decisions about whether and how they will be controlled in the experiment.
  - i **Design factors**: factors that we will manipulate in the experiment. The factors we've discussed in the Nike and Nixon examples are design factors.
  - ii **Nuisance factors**: factors that we expect to influence the response, but whose effect we do not care to quantify. Instead, we try to eliminate their effects with *blocking*.
  - iii **Allowed-to-vary factors**: factors that we *cannot* control and factors that we are unaware of in an experiment.
    - *Nixon Example*: users' age, gender, nationality.
- Choose the experimental units. These are what we measure the response variable on.
- Choose the sample size and sampling mechanism.
  - Sample size: how many units per experimental condition?
  - Sampling mechanism: how are they selected?

Data: In this stage, we collect the data according to the **Plan**. It is extremely important that we do this step correctly; the suitability and effectiveness of the analysis relies on the correctness of the data. Computer scientists often use the phrase "garbage in, garbage out" to describe the phenomenon whereby poor quality input will always provide faulty output.

- A/A Test: we assign units to one of two *identical* conditions.
  - We do this to ensure the assignment of units to conditions is truly random.
  - Two groups should be indistinguishable in terms of response distribution and other demographics.
  - If things aren't indistinguishable, there is a problem.
  - *Sample Ratio Mismatch Test*: If the ratio of users (or any randomization unit) between the variants is not close to the designed ratio, the experiment suffers from a Sample Ratio Mismatch (SRM).
    - Hypothesis test can be used to determine whether the proportion of units in each condition match what would have been expected under random assignment.

Analysis: In this stage, we statistically analyze **Data** to provide an objective answer to the **Question**.

- This is typically achieved by way of estimating parameters, fitting models, and carrying out statistical hypothesis tests. This is where we spend most of our time in the course.
- If the experiment was well-designed, and we collected the data correctly, this step should be straight-forward.

Conclusion: In this stage, we consider the results of the **Analysis**, and one must draw conclusions about what has been learned.

- We clearly communicate these conclusions to all parties involved in — or impacted by — the experiment.
- Communicating "wins" and "loses" will help to foster the culture of experimentation.

1.4 Fundamental Principles of Experimental Design

DEFINITION 1.4.1: Randomization

Randomization refers both to the manner in which we select experimental units for *inclusion* in the experiment and the manner in which we *assign* them to *experimental conditions*.

REMARK 1.4.2

Typically, we don't include the entire target/study population.

Thus, we have two levels of randomization:

- The first level of randomization exists to ensure the sample of units included in the experiment is *representative of those that were not*.
  - Allows us to generalize conclusions beyond just the experimental units to units in the population not in the experiment.
- The second level of randomization exists to *balance* the effects of *extraneous variables* not under study (i.e., the allowed-to-vary factors).
  - Balancing the effects of allowed-to-vary factors makes our conditions homogenous and thus best mimics the counterfactual, thereby making causal inference easy.

DEFINITION 1.4.3: Replication

Replication refers to the existence of multiple response observations within each experimental condition and thus corresponds to the situation in which we assign more than one unit to each condition.

- Assigning multiple units to each condition provides *assurance* that the observed results are genuine, and *not just due to chance*.
- For instance, consider the *Nike experiment* introduced previously. Suppose the CORs in the *list view* and *tile view* conditions were 0.5 and 1 respectively. This conclusion would be a lot more convincing if each condition had $n = 1000$ units as opposed to $n = 2$, where n is the sample size in *each* condition.

- How much replication do we need?
  - How big a sample size do we need?
  - Power analysis + sample size calculations will help answer this.

DEFINITION 1.4.4: Blocking

Blocking is the mechanism that we control the nuisance factors.

- To *eliminate* the influence of nuisance factors, we hold them fixed during the experiment.
- Thus, we run the experiment *at fixed levels of the nuisance factors*, i.e., within **blocks**.

EXAMPLE 1.4.5: GAP — Email Promotion

Consider an experiment in which the primary goal is to test different variations of the *message in the subject* line with the goal of maximizing '*open rate*.' However, suppose we know that the 'open rate' is also influenced by the "send time" (time of the day and the day of the week) of an email.

We send all the emails at the same time of day and on the same day of week to control/eliminate the effect of time/day nuisance factor. By *blocking*, in this way, the nuisance factor can't confound our conclusions.

Chapter 2

EXPERIMENTS WITH TWO CONDITIONS

Week 2

Anatomy of an A/B Test

- One design factor at two levels.
- We now consider the design and analysis of an experiment consisting of two experimental conditions — or what many data scientists broadly refer to as "A/B Testing" which is synonymous with "experimentation" in data science.
  - Canonical A/B test:

Figure 2.1: Canonical Button Colour Test.

Here, the metric of interest might be click-through-rate, which we're interested in maximizing.

- Other, more tangible examples:
  - Amazon
    - Checkout reassurances
    - List view vs. tile view
  - Airbnb
    - Host landing page redesign
    - Next available date
- Typically, the goal of such an experiment is to decide which condition is optimal with respect to some metric of interest θ. This could be a
  - mean (e.g., average time on page, average purchase size, average revenue per customer)
  - proportion (e.g., CTR, bounce rate, retention rate)
  - variance
  - quantile (e.g., median, 95^{th} percentile of page load time)

  - technically any statistic that can be from sample data

- Consider the button-colour example: imagine the observed click-through-rates (CTR) of the two conditions are: $\hat{\theta}_1 = 0.12$ (red) and $\hat{\theta}_2 = 0.03$ (blue).
  - Obviously, $\hat{\theta}_1 > \hat{\theta}_2$, but does that mean that $\theta_1 > \theta_2$?

- Formally, we phrase such a question as a statistical hypothesis that we test using the data collected from the experiment.
  - $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 \neq \theta_2$ (two-sided).
  - $\mathbf{H}_0$: $\theta_1 \leq \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 > \theta_2$ (one-sided).
  - $\mathbf{H}_0$: $\theta_1 \geq \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 < \theta_2$ (one-sided).

- "Absence of evidence $\neq$ evidence of absence."

- No matter which hypothesis is appropriate, the goal is always the same: based on the observed data, we will decide to *reject* $\mathbf{H}_0$ or *not reject* $\mathbf{H}_0$.

- In order to draw such a conclusion, we will define a **test statistic**.

DEFINITION 2.0.1: Test statistic

The **test statistic**, denoted T, is a random variable that satisfies three properties:

(i) It must be a function of the observed data.

(ii) It must be a function of the parameters θ_1 and θ_2.

(iii) Its distribution must not depend on θ_1 or θ_2.

- Assuming the null hypothesis is true, the test statistic T follows a particular distribution which we call the **null distribution**. For example, $\mathcal{N}(0, 1)$, $t(\texttt{df})$, $F(\texttt{df1}, \texttt{df2})$, $\chi^2(\texttt{df})$.

- We then calculate t, the observed value of the test statistic, and evaluate its extremity relative to the null distribution.
  - If t is very extreme, this suggests that perhaps the null hypothesis is not true.
  - If t appears as though it could have come from the null distribution, then there is no reason to disbelieve the null hypothesis.

- We formalize the extremity of t using the ***p*-value** of the test.

DEFINITION 2.0.2: p-value

The probability of observing a value of the test statistic *at least as extreme* as the value we observed, if the null hypothesis is true.

  - Thus, the p-value formally quantifies how "extreme" the observed test statistic is.
  - The more extreme the value of t, the smaller the p-value, and the more evidence we have against it.

- How "extreme" t must be, and hence how small the p-value must be to reject $\mathbf{H}_0$, is determined by the **significance level** of the test, denoted α.
  - If p-value $\leq \alpha$, we reject $\mathbf{H}_0$.
  - If p-value $> \alpha$, we do not reject $\mathbf{H}_0$.

REMARK 2.0.3

Common choices of α are 0.05 and 0.01.

- In order to choose α, one must understand the two types of errors that can be made when drawing conclusions in the context of a hypothesis test.
- Recall that by design, either $\mathbf{H}_0$ or $\mathbf{H}_A$ is true. This means that there are four possible outcomes when using data to decide which statement is true:
  (1) No Error: $\mathbf{H}_0$ is true, and we correctly do not reject it.
  (2) Type I Error: $\mathbf{H}_0$ is true, and we incorrectly reject it.
  (3) Type II Error: $\mathbf{H}_0$ is false, and we incorrectly do not reject it.
  (4) No Error: $\mathbf{H}_0$ is false, and we correctly reject it.
- We would like to reduce the likelihood of making either type of error.
  - But there are different consequences of each type of error.
  - So we may wish to treat them differently.

EXAMPLE 2.0.4: Pregnancy Test

$\mathbf{H}_0$: person is not pregnant versus $\mathbf{H}_A$: person is pregnant.

- Type I Error: a non-pregnant person is pregnant (false positive).
- Type II Error: a pregnant person is not pregnant (false negative).

EXAMPLE 2.0.5: Courtroom

Consider a courtroom analogy where we assume the defendant is innocent until proven guilty. Formally,

$\mathbf{H}_0$: the defendant is innocent versus $\mathbf{H}_A$: the defendant is guilty.

- Type I Error: sentencing an innocent person to jail.
- Type II Error: letting a guilty person go free.

DEFINITION 2.0.6: Significance level

The **significance level** of a test is $\alpha = \mathbb{P}(\text{Type I Error})$.

DEFINITION 2.0.7: Power

The **power** of a test is $1 - \beta$ where $\beta = \mathbb{P}(\text{Type II Error})$.

- Fortunately, it is possible to control the frequency in which these types of errors occur.
- It is desirable to have a test with a small significance level, and a large power.

2.1 Comparing Means in Two Conditions

- Here, we restrict attention to the situation in which we measure the response variable of interest on a continuous scale.

- We assume that the response observations collected in the two conditions follow normal distributions, and in particular

$$Y_{i1} \sim \mathcal{N}(\mu_1, \sigma^2) \text{ and } Y_{i2} \sim \mathcal{N}(\mu_2, \sigma^2), \quad i = 1, 2, \ldots, n_j \text{ for } j = 1, 2.$$

  - Y_{ij} = response observation for unit i in condition j.

- Using the observed data, we test hypotheses of the form:
  - $\mathbf{H}_0$: $\mu_1 = \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 \neq \mu_2$.
  - $\mathbf{H}_0$: $\mu_1 \leq \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 > \mu_2$.
  - $\mathbf{H}_0$: $\mu_1 \geq \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 < \mu_2$.

2.1.1 The Two-Sample t-Test

STATISTICAL TEST 2.1.1: Student's t-test

- *Purpose*: Compare μ_1 versus μ_2 (assuming $\sigma_1 = \sigma_2$ are unknown).
- *Test Statistic*:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \overbrace{(\mu_1 - \mu_2)}^{0}}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

  - $\hat{\sigma}$ is our estimator.
  - $t(n_1 + n_2 - 2)$ is our null distribution.

- *Observed Version*:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  - $\bar{y}_j = \dfrac{1}{n_j}\displaystyle\sum_{i=1}^{n_j} y_{ij} = \hat{\mu}_j$.
  - $\hat{\sigma}^2 = \dfrac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$.
  - $\hat{\sigma}_j^2 = \dfrac{1}{n_j - 1}\displaystyle\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$.

- *p-value Calculation*:
  - For $\mathbf{H}_0$: $\mu_1 = \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 \neq \mu_2$, we compute p-value $= \mathbb{P}(T \geq |t|) + \mathbb{P}(T \leq -|t|)$.
  - For $\mathbf{H}_0$: $\mu_1 \leq \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 > \mu_2$, we compute p-value $= \mathbb{P}(T \geq t)$.
  - For $\mathbf{H}_0$: $\mu_1 \geq \mu_2$ versus $\mathbf{H}_\text{A}$: $\mu_1 < \mu_2$, we compute p-value $= \mathbb{P}(T \leq t)$.

REMARK 2.1.2

In all cases above, $T \sim t(n_1 + n_2 - 2)$.

2.1.2 When Assumptions are Invalid

STATISTICAL TEST 2.1.3: Welch's t-test

- *Purpose*: Compare μ_1 versus μ_2 (assuming $\sigma_1 \neq \sigma_2$ are unknown).
- *Test Statistic*: "Approximately," we have

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \overbrace{(\mu_1 - \mu_2)}^{0}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \dot{\sim} t(\nu)$$

  where

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1 - 1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2}} \approx \min(n_1, n_2) - 1$$

- *Observed Version*:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

- *p-value Calculation*: Same as Statistical Test 2.1.1, but where the null distribution is $T \sim t(\nu)$.

STATISTICAL TEST 2.1.4: F-test for Variances

- *Purpose*:
  - $\mathbf{H}_0$: $\sigma_1^2 = \sigma_2^2$ versus $\mathbf{H}_\mathrm{A}$: $\sigma_1^2 \neq \sigma_2^2$.
  - $\mathbf{H}_0$: $\sigma_1^2/\sigma_2^2 = 1$ versus $\mathbf{H}_\mathrm{A}$: $\sigma_1^2/\sigma_2^2 \neq 1$.
- *Test Statistic*:

$$T = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F(n_1 - 1, n_2 - 1)$$

- *Observed Version*:

$$t = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \in \mathbf{R}$$

- *p-value Calculation*:
  - If $t \geq 1$, then p-value $= \mathbb{P}(T \geq t) + \mathbb{P}(T \leq 1/t)$.
  - If $t < 1$, then p-value $= \mathbb{P}(T \leq t) + \mathbb{P}(T \geq 1/t)$.

REMARK 2.1.5

In all cases above, $T \sim F(n_1 - 1, n_2 - 1)$.

2.1.3 Example: Instagram Ad Frequency

EXAMPLE 2.1.6: Instagram Ad frequency

- Suppose that you are a data scientist at Instagram, and you are interested in running an experiment to learn about the influence of ad frequency on user engagement.
- Currently, users see an ad every 8 posts in their social feed, but, in order to increase ad revenue, your manager is pressuring your team to show an ad every 5 posts.
  - Condition 1: 7:1 Ad Frequency
  - Condition 2: 4:1 Ad Frequency
- You are justifiably nervous about this change, and you worry that this will substantially decrease user engagement and hurt the overall user experience.
- The metric of interest you choose to optimize for is μ = average session time (where y = the length of time a user engages within the app, in minutes).
- The hypothesis here is:
$$\mathbf{H}_0\text{: } \mu_1 \leq \mu_2 \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_1 > \mu_2$$
- The data summaries are:
  - $n_1 = 500$, $\hat{\mu}_1 = \bar{y}_1 = 4.916$, $\hat{\sigma}_1 = s_1 = 0.963$.
  - $n_2 = 500$, $\hat{\mu}_2 = \bar{y}_2 = 3.052$, $\hat{\sigma}_2 = s_2 = 0.995$.

F-test:

- $t = \hat{\sigma}_1^2/\hat{\sigma}_2^2 = (0.963)^2/(0.995)^2 = 0.938$.
- p-value $= \mathbb{P}(T \leq 0.938) + \mathbb{P}(T \geq 1/0.938) = 0.472$ where $T \sim F(499, 499)$.
- This p-value is larger than any ordinary α, so we do not reject $\mathbf{H}_0$: $\sigma_1^2 = \sigma_2^2$, and so we continue with Student's t-test.

Student's t-test:

- $\hat{\sigma}^2 = \dfrac{499(0.963)^2 + 499(0.995)^2}{998} = (0.979)^2$.
- $t = \dfrac{4.916 - 3.052}{0.979\sqrt{\frac{1}{500} + \frac{1}{500}}} = 30.101$.
- p-value $= \mathbb{P}(T \geq 30.101) = 1.838 \times 10^{-142}$ where $T \sim t(998)$.
- This p-value is much smaller than any typical α, and so we reject $\mathbf{H}_0$: $\mu_1 \leq \mu_2$, and conclude that increasing ad frequency significantly reduces average session duration.

[R Code] `Comparing_two_means`

2.2 Comparing Proportions in Two Conditions

- Here, we restrict attention to the situation in which the response variable of interest is binary, indicating whether an experimental unit did, or did not, perform some action of interest. In cases like these, we let

$$Y_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in condition } j \text{ performs an action of interest} \\ 0 & \text{if unit } i \text{ in condition } j \text{ does not perform an action of interest} \end{cases} \quad \begin{matrix} i = 1, 2, \ldots, n_j \\ j = 1, 2 \end{matrix}$$

- Because the Y_{ij}'s are binary, it is common to assume that they follow a Bernoulli distribution; that is, $Y_{ij} \sim \text{Binomial}(1, \pi_j)$ where π_j represents the probability that $Y_{ij} = 1$. That is, the probability that unit i from condition j performs the "action of interest."
- Using the observed data, we test hypotheses of the form:
  - $\mathbf{H}_0$: $\pi_1 = \pi_2$ versus $\mathbf{H}_\text{A}$: $\pi_1 \neq \pi_2$.
  - $\mathbf{H}_0$: $\pi_1 \leq \pi_2$ versus $\mathbf{H}_\text{A}$: $\pi_1 > \pi_2$.
  - $\mathbf{H}_0$: $\pi_1 \geq \pi_2$ versus $\mathbf{H}_\text{A}$: $\pi_1 < \pi_2$.

2.2.1 Z-tests for Proportions

STATISTICAL TEST 2.2.1: Z-test for Proportions

- *Purpose*: Compare π_1 versus π_2.
- *Test Statistic*: "Approximately," we have

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \overbrace{(\pi_1 - \pi_2)}^{0}}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \dot{\sim} \mathcal{N}(0,1)$$

  where $\hat{\pi} = \dfrac{n\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \dfrac{\#\text{ units who performed action}}{\text{total }\#\text{ units in exp.}}$ and $\hat{\pi}_j = \bar{y}_j$.

- *Observed Version*:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- *p-value Calculation*: The p-values are calculated in the same way as in the t-tests, except here that $T \sim \mathcal{N}(0,1)$.

2.2.2 Example: Optimizing Optimizely

EXAMPLE 2.2.2: Optimizing Optimizely

- During a website redesign, Optimizely was interested in how new versions of certain pages influenced things like conversion and engagement relative to the old version.
- One outcome they were interested in was whether the redesigned homepage lead to a significant increase in the number of new accounts created.
  - Condition 1: original homepage.

  - Condition 2: redesigned homepage.

- The metric of interest here is π = conversion rate (where $y = 1$ if a homepage visitor signed up and 0 otherwise).
- The hypothesis tested here is:
$$\mathbf{H}_0\text{: } \pi_1 \geq \pi_2 \text{ versus } \mathbf{H}_A\text{: } \pi_1 < \pi_2$$
- We summarize the data from this experiment in a 2×2 contingency table:

		Condition 1	2	
Conversion	Yes	280	399	679
	No	8592	8243	16835
		8872	8642	17514

- $\hat{\pi}_1 = 280/8872 = 0.032$ and $\hat{\pi}_2 = 399/8642 = 0.046$. Thus,
$$\hat{\pi} = \frac{8872(0.032) + 8642(0.046)}{17514} = 0.039$$
$$t = \frac{0.032 - 0.046}{\sqrt{(0.039)(1-0.039)(1/8872 + 1/8642)}} = -5.007$$
- p-value $= \mathbb{P}(T \leq -5.007) = 2.758 \times 10^{-7}$ where $T \sim \mathcal{N}(0,1)$.
- We reject $\mathbf{H}_0$ and conclude that the redesigned homepage significantly increases conversion rate.
- [R Code] `Comparing_two_proportions`

2.3 Power Analysis and Sample Size Calculations

- Used to control Type II Error.
- Power analyses help determine required sample sizes.
- Suppose, for illustration, that we are interested in testing the hypothesis:
$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 \text{ versus } \mathbf{H}_A\text{: } \theta_1 \neq \theta_2$$
- Suppose, also for illustration, that the test statistic associated with this test has the form:
$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \overbrace{(\theta_1 - \theta_2)}^{0}}{\sqrt{\frac{\mathbb{V}(Y_1)}{n} + \frac{\mathbb{V}(Y_2)}{n}}} \sim \mathcal{N}(0,1)$$

DEFINITION 2.3.1: Rejection region

The **rejection region**, denoted $\mathcal{R}$, is all the values of the observed test statistic t that would lead to the rejection of $\mathbf{H}_0$:
$$\mathcal{R} = \{t : \mathbf{H}_0 \text{ is rejected}\}$$

- If $t \in \mathcal{R}$, we reject $\mathbf{H}_0$.

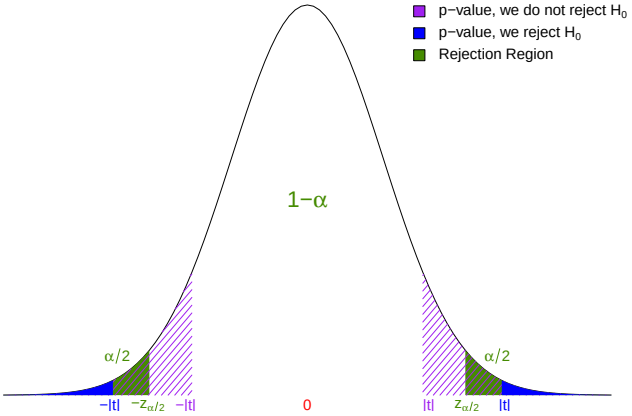


Figure 2.2: $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 \neq \theta_2$
$\mathcal{R} = \{t : t \leq -z_{\alpha/2} \text{ or } t \geq z_{\alpha/2}\}$

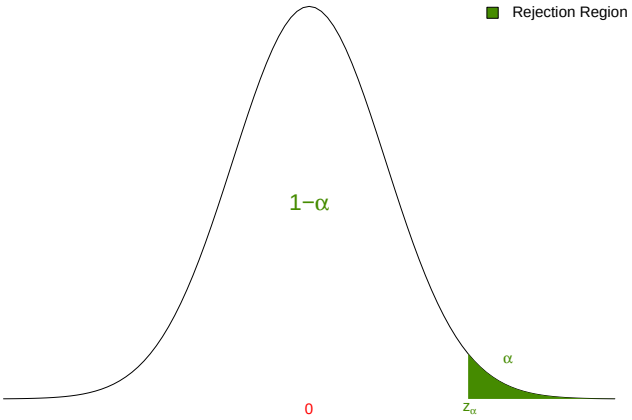


Figure 2.3: $\mathbf{H}_0$: $\theta_1 \leq \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 > \theta_2$
$\mathcal{R} = \{t : t \geq z_\alpha\}$

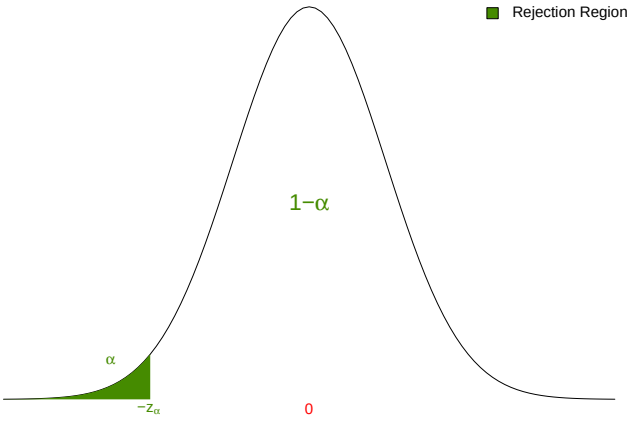


Figure 2.4: $\mathbf{H}_0$: $\theta_1 \geq \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 < \theta_2$
$\mathcal{R} = \{t : t \leq -z_\alpha\}$

- If $t \in \mathcal{R}^c$, we do not reject $\mathbf{H}_0$.

- Defining Type I and Type II error rates in terms of a rejection region is also useful:
  - $\alpha = \mathbb{P}(\text{Type I Error}) = \mathbb{P}(\text{Reject } \mathbf{H}_0 \mid \mathbf{H}_0 \text{ is true}) = \mathbb{P}(T \in \mathcal{R} \mid \mathbf{H}_0 \text{ is true})$.
  - $\beta = \mathbb{P}(\text{Type II Error}) = \mathbb{P}(\text{Do Not Reject } \mathbf{H}_0 \mid \mathbf{H}_0 \text{ is false}) = \mathbb{P}(T \in \mathcal{R}^c \mid \mathbf{H}_0 \text{ is false})$.

$$
\begin{aligned}
1-\beta &= \text{Power} \\
&= 1 - \mathbb{P}(\text{Type II Error}) \\
&= 1 - \mathbb{P}(T \in \mathcal{R}^c \mid \mathbf{H}_0 \text{ is false}) \\
&= \mathbb{P}(T \in \mathcal{R} \mid \mathbf{H}_0 \text{ is false}) \\
&= \mathbb{P}\Big(T \geq z_{\alpha/2} \cup T \leq -z_{\alpha/2} \,\Big|\, \mathbf{H}_0 \text{ is false}\Big) \\
&= \mathbb{P}\Big(T \geq z_{\alpha/2} \,\Big|\, \mathbf{H}_0 \text{ is false}\Big) + \mathbb{P}\Big(T \leq -z_{\alpha/2} \,\Big|\, \mathbf{H}_0 \text{ is false}\Big) \\
&= \mathbb{P}\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \geq z_{\alpha/2} \,\middle|\, \mathbf{H}_0 \text{ is false}\right) + \mathbb{P}\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \leq -z_{\alpha/2} \,\middle|\, \mathbf{H}_0 \text{ is false}\right)
\end{aligned}
$$

Assuming $\mathbf{H}_0$ is true, $\theta_1 - \theta_2 = 0$ and $\dfrac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \sim \mathcal{N}(0,1)$. However, $\mathbf{H}_0$ is false, which means that $\theta_1 - \theta_2 = \delta$ for some $\delta \neq 0$. Thus,

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \sim \mathcal{N}(0,1)$$

Therefore, we need to account for this. Let $Z \sim \mathcal{N}(0,1)$, then

$$
\begin{aligned}
1-\beta &= \mathbb{P}\left(\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}\right) + \mathbb{P}\left(\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \leq -z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}\right) \\
&= \mathbb{P}\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}\right) + \mathbb{P}\left(Z \leq -z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}\right)
\end{aligned}
$$

Think about what happens to these terms when δ is positive versus negative. Without loss of generality, assume $\delta > 0$, in which case

$$1-\beta = \mathbb{P}\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}\right)$$

We know that $\mathbb{P}\big(Z \geq z_{1-\beta}\big) = 1-\beta$, therefore

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}}$$

Doing some algebra yields

$$n = \frac{(z_{\alpha/2} - z_{1-\beta})^2\big[\mathbb{V}(Y_1) + \mathbb{V}(Y_2)\big]}{\delta^2}$$

- $\mathbb{V}(Y_1)$ and $\mathbb{V}(Y_2)$ are the variances of the response in the two conditions. This needs to be guessed or determined by historical information.

- $\delta = \theta_1 - \theta_2$ is called the **minimum detectable effect** (MDE).

DEFINITION 2.3.2: Minimum detectable effect (MDE)

The **minimum detectable effect**, denoted δ, is the smallest difference between conditions (i.e, between θ_1 and θ_2) that we find to be practically relevant and that we would like to detect as being statistically significant.

WEEK 3

2.4 Permutation and Randomization Tests

- All the previous tests have made some kind of distributional assumption for the response measurements, such as $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ or $Y_{ij} \sim \text{Binomial}(1, \pi_j)$.
- It would be preferable to have a test that does not rely on *any* assumptions.
- This is precisely the purpose of permutation and randomization tests.
  - These tests are *non-parametric* and rely on resampling.
  - The motivation is that if $\mathbf{H}_0$: $\theta_1 = \theta_2$ is true, any random rearrangement of the data is *equally likely to have been observed*. If $\mathbf{H}_0$ is true, then we have a single population/distribution from which our data has been drawn.
  - With n_1 and n_2 units in each condition, there are

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$

    arrangements of the $n_1 + n_2$ observations into two groups of size n_1 and n_2 respectively.

$$n_1 = n_2 = 50 \implies \binom{n_1 + n_2}{n_1} = \binom{100}{50} = 1.009 \times 10^{29}$$

- A true **permutation test** considers *all possible rearrangements* of the original data.
  - The test statistic t is calculated on the original data and on every one of its rearrangements.
  - This collection of test statistic values generate the empirical null distribution.
- In a **randomization test**, we do not consider all possible rearrangements.
  - We just consider a large number N of them.
  - We use this in practice instead of a permutation test because the exact permutation tests have too many permutations to consider.

Randomization Test Algorithm

1. Collect response observations in each condition.

$$\{y_{11}, y_{21}, \dots, y_{n_1 1}\} \to \hat{\theta}_1$$

$$\{y_{12}, y_{22}, \dots, y_{n_2 2}\} \to \hat{\theta}_2$$

2. Calculate the test statistic t on the original data.

$$t = \hat{\theta}_1 - \hat{\theta}_2 \quad \text{or} \quad t = \frac{\hat{\theta}_1}{\hat{\theta}_2}$$

3. Pool all the observations together and randomly sample (without replacement) n_1 observations which will be assigned to "Condition 1" and the remaining n_2 observations that are assigned to "Condition 2." Repeat this N times.

$$\{y^\star_{11}, y^\star_{21}, \dots, y^\star_{n_1 1}\} \to \hat{\theta}^\star_1$$

$$\{y^\star_{12}, y^\star_{22}, \dots, y^\star_{n_2 2}\} \to \hat{\theta}^\star_2$$

4. Calculate the test statistic $t^\star_k$ on each of the "shuffled" datasets, $k = 1, 2, \dots, N$.

$$t^\star_k = \hat{\theta}^\star_{1,k} - \hat{\theta}^\star_{2,k} \quad \text{or} \quad t^\star_k = \frac{\hat{\theta}^\star_{1,k}}{\hat{\theta}^\star_{2,k}}$$

5. Compare to t to $\{t^\star_1, t^\star_2, \dots, t^\star_N\}$, the empirical null distribution, and calculate the p-value:

$$p\text{-value} = \frac{\#\text{ of } t^\star\text{'s that are at least as extreme as } t}{N} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\{t^\star_k \text{ at least as extreme as } t\}$$

- $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 \neq \theta_2$. If $t = \hat{\theta}_1 - \hat{\theta}_2$, then the p-value is:

$$p\text{-value} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\{t^\star_k \geq |t| \cup t^\star_k \leq -|t|\}$$

- $\mathbf{H}_0$: $\theta_1 \geq \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 < \theta_2$. If $t = \hat{\theta}_1 - \hat{\theta}_2$, then the p-value is:

$$p\text{-value} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\{t^\star_k \leq t\}$$

- $\mathbf{H}_0$: $\theta_1 \leq \theta_2$ versus $\mathbf{H}_\text{A}$: $\theta_1 > \theta_2$. If $t = \hat{\theta}_1 - \hat{\theta}_2$, then the p-value is:

$$p\text{-value} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\{t^\star_k \geq t\}$$

EXAMPLE 2.4.1: Pokémon Go

- Suppose that Niantic Inc, is experimenting with two different promotions within Pokémon Go:
  - Condition 1: Give users nothing.
  - Condition 2: Give users 200 free Pokécoins.
  - Condition 3: Give users a 50 % discount on Shop purchases.
- In a small pilot experiment, we randomize $n_1 = n_2 = n_3 = 100$ users to each condition.
- For each user, we record the amount of real money (in USD) they spend in the 30 days following the experiment.
- The data summaries are:
  - $\bar{y}_1 = \$\,10.740$, $Q_{y_1}(0.5) = \$\,9.000$.
  - $\bar{y}_2 = \$\,9.530$, $Q_{y_2}(0.5) = \$\,8.000$.

- $\bar{y}_3 = \$\,13.410$, $Q_{y_3}(0.5) = \$\,10.000$.

Using R, we performed a randomization test with $N = 10\,000$ with respect to the mean we found that the control and free coin conditions did not significantly differ. But there was a significant increase in the amount of money spent in the discount condition relative to the other two.

The hypotheses that we tested to determine these conclusions were:

$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_1 \neq \mu_2$$
$$\mathbf{H}_0\text{: } \mu_1 \geq \mu_2 \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_1 < \mu_2$$

Interestingly, when you run these same tests, but on the basis of the median, we find no significant difference between any of the conditions.

- [R Code] `Randomization_test`

Chapter 3

EXPERIMENTS WITH MORE THAN TWO CONDITIONS

Anatomy of an "A/B/m" Test

- One design factor at m levels.
- We will now consider a design and analysis of an experiment consisting of more than two experimental conditions — or what many data scientists broadly refer to as "A/B/m Testing."
  - Canonical A/B/m test:

Figure 3.1: Canonical Button Colour Test.

    What colour maximizes click-through rate?

- Other, more tangible, examples:
  - Netflix.
  - Etsy.
- Typically, the goal of such an experiment is to decide which condition is optimal with respect to some metric of interest θ. This could be a:
  - mean
  - proportion
  - variance
  - quantile
  - technically any statistic that can be calculated from sample data
- From a design standpoint, such an experiment is *very* similar to a two-condition experiment.
  1. Choose a metric of interest θ which addresses the question you are trying to answer.
  2. Determine the response variable y that must be measured on each unit to estimate $\hat{\theta}$.
  3. Choose the design factor x and the m levels you will experiment with.

4. Choose $n_1, n_2, \ldots, n_m$ and assign units to conditions at random.
5. Collect the data and estimate the metric of interest in each condition:

$$\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$$

- Determining which condition is optimal typically involves a series of pairwise comparisons: t-tests, z-tests, or randomization tests.
- But it is useful to begin such an investigation with a *gatekeeper* test (test of overall equality) which serves to determine whether there is *any* difference between the m experimental conditions. Formally, we phrase such a question as the following statistical hypothesis:

$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } \mathbf{H}_\text{A}\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k$$

In the case of means:

$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_m \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$

In the case of proportions:

$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m \text{ versus } \mathbf{H}_\text{A}\text{: } \pi_j \neq \pi_k \text{ for some } j \neq k$$

3.1 Comparing Means in Multiple Conditions

- We assume that our response variable follows a normal distribution, and we assume that the mean of the distribution depends on the condition in which we take the measurements, and that the variance is the same across all conditions.

$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2) \quad \text{for } i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, m$$

- We use an F-test to test for means:

$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_m \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$

3.1.1 The F-test for Overall Significance in a Linear Regression

- In particular, we use the F-test for overall significance in an *appropriately defined linear regression model*:
  - The *appropriately defined linear regression model* in this situation is one in which the response variable depends on $m - 1$ indicator variables:

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is in condition } j \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } j = 1, 2, \ldots, m-1.$$

  - For a particular unit i, we adopt the model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{m-1} x_{i,m-1} + \varepsilon_i$$

    * Y_i = response observation for unit $i = 1, 2, \ldots, N = \sum_{j=1}^{m} n_j$.
    * ε_i = random error term which we assume follows a $\mathcal{N}(0, \sigma^2)$ distribution independently for all $i = 1, 2, \ldots, N$.
    * Because we're about to do a regression analysis, the usual *residual diagnostics* are relevant.
  - In this model the β's are unknown parameters, and we interpret them in the context of the following expectations:
    * Expected response in condition m:

$$\mathbb{E}[Y_i \mid x_{i1} = x_{i2} = \cdots = x_{i,m-1} = 0] = \beta_0 = \mu_m$$

* Expected response in condition j:

$$\mathbb{E}[Y_i \mid x_{ij} = 1] = \beta_0 + \beta_j = \mu_j \quad \text{for } j = 1, 2, \ldots, m-1$$

* β_0 is the expected response in condition m.
* β_j is the expected difference in response value in condition j versus condition m for $j = 1, 2, \ldots, m-1$.

$$\begin{aligned} \mu_1 &= \beta_0 + \beta_1 \\ \mu_2 &= \beta_0 + \beta_2 \\ &\vdots \\ \mu_{m-1} &= \beta_0 + \beta_{m-1} \\ \mu_m &= \beta_0 \end{aligned}$$

– Based on these assumptions $\mathbf{H}_0$: $\theta_1 = \theta_2 = \cdots = \theta_m$ is true if and only if $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$, and hence is equivalent to testing:

$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j$$

– This hypothesis corresponds, as noted, to the F-test for overall significance in the model.

- In regression parlance, the test statistic is the ratio of the regression mean squares (MSR) to the mean squared error (MSE) in a standard regression-based analysis of variance (ANOVA):

$$t = \frac{\text{MSR}}{\text{MSE}}$$

- In our setting we can more intuitively think of the test statistic as comparing the response variability between conditions to the response variability within conditions:

– Average response in condition j: $\bar{y}_{\bullet j} = \dfrac{1}{n_j} \displaystyle\sum_{i=1}^{n_j} y_{ij}$.

– Overall average response: $\bar{y}_{\bullet\bullet} = \dfrac{1}{N} \displaystyle\sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij} = \dfrac{1}{N} \sum_{j=1}^{m} n_j \bar{y}_{\bullet j}$.

– Quantifies variability <u>*between*</u> conditions: $\text{SS}_\text{C} = \displaystyle\sum_{j=1}^{m} \sum_{i=1}^{n_j} (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2$.

– Quantifies variability <u>*within*</u> conditions: $\text{SS}_\text{E} = \displaystyle\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet j})^2$.

– Quantifies <u>*overall*</u> variability: $\text{SS}_\text{T} = \displaystyle\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\bullet\bullet})^2 = \text{SS}_\text{C} + \text{SS}_\text{E}$.

- The null distribution for this test is $F(m-1, N-m)$.
- p-value $= \mathbb{P}(T \geq t)$ where $T \sim F(m-1, N-m)$.
- If $\mathbf{H}_0$: $\mu_1 = \cdots = \mu_m$ is true, then $\mathbb{E}[\text{MS}_\text{C}] = \sigma^2$ and $\mathbb{E}[\text{MS}_\text{E}] = \sigma^2$.

Table 3.1: ANOVA Table

Source	SS	d.f.	MS	Test Statistic
Condition	SS_C	$m-1$	$MS_C = SS_C/(m-1)$	$t = MS_C/MS_E$
Error	SS_E	$N-m$	$MS_E = SS_E/(N-m)$	
Total	SS_T	$N-1$		

3.1.2 Example: Candy Crush Boosters

- Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the colour bomb.
- We randomize each user to one of these three conditions ($n_1 = 121$, $n_2 = 135$, $n_3 = 117$) and they receive (for free) 5 boosters corresponding to their condition. Interest lies in evaluating the effect of these different boosters on the length of time a user plays the game.
- Let μ_j represent the average length of game play (in minutes) associated with booster condition $j = 1, 2, 3$. While interest lies in finding the condition associated with the longest average length of game play, here we first rule out the possibility that booster type does not influence the length of game play (i.e., $\mu_1 = \mu_2 = \mu_3$).
- In order to do this we fit the linear regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

  where x_1 and x_2 are indicator variables indicating whether we observe a particular value of the response in the jelly fish or colour bomb conditions, respectively. The lollipop hammer is therefore the reference condition.
- In R, we found that the test statistic for testing:

$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \mu_3 \text{ versus } \mathbf{H}_A\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$

  was $t = 851.895$ and the null distribution was $T \sim F(2, 370)$. The corresponding p-value was:

$$p\text{-value} = \mathbb{P}(T \geq 851.895) = 3.280 \times 10^{-139}$$

- Therefore, we have very strong evidence against $\mathbf{H}_0$ and conclude that the average length of game play is not the same in the three booster conditions.
- [R Code] `Comparing_multiple_means`

3.2 Comparing Proportions in Multiple Conditions

- As is always the case when comparing proportions is of interest, we assume that our response variable is binary:

$$Y_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in condition } j \text{ performs an action of interest} \\ 0 & \text{if unit } i \text{ in condition } j \text{ does not perform an action of interest} \end{cases} \quad \begin{matrix} i = 1, 2, \ldots, n_j \\ j = 1, 2, \ldots, m \end{matrix}$$

- $Y_{ij} \sim \text{Binomial}(1, \pi_j)$ where π_j is the probability of a unit in condition j performing the action.
- We use a **chi-squared test of independence** (Pearson χ^2 test) to test for proportions:

$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m \text{ versus } \mathbf{H}_A\text{: } \pi_j \neq \pi_k \text{ for some } j \neq k.$$

3.2.1 The Chi-squared Test of Independence

- We use the chi-squared test of independence as a test for 'no association' between two categorical variables that are summarized in a *contingency table*.
- We apply this methodology here to test the independence of the binary outcome (whether a unit performs the action of interest) and the particular condition they are in.
- To start, let's assume that $m = 2$, and let's use the Optimizely experiment as a reference.
  - If $\pi_1 = \pi_2 = \pi$, then we would expect the conversion rate in each condition to be the same.
  - An estimate of the pooled conversion rate in this case is $\hat{\pi} = 679/17514 = 0.039$.
  - Let X = number of conversions in a condition with n units, therefore $X \sim \text{Binomial}(n, \pi)$ where $\mathbb{E}[X] = n\pi$.
  - Therefore, we would expect $n_1\hat{\pi} = 8872(0.039) = 343.958$ conversions in condition 1, and $n_2\hat{\pi} = 8642(0.039) = 335.042$ conversions in condition 2.
  - The chi-squared test formally evaluates if the difference between what was observed and what is expected under the null hypothesis is large enough to be considered *significantly* different.
  - The *general* 2×2 contingency table for a scenario like this is shown in Table 3.2.

Table 3.2: A General 2×2 Contingency Table

		Condition 1	2	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	O_0
		n_1	n_2	$n_1 + n_2$

    - $O_{\ell,j}$: observed number of conversions ($\ell = 1$), and the observed number of non-conversions ($\ell = 0$) in condition $j = 1, 2$.
    - O_ℓ: overall number of conversions ($\ell = 1$) or non-conversions ($\ell = 0$)
  - So,
$$\hat{\pi} = \frac{O_1}{n_1 + n_2} \quad \text{and} \quad 1 - \hat{\pi} = \frac{O_0}{n_1 + n_2}$$
represent the overall proportions of units that did or did not convert, and they are estimates of overall conversion and non-conversion rates.
  - Let $E_{1,j}$ and $E_{0,j}$ represent the expected number of conversions and non-conversions in condition $j = 1, 2$,
$$E_{1,j} = n_j\hat{\pi} \quad \text{and} \quad E_{0,j} = n_j(1 - \hat{\pi})$$
    - This is what we expect if $\mathbf{H}_0$: $\pi_1 = \pi_2$ is true.
  - The χ^2 test statistic compares the observed count in each cell to the corresponding expected count, and is defined as
$$T = \sum_{\ell=0}^{1}\sum_{j=1}^{2} \frac{(O_{\ell_j} - E_{\ell,j})^2}{E_{\ell,j}} \sim \chi^2(1)$$
  - p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(1)$.
  - Returning to the Optimizely example, the *expected* table is Table 3.3.
  - And the resultant test statistic and p-value are:
$$t = \frac{(280 - 343.958)^2}{343.958} + \frac{(399 - 335.042)^2}{335.042} + \frac{(8592 - 8528.042)^2}{8528.042} + \frac{(8243 - 8306.958)^2}{8306.958} = 25.075$$
$$p\text{-value} = \mathbb{P}(T \geq 25.075) = 5.516 \times 10^{-7} \quad \text{where } T \sim \chi^2(1)$$

Table 3.3: 2×2 Contingency Table for Optimizely's Homepage Experiment

		Condition 1	2	
Conversion	Yes	343.958	335.042	679
	No	8528.042	8306.958	16835
		8872	8642	17514

- Let's now extend this for $m > 2$.
  - We've used the chi-squared test is a test of 'no association' between the binary outcome (whether a unit performs the action of interest) and the particular condition they are in.
    - But there is no requirement that there be only two conditions.
    - Here we generalize the test to any number of experimental conditions.
  - The information associated with this test can be summarized in a $2 \times m$ contingency table as seen in Table 3.4.

Table 3.4: A General $2 \times m$ Contingency Table

		Condition 1	2	$\cdots$	m	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	$\cdots$	$O_{1,m}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	$\cdots$	$O_{0,m}$	O_0
		n_1	n_2	$\cdots$	n_m	$N = \sum_{j=1}^{m} n_j$

    - # of conversions ($\ell = 1$) or non-conversions ($\ell = 0$) is condition $j = 1, 2$.
    - $\hat{\pi} = O_1/N$.
    - $1 - \hat{\pi} = O_0/N$.
  - We compare each of the observed frequencies $O_{1,j}$ with the corresponding expected frequency $E_{\ell,j}$.

$$E_{1,j} = n_j\hat{\pi} \quad \text{and} \quad E_{0,j} = n_j(1 - \hat{\pi})$$

    - Expected number of conversions/non-conversions in condition j assuming $\mathbf{H}_0$: $\pi_1 = \pi_2 = \cdots = \pi_m$ is true.
  - The χ^2 test statistic compares the observed count in each cell to the corresponding expected count, and is defined as:

$$T = \sum_{\ell=0}^{1}\sum_{j=1}^{m} \frac{(O_{\ell,j} - E_{\ell,j})^2}{E_{\ell_j}} \sim \chi^2(m-1)$$

  - p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(m-1)$.

3.2.2 Example: Nike SB Video Ads

- Suppose that Nike is running an ad campaign for Nike SB, their skateboarding division, and the campaign involves $m = 5$ different video ads that are being shown in Facebook newsfeeds.
- A video ad is 'viewed' if it is watched for longer than 3 seconds, and interest lies in determining which ad is most popular and hence most profitable by comparing the viewing rates of the five different videos.
- We show each of these 5 videos to $n_1 = 5014$, $n_2 = 4971$, $n_3 = 5030$, $n_4 = 5007$, and $n_5 = 4980$ users, and summarize the results in Table 3.5.

Table 3.5: A 2×5 Observed Contingency Table for the Nike Example

		Condition					
		1	2	3	4	5	
View	Yes	160	95	141	293	197	886
	No	4854	4876	4889	4714	4783	24116
		5014	4971	5030	5007	4980	25002

- The overall watch rate (and its complement) are:

$$\hat{\pi} = \frac{O_1}{N} = \frac{886}{25002} = 0.0354 \quad \text{and} \quad 1 - \hat{\pi} = \frac{24116}{25002} = 0.9649$$

- We multiply n_j by $\hat{\pi}$ and $(1 - \hat{\pi})$ for $j = 1, 2, 3, 4, 5$ to get the expected cell frequencies in Table 3.6.

Table 3.6: A 2×5 Expected Contingency Table for the Nike Example

		Condition					
		1	2	3	4	5	
View	Yes	177.68	176.16	178.25	177.43	176.48	886
	No	4836.32	4794.84	4851.75	4829.57	4803.52	24116
		5014	4971	5030	5007	4980	25002

- The resultant test statistic and p-value (where $T \sim \chi^2(4)$) are:

$$t = \sum_{\ell=0}^{1} \sum_{j=1}^{m} \frac{(O_{\ell,j} - E_{\ell,j})^2}{E_{\ell,j}} = 129.1686$$

$$p\text{-value} = \mathbb{P}(T \geq 129.1686) = 5.86 \times 10^{-27}$$

- Therefore, we reject $\mathbf{H}_0$: $\pi_1 = \pi_2 = \cdots = \pi_5$ and conclude that the "watch-rate" is not the same for each of the video ads.
- [R Code] `Comparing_multiple_proportions`

WEEK 4

3.3 The Problem of Multiple Comparisons

- We have seen that "gatekeeper" tests of overall equality such as:
$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } \mathbf{H}_\text{A}\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k$$
are often rejected.
- We may follow this up with a series of pairwise comparisons to determine which condition(s) is (are) optimal.
  - We already know how to do this!
    - Z-tests, t-tests, F-tests, χ^2-tests, randomization tests.
- HOWEVER, when doing multiple comparisons like this, we encounter the **multiple comparison** or **multiple testing problem**.
  - Type I Errors are more likely to occur in a family of tests than an individual test.
- To frame this discussion, let's define some notation:

  - M: the number of hypotheses tested.
  - M_0: the number of true null hypotheses.
  - M_A: the number of false null hypotheses.
  - R: the number of null hypotheses that we reject.
  - $M - R$: the number of null hypotheses that we don't reject.
  - V: the number of true null hypotheses that were incorrectly rejected; that is, the number of Type I Errors.
  - S: the number of false null hypotheses that were incorrectly rejected.
  - U: the number of true null hypotheses that were correctly accepted.
  - T: the number of false null hypotheses that were incorrectly accepted; that is, the number of Type II Errors.
  - $M = M_0 + M_A$.

- We summarize the outcomes of these M decisions in Table 3.7.

Table 3.7: Outcomes From M Simultaneous Hypothesis Tests

		Decision		
		Reject $\mathbf{H}_0$	Accept $\mathbf{H}_0$	
Truth	$\mathbf{H}_0$ is True	V	U	M_0
	$\mathbf{H}_0$ is False	S	T	M_A
		R	$M - R$	M

  - R and $M - R$ are observable.
  - M_0, M_A, V, U, S, T are random variables; that is, the random process of collecting data and testing the M hypotheses determines their values. Therefore, they are all unobservable.

- Ideally, we would like V and T to be small.
  - T is controlled via sample size as it is related to power.
  - We control functions of V with sophisticated and clever statistical methods.

3.3.1 Family-Wise Error Rate

DEFINITION 3.3.1: Family-wise error rate

The **family-wise error rate** is the probability of committing a Type I Error in *any* of the M hypothesis tests.

$$\text{FWER} = \mathbb{P}(V \geq 1)$$

That is, the probability of making at least one Type I Error in M tests.

- If we use a significance level of α for each of the M tests, the FWER will be much greater than α.
- Boole's Inequality, which is $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, provides an upper

bound:

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}(\text{At least one Type I Error in } M \text{ tests}) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{M} \text{Type I Error on test } k\right) \\
&\leq \sum_{k=1}^{M} \mathbb{P}(\text{Type I Error on test } k) && \text{Boole's Inequality} \\
&= \sum_{k=1}^{M} \alpha \\
&= M\alpha
\end{aligned}$$

EXAMPLE 3.3.2: FWER

If $M = 10$ and $\alpha = 0.05$, then FWER ≤ 0.5.

- If we're willing to assume that the M tests are independent then:

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}(\text{At least one Type I Error in } M \text{ tests}) \\
&= 1 - \mathbb{P}(\text{No Type I Error in } M \text{ tests}) \\
&= 1 - \mathbb{P}\left(\bigcap_{k=1}^{M} \text{No Type I Error on test } k\right) \\
&= 1 - \prod_{k=1}^{M} \mathbb{P}(\text{No Type I Error on test } k) && \text{by independence} \\
&= 1 - \prod_{k=1}^{M} (1-\alpha) \\
&= 1 - (1-\alpha)^M
\end{aligned}$$

- This error rate, as a function of M can be seen in Figure 3.2. As M increases, FWER also increases. In fact, $\lim_{M \to \infty} \text{FWER} = 1$.
- A common value of M is $\binom{m}{2}$: the number of pairwise comparisons necessary to compare each condition to every other condition.

EXAMPLE 3.3.3

If $m = 5$ and $\alpha = 0.05$, then $M = \binom{5}{2} = 10$. Therefore, FWER $= 1 - (1 - 0.05)^{10} = 0.4013$.

- Available to us are a variety of different statistical techniques that may be used to ensure the FWER does not exceed some threshold.

$$\text{FWER} \leq \alpha^\star \in [0, 1]$$

REMARK 3.3.4: General Notation

- Denote the M null hypotheses as: $\mathbf{H}_{0,1}, \mathbf{H}_{0,2}, \ldots, \mathbf{H}_{0,M}$.
- Denote their corresponding p-values as: $p_1, p_2, \ldots, p_M$.

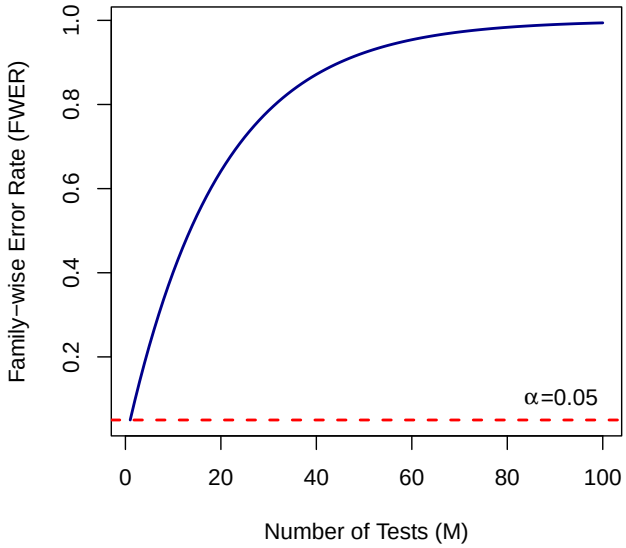


Figure 3.2: Family-Wise Error Rate Versus the Number of Hypothesis Tests, M.

EXAMPLE 3.3.5

Suppose we test $M = 4$ hypotheses, and the resulting p-values are $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$.

The Bonferroni Correction

- This is the simplest method.
- Reject $\mathbf{H}_{0,k}$ if

$$p_k \leq \frac{\alpha^\star}{M} \quad \text{for } k = 1, 2, \ldots, M$$

  So, we test all M hypotheses at a significance level of $\alpha^\star/M$.

- The procedure ensures FWER $\leq \alpha^\star$. From Boole's Inequality, we know that

$$\text{FWER} \leq M\left(\frac{\alpha^\star}{M}\right) = \alpha^\star$$

- If we assume independence, the Bonferroni-corrected FWER becomes

$$1 - \left(1 - \frac{\alpha^\star}{M}\right)^M$$

  Taking the limit of $M \to \infty$ yields,

$$\lim_{M\to\infty}\left[1 - \left(1 - \frac{\alpha^\star}{M}\right)^M\right] = 1 - e^{-\alpha^\star}$$

  which for typical values of $\alpha^\star$ in the range of $]0, 0.1]$ is approximately equal to $\alpha^\star$. For example, if $\alpha^\star = 0.1$, then the error is ≈ 0.005. The asymptotic error rate and line of equality can be seen in Figure 3.3.

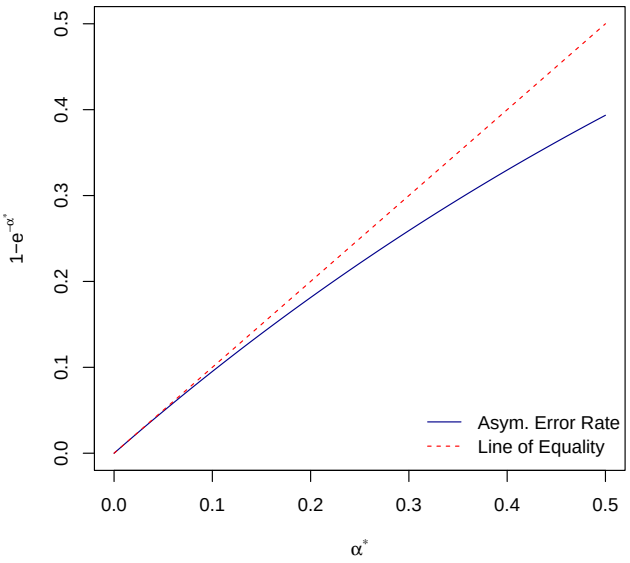


Figure 3.3: Illustration of the Bonferroni Correction for Asymptotically Large M.

EXAMPLE 3.3.6: Four-test Example — Bonferroni Correction

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$. Suppose that we wish to ensure FWER $\leq \alpha^\star = 0.05$.

Under the Bonferroni Correction, we compare each p-value to $\alpha^\star/M = 0.05/4 = 0.0125$. Only $p_3 < 0.0125$, and hence only $\mathbf{H}_{0,3}$ is rejected.

The Šidák Correction

- This approach exploits the FWER formula derived when we assumed the M tests were independent.
- Reject $\mathbf{H}_{0,k}$ if

$$p_k \leq 1 - (1-\alpha^\star)^{1/M} \quad \text{for } k = 1, 2, \ldots, M$$

REMARK 3.3.7

Where does the Šidák Correction come from?

$$\begin{aligned} \alpha^\star = \text{FWER} = 1-(1-\alpha)^M &\iff 1-\alpha^\star = (1-\alpha)^M \\ &\iff (1-\alpha^\star)^{1/M} = 1-\alpha \\ &\iff \alpha = 1-(1-\alpha^\star)^{1/M} \end{aligned}$$

- This is actually not much different from the Bonferroni correction since

$$\frac{\alpha^\star}{M} \approx 1-(1-\alpha^\star)^{1/M}$$

EXAMPLE 3.3.8: Bonferroni versus Šidák Correction

Let $\alpha^\star = 0.05$ and $M = 10$. Then, $\alpha^\star/M = 0.005$, and $1-(1-\alpha^\star)^{1/M} = 0.005116$.

EXAMPLE 3.3.9: Four-test Example — Šidák Correction

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$. Suppose that we wish to ensure FWER $\leq \alpha^\star = 0.05$.

Under the Šidák Correction, we have

$$1 - (1 - \alpha^\star)^{1/M} = 1 - (0.95)^{0.25} = 0.012741$$

Therefore, we only reject $\mathbf{H}_{0,3}$ since only $p_3 < 0.012741$.

Holm's "Step-Up" Procedure

- The Bonferroni and Šidák corrections methods are very strict for large M.
  - In these cases *most* null hypotheses will not be rejected.
  - If we're too strict, we basically stop rejecting null hypotheses thereby eliminating Type I Errors, but we increase the Type II Errors.
- Ideally we would have an approach that is less strict but still controls the FWER at some $\alpha^\star$.
- This is exactly what Holm's Procedure gives us!

1. Order the M p-values from smallest to largest:

   $$p_{(1)}, p_{(2)}, \cdots, p_{(M)}$$

   where $p_{(k)}$ is the k^{th} smallest p-value.
2. Starting from $k = 1$ and continuing incrementally, compare $p_{(k)}$ to $\alpha^\star/(M - k + 1)$. Determine $k^\star$, the smallest value of k such that

   $$p_{(k)} > \frac{\alpha^\star}{M - k + 1}$$

3. Reject the null hypotheses $\mathbf{H}_{0,(1)}, \ldots, \mathbf{H}_{0,(k^\star-1)}$ and do not reject $\mathbf{H}_{0,(k^\star)}, \ldots, \mathbf{H}_{0,(M)}$.

- What's really happening?

$$\begin{aligned} p_{(1)} &\text{ versus } \alpha^\star/M \\ p_{(2)} &\text{ versus } \alpha^\star/(M-1) \\ p_{(3)} &\text{ versus } \alpha^\star/(M-2) \\ &\vdots \\ p_{(M)} &\text{ versus } \alpha^\star \end{aligned}$$

We compare each p-value to a Bonferroni-Corrected significance level based on the number of comparisons that remain to be made at a particular "step."

THEOREM 3.3.10

Holm's procedure controls the family-wise error rate.

We need to show that FWER $= \mathbb{P}(V \geq 1) \leq \alpha^\star \in [0, 1]$ when using the Holm's procedure.

Let $p_{(1)}, p_{(2)}, \ldots, p_{(M)}$ be the ordered p-values and let $\mathbf{H}_{0,(1)}, \mathbf{H}_{0,(2)}, \ldots, \mathbf{H}_{0,(M)}$ be the corresponding null hypotheses.

Define $K_0 \subset \{1, 2, \ldots, M\}$ to be the subset of indices which correspond to true null hypotheses; that is, $\mathbf{H}_{0,k}$ is true for $k \in K_0$. We can visualize the sequential decisions made in Holm's Procedure as follows:

$$\overbrace{\underbrace{\mathbf{H}_{0,(1)} \cdots \mathbf{H}_{0,(h-1)}}_{\text{these are false } \mathbf{H}_0\text{'s}} \mathbf{H}_{0,(h)} \cdots \mathbf{H}_{0,(R)}}^{\text{these are rejected}} \mid \underbrace{\mathbf{H}_{0,(R+1)} \cdots \mathbf{H}_{0,(M)}}_{\text{these are not rejected}}$$

Let $\mathbf{H}_{0,(h)}$ be the first *true* $\mathbf{H}_0$ that was rejected. Since it was rejected by Holm's procedure, we know that

$$p_{(h)} \leq \frac{\alpha^\star}{M - h + 1}$$

Clearly we must have $h - 1 \leq M - M_0$ since $M - M_0$ is the total number of false $\mathbf{H}_0$'s and $h - 1$ is the number of false $\mathbf{H}_0$'s encountered by test h. And so,

$$M_0 \leq M - h + 1 \iff \frac{1}{M_0} \geq \frac{1}{M - h + 1} \iff \frac{\alpha^\star}{M_0} \geq \frac{\alpha^\star}{M - h + 1}$$

Thus, we must have $p_{(h)} \leq \alpha^\star/(M - h + 1) \leq \alpha^\star/M_0$. Therefore,

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}(\text{At least one Type I Error in } M \text{ tests}) \\
&= \mathbb{P}(\text{Reject at least one true } \mathbf{H}_0) \\
&= \mathbb{P}\left(\exists\, k \in K_0 \text{ such that } p_k \leq \frac{\alpha^\star}{M_0}\right) \\
&= \mathbb{P}\left(\bigcup_{k \in K_0} p_k \leq \frac{\alpha^\star}{M_0}\right) \\
&\leq \sum_{k \in K_0} \mathbb{P}\left(p_k \leq \frac{\alpha^\star}{M_0}\right) \\
&= \sum_{k \in K_0} \frac{\alpha^\star}{M_0} \\
&= M_0 \left(\frac{\alpha^\star}{M_0}\right) \\
&= \alpha^\star
\end{aligned}$$

where we used the fact that p-values for true null hypotheses follow a $\mathcal{U}[0, 1]$ distribution.

EXAMPLE 3.3.11: Four-test Example ($M = 4$) — Holm's Procedure

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$. Suppose that we wish to ensure FWER $\leq \alpha^\star = 0.05$.

$$\begin{aligned}
p_{(1)} &= p_3 = 0.008 \text{ versus } \alpha^\star/M = 0.05/4 = 0.0125 \\
p_{(2)} &= p_1 = 0.015 \text{ versus } \alpha^\star/(M-1) = 0.05/3 = 0.0167 \\
p_{(3)} &= p_4 = 0.026 \text{ versus } \alpha^\star/(M-2) = 0.05/2 = 0.025 \\
p_{(4)} &= p_2 = 0.029 \text{ versus } \alpha^\star/(M-3) = 0.05/1 = 0.05
\end{aligned}$$

We reject $\mathbf{H}_{0,(1)} = \mathbf{H}_{0,3}$ and $\mathbf{H}_{0,(2)} = \mathbf{H}_{0,1}$. We do not reject $\mathbf{H}_{0,(3)} = \mathbf{H}_{0,4}$ or $\mathbf{H}_{0,(4)} = \mathbf{H}_{0,2}$. Note that $k^\star = 3$.

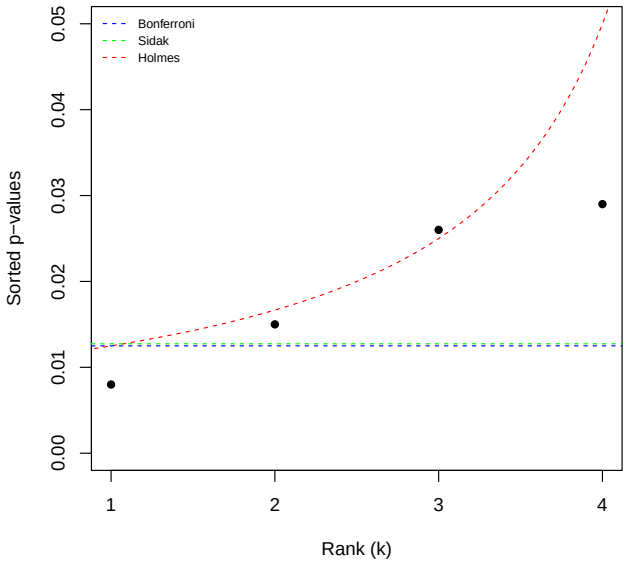


Figure 3.4: Significance Thresholds for Several Methods of Correction (1).

- The decision process for all three of these methods can be visualized by plotting the ordered p-values $p_{(k)}$ versus their ranks $k = 1, 2, \ldots, M$ and overlay the significance thresholds which can be seen in Figure 3.4.

- The Bonferroni correction is most strict, followed by the Šidák correction, then by Holm's procedure.

Adjusted p-values

- So far we have framed each of the correction procedures above as an adjustment to the significance threshold against which each p-value is compared.
- Alternatively (and equivalently) we could invert this process and frame the decision in terms of a comparison of *adjusted p-values* to $\alpha^\star$.
- This is more familiar (compare our p-values to some constant threshold $\alpha^\star$).
  - We just need to adjust our p-values first.
- The decisions made with the following adjusted p-values are identical to that achieved by comparing unadjusted p-values to the methods' adjusted significance thresholds.
  - Bonferroni: Reject $\mathbf{H}_{0,k}$ if $p_k^\star \leq \alpha^\star$ where

$$p_k^\star = Mp_k$$

EXAMPLE 3.3.12: Bonferroni's Adjusted p-values

In our four-test example, $p_1^\star = 0.06$, $p_2^\star = 0.116$, $p_3^\star = 0.032$, and $p_4^\star = 0.104$. Comparing to $\alpha^\star = 0.05$, we reject $\mathbf{H}_{0,3}$.

  - Šidák: Reject $\mathbf{H}_{0,k}$ if $p_k^\star \leq \alpha^\star$ where

$$p_k^\star = 1 - (1 - p_k)^M$$

EXAMPLE 3.3.13: Šidák's Adjusted p-values

In our four-test example, $p_1^\star = 0.0587$, $p_2^\star = 0.1111$, $p_3^\star = 0.0316$, and $p_4^\star = 0.1$. Comparing to $\alpha^\star = 0.05$, we reject $\mathbf{H}_{0,3}$.

– Holm: Reject $\mathbf{H}_{0,k}$ if $p_{(k)}^\star \leq \alpha^\star$ where

$$p_{(k)}^\star = \max_{j \leq k} \Big\{ p_{(j)}(M - j + 1) \Big\}$$

EXAMPLE 3.3.14: Holm's Adjusted p-values

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$.

k	$p_{(k)}$	$M-k+1$	$p_{(k)}(M-k+1)$	$p_{(k)}^\star = \max_{j \leq k}\Big\{p_{(j)}(M-j+1)\Big\}$
1	0.008	4	0.032	$\max\{0.032\} = 0.032 = p_{(1)}^\star$
2	0.015	3	0.045	$\max\{0.032, 0.045\} = 0.045 = p_{(2)}^\star$
3	0.026	2	0.052	$\max\{0.032, 0.045, 0.052\} = 0.052 = p_{(3)}^\star$
4	0.029	1	0.029	$\max\{0.032, 0.045, 0.052, 0.029\} = 0.052 = p_{(4)}^\star$

Thus, $p_1^\star = p_{(2)}^\star = 0.045$, $p_2^\star = p_{(4)}^\star = 0.052$, $p_3^\star = p_{(1)}^\star = 0.032$, and $p_4^\star = p_{(3)}^\star = 0.052$. Comparing to $\alpha^\star = 0.05$, we reject $\mathbf{H}_{0,1}$ and $\mathbf{H}_{0,3}$.

- Implemented in R with `p.adjust()`.

3.3.2 False Discovery Rate

- In the mid-1900s, Statisticians developed FWER methods with $M \approx 20$ comparisons in mind.
- In the era of Big Data, much larger values of M are typical.
- For larger values of M, traditional methods tend to be very conservative, and so FWER is perhaps not the best metric to control.
- More recently, emphasis has been placed on controlling the *rate* at which Type I Errors occur.

DEFINITION 3.3.15: False discovery proportion

The **false discovery proportion** (FDP) is

$$Q = \frac{V}{R}$$

Thus, Q is the proportion of all rejected null hypotheses that were rejected in error.

- In particular, interest lies in controlling the **false discovery rate** (FDR).

DEFINITION 3.3.16: False discovery rate

The **false discovery rate** is

$$\mathbb{E}[Q] = \mathbb{E}\left[\frac{V}{R}\right]$$

- Unlike the FWER, the FDR is adaptive in the sense that the number of Type I Errors V has different implications depending on the size of M. That is,

  - Two Type I Errors in 10 tests might be unacceptable.
  - Two Type I Errors in 100 tests might be okay.

- Methods that control the FDR are less strict than methods that control FWER.
  - More Type I Errors will occur with such methods, but this is viewed as acceptable when M is very large.

Benjamini-Hochberg Procedure

- The Benjamini-Hochberg (BH) procedure for controlling FDR is a sequentially rejective procedure much like Holm's procedure for controlling FWER.
- We summarize the BH procedure, which aims to ensure FDR $\leq \alpha^\star$:

1. Order the M p-values from smallest to largest:
$$p_{(1)}, p_{(2)}, \cdots, p_{(M)}$$
where $p_{(k)}$ is the k^{th} smallest p-value.
2. Starting from $k = 1$ and continuing incrementally, compare $p_{(k)}$ to $k\alpha^\star/M$. Determine $k^\star$ the largest value of k such that
$$p_{(k)} \leq \frac{k\alpha^\star}{M}$$
3. Reject the null hypotheses $\mathbf{H}_{0,(1)}, \ldots, \mathbf{H}_{0,(k^\star)}$ and do not reject $\mathbf{H}_{0,(k^\star+1)}, \ldots, \mathbf{H}_{0,(M)}$.

- The decision process associated with this procedure is best visualized with a plot of the ordered p-values $p_{(k)}$ versus their ranks $k = 1, 2, \ldots, M$ with the significance threshold overlaid which can be seen in Figure 3.5.
  - The BH significance threshold is the line with intercept 0 and slope $\alpha^\star/M$.

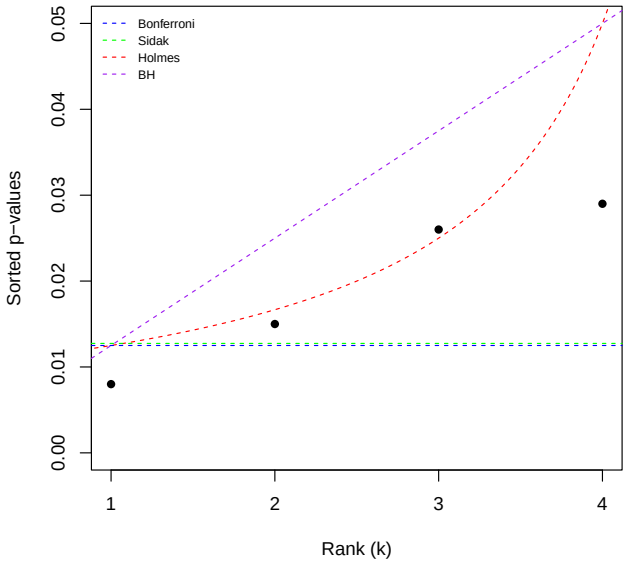


Figure 3.5: Significance Thresholds for Several Methods of Correction (2).

EXAMPLE 3.3.17: Four-test Example — Benjamini-Hochberg Procedure

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$. Suppose that we wish to ensure FDR $\leq \alpha^\star = 0.05$. Since all p-values fall below the purple line in Figure 3.5, we reject all four null hypotheses.

- This threshold is much less strict than any of the FWER-control thresholds, but this is the appeal of the approach.
- The proof that this procedure guarantees FDR $\leq \alpha^\star$ is outside the scope of this course, but the interested reader is referred to Benjamini and Hochberg (1995) and Storey et al. (2004).
- Like the FWER controlling methods we can define Benjamini-Hochberg-adjusted p-values and invert the decision framework by comparing the adjusted p-values to $\alpha^\star$.
  - Reject $\mathbf{H}_{0,(k)}$ if $p^\star_{(k)} \leq \alpha^\star$ where

$$p^\star_{(k)} = \min_{j \geq k} \left\{ \frac{M p_{(j)}}{j} \right\}$$

EXAMPLE 3.3.18: Benjamini-Hochberg Procedure's Adjusted p-values

Let $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$.

k	$p_{(k)}$	$Mp_{(k)}/k$	$p^\star_{(k)} = \min_{j \geq k}\{Mp_{(j)}/j\}$
1	0.008	0.032	$\min\{0.032, 0.030, 0.035, 0.029\} = 0.029 = p^\star_{(1)}$
2	0.015	0.030	$\min\{0.030, 0.035, 0.029\} = 0.029 = p^\star_{(2)}$
3	0.026	0.035	$\min\{0.035, 0.029\} = 0.029 = p^\star_{(3)}$
4	0.029	0.029	$\min\{0.029\} = 0.029 = p^\star_{(4)}$

Thus, $p^\star_1 = p^\star_{(2)} = 0.029$, $p^\star_2 = p^\star_{(4)} = 0.029$, $p^\star_3 = p^\star_{(1)} = 0.029$, and $p^\star_4 = p^\star_{(3)} = 0.029$. Comparing to $\alpha^\star = 0.05$, we reject all $\mathbf{H}_0$'s.

- [R Code] `Multiple_testing_example`

3.3.3 Sample Size Determination

- So what does all of this mean for power analyses and sample size calculations?
- There is a duality between significance level and power.
  - All else equal, reducing a test's significance level will increase the Type II Error rate and hence decrease power.
  - Play around with this interactive app to gain comfort with this notion.
- Thus, all the correction procedures discussed (which decrease the effective significance level) negatively impact power.
- In order to maintain power at some pre-specified level, we must compensate by increasing the sample size.
- Therefore, the more complicated your experiment (i.e., the more conditions it has), the larger your sample sizes will need to be.
  - Such modifications can be accounted for when selecting a sample size.

  - The significance level you use in your sample size calculations should be the adjusted one based on the correction method you use.
  - This is easier to do with *some* correction methods than others.

WEEK 5

Primer on Logistic Regression

- Linear regression is an effective method of modelling the relationship between a single response variable (Y), and one or more explanatory variables ($x_1, x_2, \dots, x_p$).
  - However, ordinary linear regression assumes that the response variable follows a normal distribution (i.e., $Y \sim \mathcal{N}(\mu, \sigma^2)$).
  - When the response variable is binary, this assumption is no longer valid.
- When we have a binary response, the Bernoulli distribution (i.e., $Y \sim \text{Binomial}(1, \pi)$) is a much more appropriate distributional assumption.
  - But ordinary linear regression is no longer appropriate.
  - Instead, we use **Logistic Regression**.
- In the context of a linear regression model, the expected response (given the values of the explanatory variables) is equated to the **linear predictor** $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$:

$$\mathbb{E}[Y \,|\, x_1, x_2, \dots, x_p] = \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- In the context of Logistic Regression we also want to relate the expected response to the linear predictor.
  - But now, $\mathbb{E}[Y] = \pi \in [0, 1]$.
  - And equating π and $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ does not make sense. In general, the linear predictor need not lie in $[0, 1]$.

$$\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad \text{not a good thing to do}$$

- Instead, we relate the linear predictor to $\mathbb{E}[Y] = \pi$ through a monotonic differentiable **link function** that maps $[0, 1] \to \mathbf{R}$.
  - Logistic Regression arises when this link function is the **logit** function:

$$\text{logit}(\pi) = \log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - Inverting this yields the expected response (given the values of the explanatory variables):

$$\mathbb{E}[Y \,|\, \widehat{x_1, x_2}, \dots, x_p] = \hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}} = \text{expit}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

- To interpret β_0, we set each explanatory variable to zero (i.e., $x_1 = x_2 = \cdots = x_p = 0$).
  - We see that β_0 is the **log-odds** that $Y = 1$ when $x_1 = x_2 = \cdots = x_p = 0$.

$$\log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0$$

  - Equivalently, e^{β_0} is the **odds** that the response would equal 1 when $x_1 = x_2 = \cdots = x_p = 0$. Exponentiating both sides yields

$$\frac{\pi}{1-\pi} = e^{\beta_0}$$

DEFINITION 3.3.19: Odds

The **odds** of an event A is:

$$\frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}$$

- The interpretation of β_j, for $j = 1, 2, \ldots, p$, is uncovered by considering the Logistic Regression equation for different values of x_j.
  - Let π_x be the value of π when $x_j = x$ and let π_{x+1} be the value of π when $x_j = x + 1$.

$$\begin{aligned}\log\left(\frac{\pi_{x+1}}{1-\pi_{x+1}}\right) - \log\left(\frac{\pi_x}{1-\pi_x}\right) &= (\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x+1) + \cdots + \beta_p x_p) \\ &\quad - (\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p) \\ &= \beta_j\end{aligned}$$

  - Thus:

$$\log\left(\frac{\pi_{x+1}}{1-\pi_{x+1}} \Bigg/ \frac{\pi_x}{1-\pi_x}\right) = \beta_j$$

    and so β_j is interpreted as a **log-odds ratio**, comparing the odds that $Y = 1$ when $x_j = x + 1$ versus $x_j = x$ (all else being equal).
  - Equivalently, e^{β_j} is interpreted as the **odds ratio**, comparing the odds that $Y = 1$ when $x_j = x+1$ versus $x_j = x$ (all else being equal). Exponentiating yields

$$\frac{\pi_{x+1}}{1-\pi_{x+1}} \Bigg/ \frac{\pi_x}{1-\pi_x} = e^{\beta_j}$$

- **Maximum likelihood estimation** is a method that is used to estimate parameters in Logistic Regression.
  - This means that the $\hat{\beta}$'s are maximum likelihood estimates, whose corresponding estimators have nice properties, such as:

$$\tilde{\beta} \dot{\sim} \mathcal{N}\left(\beta, \frac{1}{J(\beta)}\right)$$

    where $J(\beta)$ is the Fisher Information.
  - A consequence of this is that hypotheses of the form

$$\mathbf{H}_0\colon\ \beta_j = 0 \text{ versus } \mathbf{H}_\mathrm{A}\colon\ \beta_j \neq 0$$

    are done with *Z-tests* with test statistics given by

$$t = \frac{\hat{\beta}_j - 0}{\mathrm{Se}\left(\hat{\beta}_j\right)} \dot{\sim} \mathcal{N}(0,1)$$

  - In order to test hypotheses about several β's being simultaneously equal to zero, we use *likelihood ratio tests*.

Chapter 4

BLOCKING

- In the context of designed experiments we categorize factors as either:
  - *Design* factors: we manipulate these to quantify their impact on the response. They define the experimental conditions.
  - *Allowed-to-vary* factors: these are unknown, or known but uncontrollable factors that are not controlled in the experiment.
  - *Nuisance* factors: we control these to eliminate their effect on the response.
- But remember: in practice, context dictates whether a factor should be considered a design factor, a nuisance factor, or if it should be allowed to vary.

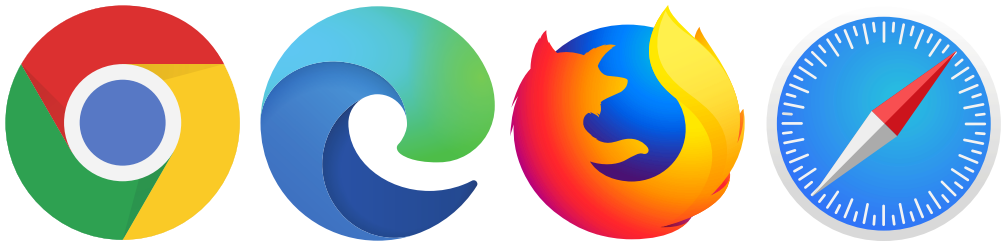


Figure 4.1: Four Levels of the *browser* Factor.

1. Usability testing involves studying the ease with which an individual uses a product or service for some intended purpose. Suppose investigators are performing a usability test to determine with which browser 70 to 80-year-old users find it easiest to look up the phone number of the nearest pharmacy. In this example, experimental units (70 to 80-year-olds) are randomly assigned to one of four browser conditions, and the investigators measure the time it takes to complete the task.
   - Browser is the design factor.
2. Suppose that Netflix is experimenting with server-side modifications to improve (reduce) the latency of Netflix.com. We hypothesize that the current infrastructure serves as a control condition and the modified infrastructure reduces median page load time. It is possible that a user's browser may also affect page load time, but this effect is not of interest to the investigators. To control for the potential impact of one's browser, Netflix initially experiments with only Firefox users.
   - Browser is the nuisance factor.
3. Suppose that Amazon.ca is experimenting with the width of their search bar. They hypothesize that a wider search bar will minimize the amount of mouse movement required to navigate to it, thereby minimizing the average time-to-query. The experimenters do not care which browser a customer uses and so this factor is uncontrolled and hence is *allowed-to-vary* during their experiment.

  - Browser is the allowed-to-vary factor.

- It's also important to understand the subtle distinction between nuisance factors and design factors in the context of a single experiment.
  - We control both factors in the experiment.
  - With a design factor we wish to quantify its influence on the response variable.
  - With a nuisance factor we do not care to quantify its effect, we wish only to *eliminate* it.

- We eliminate the effect of one or more nuisance factors with **blocking**.
  - To eliminate the effect of a nuisance factor, it cannot be allowed to vary on its own.
  - Blocking fixes the nuisance factor at one or more levels (**blocks**).
  - By holding a nuisance factor fixed, it cannot vary and hence cannot influence the response.
    * This is how Netflix handled the nuisance factor "browser" in Example 2.

4.1 Randomized Complete Block Designs

- The randomized complete block design (RCBD) is a simple experimental design that may be applied when we wish to investigate:
  - A single design factor; e.g., m levels, m conditions, while controlling for a single nuisance factor; e.g., b levels, b blocks.

- In a RCBD, we carry out each of the experimental conditions in every one of the blocks.
  - m conditions are happening inside each of the b blocks.

- The *observed* data in such an experiment is y_{ijk}.
  - Response observation for unit $i = 1, 2, \dots, n_{jk}$ in condition $j = 1, 2, \dots, m$ within block $k = 1, 2, \dots, b$.

- We assume that there are n_{jk} units in $(\text{condition}, \text{block}) = (j, k)$ and thus an overall total of $N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ units.
  - If $n_{jk} = n$ for all (j, k), we call the design "balanced."

- We tabulate the response data of this form below:

Table 4.1: Response Observations in a Randomized Complete Block Design

		Block				
		1	2	$\cdots$	b	
Condition	1	$\{y_{i11}\}_{i=1}^{n_{11}}$	$\{y_{i12}\}_{i=1}^{n_{12}}$	$\cdots$	$\{y_{i1b}\}_{i=1}^{n_{1b}}$	$\bar{y}_{\bullet 1 \bullet}$
	2	$\{y_{i21}\}_{i=1}^{n_{21}}$	$\{y_{i22}\}_{i=1}^{n_{22}}$	$\cdots$	$\{y_{i2b}\}_{i=1}^{n_{2b}}$	$\bar{y}_{\bullet 2 \bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	m	$\{y_{im1}\}_{i=1}^{n_{m1}}$	$\{y_{im2}\}_{i=1}^{n_{m2}}$	$\cdots$	$\{y_{imb}\}_{i=1}^{n_{mb}}$	$\bar{y}_{\bullet m \bullet}$
		$\bar{y}_{\bullet\bullet 1}$	$\bar{y}_{\bullet\bullet 2}$	$\cdots$	$\bar{y}_{\bullet\bullet b}$	$\bar{y}_{\bullet\bullet\bullet}$

  - Block-specific average responses: $\bar{y}_{\bullet\bullet 1}, \bar{y}_{\bullet\bullet 2}, \dots, \bar{y}_{\bullet\bullet b}$.
  - Overall average response: $\bar{y}_{\bullet\bullet\bullet}$.
  - Condition-specific average responses: $\bar{y}_{\bullet 1 \bullet}, \bar{y}_{\bullet 2 \bullet}, \dots, \bar{y}_{\bullet m \bullet}$.

- We calculate the row, column, and overall means as follows:

$$\bar{y}_{\bullet j\bullet} = \frac{1}{n_{j+}} \sum_{k=1}^{b} n_{jk} \bar{y}_{\bullet jk} \quad \text{where } n_{j+} = \sum_{k=1}^{b} n_{jk}$$

$$\bar{y}_{\bullet\bullet k} = \frac{1}{n_{+k}} \sum_{j=1}^{m} n_{jk} \bar{y}_{\bullet jk} \quad \text{where } n_{+k} = \sum_{j=1}^{m} n_{jk}$$

$$\bar{y}_{\bullet\bullet\bullet} = \frac{1}{N} \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk} \bar{y}_{\bullet jk} = \frac{1}{N} \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} y_{ijk}$$

  where $\bar{y}_{\bullet jk}$ is the average response value in $(\text{condition}, \text{block}) = (j, k)$ cell, also

$$\bar{y}_{\bullet jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}$$

- Simple summaries such as these provide a crude assessment of whether the condition-to-condition and block-to-block variation is large.
  - If the condition-specific averages are very different, this suggests that the design factor influences the response.
  - If the block-specific averages are very different, this suggests that the nuisance factor influences the response, and that blocking was appropriate.
- The primary analysis goal in a RCBD is to determine whether the expected response differs significantly from one condition to another.
  - And if so, to identify the optimal condition, while controlling for the potential effect of the nuisance factor.
- We've previously done this with gatekeeper tests of the form:

$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } \mathbf{H}_\text{A}\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k$$

- We do the same thing here, while accounting for the nuisance factor, with *appropriately defined* linear (continuous response) or logistic (binary response) regression models which contain:
  - An intercept.
  - $m-1$ indicator variables for the design factor's levels.
  - $b-1$ indicator variables for the nuisance factor's levels.
- We write the linear predictor as:

$$\alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ij} \qquad (\star)$$

  - $x_{ij} = 1$ if unit i is in condition $j = 1, 2, \ldots, m-1$, and zero otherwise.
  - $z_{ik} = 1$ if unit i is in block $k = 1, 2, \ldots, b-1$, and zero otherwise.
  - The β's jointly quantify the effect of the design factor.
  - The γ's jointly quantify the effect of the nuisance factor.
- Two relevant hypotheses are:
  (1) $\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$ versus $\mathbf{H}_\text{A}\text{: } \beta_j \neq 0$ for some j.
    - If we don't reject $\mathbf{H}_0$, this suggests the x's don't need to be in the model and hence the design factor doesn't significantly influence the response.
  (2) $\mathbf{H}_0\text{: } \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0$ versus $\mathbf{H}_\text{A}\text{: } \gamma_k \neq 0$ for some k.

  - If we don't reject $\mathbf{H}_0$, this suggests the z's don't need to be in the model and hence the nuisance factor doesn't significantly influence the response. Therefore, blocking isn't necessary.

- We test these hypotheses by comparing a *full* model and *reduced* models where the *full* model is a model with a linear predictor given by ($\star$), and a *reduced* model is a model with a linear predictor that arises when $\mathbf{H}_0$ is true.
  - We try to determine whether the full model fits the data significantly better than the reduced one.

4.1.1 RCBD to Compare Means

- Here, we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factor):

$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_m \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$

  where μ_j is the expected response in condition $j = 1, 2, \ldots, m$.

- We do this by testing:

$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j$$

  with an ANOVA in the context of the following linear regression model.

$$Y_i = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik} + \varepsilon_i$$

  where Y_i is the response observation for unit $i = 1, 2, \ldots, N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is a random error term.

- The relevant ANOVA table is Table 4.2.

Table 4.2: Two-Way ANOVA Table Associated With a Randomized Complete Block Design

Source	SS	d.f.	MS	Test Statistic
Condition	SS_C	$m-1$	$\text{MS}_\text{C} = \text{SS}_\text{C}/(m-1)$	$t_\text{C} = \text{MS}_\text{C}/\text{MS}_\text{E}$
Block	SS_B	$b-1$	$\text{MS}_\text{B} = \text{SS}_\text{B}/(b-1)$	$t_\text{B} = \text{MS}_\text{B}/\text{MS}_\text{E}$
Error	SS_E	$N-m-b+1$	$\text{MS}_\text{E} = \text{SS}_\text{E}/(N-m-b+1)$	
Total	SS_T	$N-1$		

- The sums of squares given in Table 4.2 are:
  - Total sum of squares (quantifies overall response variation):

$$\text{SS}_\text{T} = \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = \text{SS}_\text{C} + \text{SS}_\text{B} + \text{SS}_\text{E}$$

  - Condition sum of squares (quantifies condition-to-condition response variation):

$$\text{SS}_\text{C} = \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$$

  - Block sum of squares (quantifies block-to-block response variation):

$$\text{SS}_\text{B} = \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2$$

- Error sum of squares (quantifies residual response variation not accounted for by conditions of blocks):

$$\text{SS}_\text{E} = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}}(y_{ijk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2$$

- So how do we use this table?
  - We test: $\mathbf{H}_0$: $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$ using $t_\text{C} = \text{MS}_\text{C}/\text{MS}_\text{E}$.
    - p-value $= \mathbb{P}(T \geq t_\text{C})$ where $T \sim F(m-1, N-m-b+1)$.
  - We test: $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0$ using $t_\text{B} = \text{MS}_\text{B}/\text{MS}_\text{E}$.
    - p-value $= \mathbb{P}(T \geq t_\text{B})$ where $T \sim F(b-1, N-m-b+1)$.

4.1.2 Example: Promotions at The Gap

EXAMPLE 4.1.1: Promotions at The Gap

The Gap has three versions of an online weekday promotion that a customer sees when they go to gapcanada.ca:

- Version 1: 50% discount on one item.
- Version 2: 20% discount on your entire order.
- Version 3: Spend $50 and get a $10 gift card.

Interest lies in determining whether there is a difference in the average purchase total (i.e, the average dollar value of a customer's purchase) between promotion versions. However, the amount of money one spends may also be influenced by the nuisance factor, day of week. As such, we ran a randomized complete block design with $m = 3$ experimental conditions (corresponding to the three promotions) and $b = 5$ blocks (corresponding to the day of the week). Here $n_{jk} = 50$ for all (j, k), and so the design was "balanced." For each visitor to gapcanada.ca, their purchase total (in dollars) was recorded. The regression model fit to these response observations is:

$$Y_i = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \gamma_4 z_{i4} + \varepsilon_i$$

where x_{i2} and x_{i3} are condition indicators for promotions 2 and 3 (promotion 1 is the baseline) and $z_{i1}, \ldots, z_{i4}$ are block indicators for Monday-Thursday (Friday is the baseline). The ANOVA Table for this experiment is Table 4.3.

Table 4.3: The Gap RCBD ANOVA Table

Source	SS	d.f.	MS	Test Statistic
Condition	49618.34	2	24809.17	$t_\text{C} = 2165.39$
Block	19258.30	4	4814.58	$t_\text{B} = 420.22$
Error	8512.67	743	11.46	
Total	77389.32	749		

- $\mathbf{H}_0$: $\beta_2 = \beta_3 = 0$ tells us whether the design factor is significant.
  - p-value $= \mathbb{P}(T \geq t_\text{C}) = \mathbb{P}(T \geq 2165.39) = 1.101 \times 10^{-310}$ where $T \sim F(2, 743)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the expected response is not the same in all conditions.
- $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$ tells us whether the nuisance factor is significant.
  - p-value $= \mathbb{P}(T \geq t_\text{B}) = \mathbb{P}(T \geq 420.22) = 4.345 \times 10^{-189}$ where $T \sim F(4, 743)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that blocking was appropriate.

[R Code] `Comparing_means_within_blocks`

4.1.3 RCBD to Compare Proportions

- Here we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factor):
$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m \text{ versus } \mathbf{H}_\text{A}\text{: } \pi_j \neq \pi_k \text{ for some } j \neq k$$
where π_j is the expected response in condition $j = 1, 2, \ldots, m$.

- We do this by testing:
$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j$$
with a likelihood ratio test (LRT) in the context of the following logistic regression model:
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik}$$
where Y_i is the response observation for unit $i = 1, 2, \ldots, N = \sum_{k=1}^{b}\sum_{j=1}^{m} n_{jk}$, and $\pi_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$.
  - The likelihood ratio test compares the full model to the one without the x's.

- Similarly, we test:
$$\mathbf{H}_0\text{: } \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} \text{ versus } \mathbf{H}_\text{A}\text{: } \gamma_k \neq 0 \text{ for some } k$$
with a LRT that compares the full model to the reduced one without the z's.
$\mathbf{H}_0$: Reduced model fits as well versus $\mathbf{H}_\text{A}$: Full model fits better than reduced model.

- The observed test statistic for both of these tests is:
$$\begin{aligned} t &= 2\log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2\Big[\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}\Big] \end{aligned}$$
which follows an approximate $\chi^2(\ell)$, if $\mathbf{H}_0$ is true, where
$$\ell = (\#\text{ parameters in full model}) - (\#\text{ parameters in reduced model})$$
  - p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(\ell)$.

4.1.4 Example: Enterprise Banner Ads

EXAMPLE 4.1.2: Enterprise Banner Ads

Enterprise is experimenting with $m = 3$ banner ads as a mechanism to drive traffic to their website. Since there are known regional differences in consumer preferences in the US, they wish to control for the nuisance factor "region" with $b = 4$ blocks corresponding to the four major US geographic regions: Northeast (NE), Northwest (NW), Southeast (SE), and Southwest (SW). We randomize a total of $n_{jk} = 5000$ for all (j, k) people to each ad condition in each region.

Interest lies in determining whether the different ads perform similarly with respect to click- through-rate (CTR) — and we wish to determine which one maximizes CTR — but we want to control for the effect of region. We do so with the following logistic regression model:
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3}$$
where x_{i2} and x_{i3} are condition indicators for ads 2 and 3 (ad 1 is the baseline), and z_{i1}, z_{i2}, z_{i3} are block indicators for NW, SE, SW regions (NE is the baseline).

- $\mathbf{H}_0$: $\beta_2 = \beta_3 = 0$.

- p-value $= \mathbb{P}(T \geq t_C) = \mathbb{P}(T \geq 249.924) = 5.367 \times 10^{-55}$ where $T \sim \chi^2(2)$.
- Therefore, we reject $\mathbf{H}_0$ and conclude that the design factor is significant and the CTR is not the same in every condition.

- $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = \gamma_3 = 0$.
  - p-value $= \mathbb{P}(T \geq t_B) = \mathbb{P}(T \geq 139.824) = 4.126 \times 10^{-30}$ where $T \sim \chi^2(3)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the nuisance factor is significant and therefore blocking was a good thing to do.

[R Code] `Comparing_proportions_within_blocks`

WEEK 6

4.2 Balanced Incomplete Block Designs

- Randomized Complete Block Designs (RCBD) were a tool for the exploration of *one* design factor (m levels) while controlling for the effect of *one* nuisance factor (b blocks).
  - In a RCBD, we carry out *every* experimental condition inside *every* block.
  - But sometimes, due to practical constraints, this is not possible.
- The Gap is experimenting with $m = 3$ promotional offers:
  - Version 1: 50% discount on one item.
  - Version 2: 20% discount on your entire order.
  - Version 3: Spend \$50 and get a \$10 gift card.
- Experimenters would like to control for a possible day-of-week effect (block by day).
  - Naturally, one might consider a RCBD; that is, suppose we observe data in *every* block-condition combination as seen in Table 4.4.

Table 4.4: Complete Block Design

		Day					
		1	2	3	4	5	6
Promotion	1	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓
	3	✓	✓	✓	✓	✓	✓

- But The Gap may only offer two of the three promotions in a single day.
  - So we must consider an **incomplete** block design; that is, suppose we observe data in only *some* block-condition combinations.
- We refer to Table 4.5 as a **balanced incomplete block design** (BIBD).

REMARK 4.2.1: Notation

- m: number of experimental conditions. In our previous example, $m = 3$.
- b: number of blocks. In our previous example, $b = 6$.

Table 4.5: Incomplete Block Design

		Day					
		1	2	3	4	5	6
Promotion	1	✓	✓	×	✓	✓	×
	2	✓	×	✓	✓	×	✓
	3	×	✓	✓	×	✓	✓

  - $m^\star$: number of experimental conditions that can be run in each block. Also known as "block size." In our previous example, $m^\star = 2$.
    - RCBD: $m^\star = m$.
    - BIBD: $m^\star < m$.
  - r: number of blocks in which each condition appears. In our previous example, $r = 4$.
  - λ: number of blocks that *any* pair of conditions appear in together. In our previous example, $\lambda = 2$.

- The BIBD is "balanced" in the sense that:
  - The number of conditions in each block is the same for every block ($m^\star$).
  - The number of blocks each condition appears in is the same for every condition (r).
  - The number of blocks each pair of conditions appear in together is the same for every possible condition pairing (λ).
- This balance allows for the comparison of a metric of interest across m conditions while still accounting for a nuisance factor with b levels
  - But despite this balance, the "incompleteness" requires some sacrifice.

4.2.1 General Comments on the Design of a BIBD

- Not just any haphazard combination of $(m, b, m^\star, r, \lambda)$ values will yield a BIBD.
- Great care must go into planning a BIBD to ensure all forms of balance.
- A variety of restrictions must be met:
  - Consequences of "incompleteness:"
    - $m^\star < m$.
    - $r < b$.
    - $\lambda < r$.
  - Number of block-condition combinations for which we observe data:
    - $mr = bm^\star$
  - For condition X (doesn't matter which condition this is), $r(m^\star - 1) = \lambda(m - 1)$ is the total number of conditions that condition X appears within the same blocks.
    - Condition X appears in r blocks, and in each, it's grouped with $m^\star - 1$ *other* conditions.
    - We pair each of the other $m - 1$ conditions with condition X λ times.
- We use these restrictions as follows:
  1. Specify m, $m^\star$, and λ.

2. Calculate $r = \lambda(m-1)/(m^\star - 1)$, noting that it must be an integer.
3. Calculate $b = mr/m^\star$, noting that it must be an integer.

EXAMPLE 4.2.2

Let $m = 3$, $m^\star = 2$, and $\lambda = 1$. We have $r = (1)(2)/(1) = 2$, and $b = (3)(2)/2 = 3$. See Table 4.6.

EXAMPLE 4.2.3: Pizza Table

Let $m = 3$, $m^\star = 2$, and $\lambda = 2$. We have $r = (2)(2)/(1) = 4$, and $b = (3)(4)/2 = 6$.

EXAMPLE 4.2.4

Let $m = 3$, $m^\star = 2$, and $\lambda = 3$. We have $r = (3)(2)/(1) = 6$, and $b = (3)(6)/2 = 9$. See Table 4.7.

- We select the design based on a trade-off between larger λ values and smaller b values.
  - Larger λ provides more information about pairwise comparisons.
  - Smaller b corresponds to fewer blocks and hence a smaller experiment.

Table 4.6: Incomplete Block Design

		Block		
		1	2	3
Condition	1	✓	✓	×
	2	✓	×	✓
	3	×	✓	✓

Table 4.7: Incomplete Block Design

		Block								
		1	2	3	4	5	6	7	8	9
Condition	1	✓	✓	✓	✓	✓	✓	×	×	×
	2	✓	✓	✓	×	×	×	✓	✓	✓
	3	×	×	×	✓	✓	✓	✓	✓	✓

4.2.2 General Comments on the Analysis of a BIBD

- Primary analysis goal:
  - Determine whether there exist significant differences among the expected response values from one experimental condition to another.
- In a RCBD, we do this by comparing the condition-specific means $\bar{y}_{\bullet j \bullet}$ to the overall mean $\bar{y}_{\bullet\bullet\bullet}$. This isn't fair in a BIBD because $\bar{y}_{\bullet\bullet\bullet}$ is calculated from data from blocks that condition j didn't appear in.
- In a BIBD, due to incompleteness, we compare $\bar{y}_{\bullet j \bullet}$ with the average response from the blocks that condition j appeared in:

$$\frac{\sum_{k\in\mathcal{B}_j}\sum_{i=1}^{n_{jk}} y_{ijk}}{\sum_{k\in\mathcal{B}_j} n_{jk}}$$

where $\mathcal{B}_j \subset \{1, 2, \ldots, B\}$ is the subset of indices indicating which blocks condition j appeared in.

- In general, the analysis of BIBDs involves an *adjustment* of this form when evaluating the effect of the design factor.

4.3 Latin Square Designs

- Until now, we have discussed experimental designs that employ blocking to control for *one* nuisance factor:
  - If we want to control for *two* nuisance factors, we should use a **Latin square design**.
  - If we want to control for *three* nuisance factors, we should use a **Graeco-Latin square design**.
  - If we want to control for *four* nuisance factors, we should use a **Hyper-Graeco-Latin square design**.
- A Latin square of order p is a $p \times p$ grid containing p unique symbols.
  - Each of these symbols occurs exactly once in each column.
  - Each of these symbols occurs exactly once in each row.
  - These "symbols" are typically denoted by Latin letters.

Table 4.8: 3×3, 4×4, and 5×5 Latin Square Examples

A	C	B
C	B	A
B	A	C

A	B	C	D
C	D	A	B
B	C	D	A
D	A	B	C

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

- A Sudoku puzzle is a special example of a 9×9 Latin square.
- We exploit this combinatorial structure to help us design experiments that facilitate blocking by two nuisance factors.
  - We arbitrarily associate the p rows with the levels of the first nuisance factor.
  - We arbitrarily associate the p columns with the levels of the second nuisance factor.
  - We arbitrarily associate the p Latin letters with the levels of the design factor.
- We present an example with $p = 4$ in Table 4.9.

Table 4.9: 4×4 Latin Square Design

		NF 2			
		1	2	3	4
NF 1	1	A	B	C	D
	2	D	A	B	C
	3	C	D	A	B
	4	B	C	D	A

  - **Limitation of LSD's**: we need to experiment with *all* of these factors at p levels.
  - $(3, 2)$ element represents the block where NF 1 is at level 3, NF 2 is at level 2, and DF is at level D.
- Each cell in this table represents a "block" in which we fix the nuisance factors' levels, and the Latin letter indicates the execution of an experimental condition.

- Rows, columns, and letters are all orthogonal, allowing us to separately estimate the effects of the design factor and each of the two nuisance factors.
- We may informally summarize these effects with the overall average and level-specific averages of the response variables.
  - Average response in a particular condition:

$$\bar{y}_{\bullet j \bullet \bullet} = \frac{1}{np} \sum_{(j,k,\ell) \in \mathcal{S}_j} \sum_{i=1}^{n} y_{ijk\ell}$$

  - Average response in a given row:

$$\bar{y}_{\bullet \bullet k \bullet} = \frac{1}{np} \sum_{(j,k,\ell) \in \mathcal{S}_k} \sum_{i=1}^{n} y_{ijk\ell}$$

  - Average response in a given column:

$$\bar{y}_{\bullet \bullet \bullet \ell} = \frac{1}{np} \sum_{(j,k,\ell) \in \mathcal{S}_\ell} \sum_{i=1}^{n} y_{ijk\ell}$$

  - Overall average:

$$\bar{y}_{\bullet \bullet \bullet \bullet} = \frac{1}{N} \sum_{(j,k,\ell) \in \mathcal{S}} \sum_{i=1}^{n} y_{ijk\ell}$$

    * $y_{ijk\ell}$ is the response observation for unit $i = 1, 2, \ldots, n$ in block (k, ℓ) and hence condition j.
    * $j, k, \ell = 1, 2, \ldots, p$.
    * n is the number of units in each block.
    * $N = np^2$.
- A comment about notation:
  - Each block contains just one condition, so each pair (k, ℓ) uniquely determines the value of j.
  - Consequently, there exist just p^2 tuples (j, k, ℓ).
  - Denote them by the set $\mathcal{S}$.
  - From Table 4.9, we have:

$(1,1,1)$	$(2,1,2)$	$(3,1,3)$	$(4,1,4)$
$(4,2,1)$	$(1,2,2)$	$(2,2,3)$	$(3,2,4)$
$(3,3,1)$	$(4,3,2)$	$(1,3,3)$	$(2,3,4)$
$(2,4,1)$	$(3,4,2)$	$(4,4,3)$	$(1,4,4)$

    * $S_{j=1} = \{(1,1,1), (1,2,2), (1,3,3), (1,4,4)\}$.
    * We also define:
      · $\mathcal{S}_j \subset \mathcal{S}$: all tuples for which the design factor is level j.
      · $\mathcal{S}_k \subset \mathcal{S}$: all tuples for which the nuisance factor 1's is level k.
      · $\mathcal{S}_\ell \subset \mathcal{S}$: all tuples for which the nuisance factor 2's is level ℓ.
- The primary analysis goal in a Latin square design is to determine whether the expected response differs significantly from one condition to another.
  - If so, to identify the optimal condition while controlling for the potential effect of the nuisance factors.

- We've previously done this with gatekeeper tests of the form:
$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_p \text{ versus } \mathbf{H}_\text{A}\text{: } \theta_j \neq \theta_{j'} \text{ for some } j \neq j'.$$
- We do the same thing here, while accounting for the nuisance factors, with *appropriately defined* linear or logistic regression models which contain:
  - An intercept.
  - $p-1$ indicator variables for the design factor's levels.
  - $p-1$ indicator variables for nuisance factor 1's levels.
  - $p-1$ indicator variables for nuisance factor 2's levels.
- We write the linear predictor as:
$$\alpha + \sum_{j=1}^{p-1} \beta_j x_{ij} + \sum_{k=1}^{p-1} \gamma_j z_{ik} + \sum_{\ell=1}^{p-1} \delta_\ell w_{i\ell}$$
  - $x_{ij} = 1$ if unit i is in condition $j = 1, 2, \ldots, p-1$ (zero otherwise).
  - $z_{ik} = 1$ if unit i is in a block for which nuisance factor 1 is at level $k = 1, 2, \ldots, p-1$ (zero otherwise).
  - $w_{i\ell} = 1$ if unit i is in a block for which nuisance factor 2 is at level $\ell = 1, 2, \ldots, p-1$ (zero otherwise).
  - The β's jointly quantify the effect of the design factor.
  - The γ's jointly quantify the effect of nuisance factor 1.
  - The δ's jointly quantify the effect of nuisance factor 2.
- Three relevant hypotheses are:
$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j.$$
  - Provides insight into whether DF is important.
$$\mathbf{H}_0\text{: } \gamma_1 = \beta_2 = \cdots = \gamma_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \gamma_k \neq 0 \text{ for some } k.$$
  - Provides insight into whether NF 1 is important.
$$\mathbf{H}_0\text{: } \delta_1 = \delta_2 = \cdots = \delta_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \delta_\ell \neq 0 \text{ for some } \ell.$$
  - Provides insight into whether NF 2 is important.
- We test these hypotheses by comparing a *full* (linear predictor) and *reduced* ($\mathbf{H}_0$ is true) model.
  - We try to determine whether the full model fits the data significantly better than the reduced one.

4.3.1 Latin Squares to Compare Means

- Here we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factors):
$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_p \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_{j'} \text{ for some } j \neq j'$$
where μ_j is the expected response in condition $j = 1, 2, \ldots, p$.
- We do this by testing:
$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j$$
with an ANOVA in the context of the following linear regression model:
$$Y_i = \alpha + \sum_{j=1}^{p-1} \beta_j x_{ij} + \sum_{k=1}^{p-1} \gamma_k z_{ik} + \sum_{\ell=1}^{p-1} \delta_\ell w_{i\ell} + \varepsilon_i \quad \text{(Full Model)}$$
  - Y_i is the response observation for unit $i = 1, 2, \ldots, N = np^2$.

- $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is the random error term.

- The relevant sums of squares are:
  - The total sum of squares, which quantifies overall variation in response values:
$$\text{SS}_\text{T} = \sum_{(j,k,\ell)\in\mathcal{S}} \sum_{i=1}^{n} (y_{ijk\ell} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 = \text{SS}_\text{C} + \text{SS}_{\text{B}_1} + \text{SS}_{\text{B}_2} + \text{SS}_\text{E}$$
  - The condition sum of squares, which quantifies variability in the response from one condition to another:
$$\text{SS}_\text{C} = \sum_{(j,k,\ell)\in\mathcal{S}} \sum_{i=1}^{n} (\bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 = np \sum_{j=1}^{p} (\bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2$$
  - The first block sum of squares, which quantifies variability in the response from one level of nuisance factor 1 to another:
$$\text{SS}_{\text{B}_1} = \sum_{(j,k,\ell)\in\mathcal{S}} \sum_{i=1}^{n} (\bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 = np \sum_{k=1}^{p} (\bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2$$
  - The second block sum of squares, which quantifies variability in the response from one level of nuisance factor 2 to another:
$$\text{SS}_{\text{B}_2} = \sum_{(j,k,\ell)\in\mathcal{S}} \sum_{i=1}^{n} (\bar{y}_{\bullet\bullet\bullet\ell} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 = np \sum_{\ell=1}^{p} (\bar{y}_{\bullet\bullet\bullet\ell} - \bar{y}_{\bullet\bullet\bullet\bullet})^2$$
  - The error sum of squares, which quantifies variability in the response that was not explained by conditions or blocks (i.e., the design and nuisance factors):
$$\text{SS}_\text{E} = \sum_{(j,k,\ell)\in\mathcal{S}} \sum_{i=1}^{n} (y_{ijk\ell} - \bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\ell} - 2\bar{y}_{\bullet\bullet\bullet\bullet})^2$$

- We show the corresponding ANOVA table in Table 4.10.

Table 4.10: Three-Way ANOVA Table Associated with a Latin Square Design

Source	SS	d.f.	MS	Test Statistic
Design Factor	SS_C	$p-1$	$\text{MS}_\text{C} = \text{SS}_\text{C}/(p-1)$	$t_\text{C} = \text{MS}_\text{C}/\text{MS}_\text{E}$
Nuisance Factor 1	SS_{B_1}	$p-1$	$\text{MS}_{\text{B}_1} = \text{SS}_{\text{B}_1}/(p-1)$	$t_{\text{B}_1} = \text{MS}_{\text{B}_1}/\text{MS}_\text{E}$
Nuisance Factor 2	SS_{B_2}	$p-1$	$\text{MS}_{\text{B}_2} = \text{SS}_{\text{B}_2}/(p-1)$	$t_{\text{B}_2} = \text{MS}_{\text{B}_2}/\text{MS}_\text{E}$
Error	SS_E	$N-3p+2$	$\text{MS}_\text{E} = \text{SS}_\text{E}/(N-3p+2)$	
Total	SS_T	$N-1$		

- So, how do we use this table?
  - We test $\mathbf{H}_0$: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ using $t_\text{C} = \text{MS}_\text{C}/\text{MS}_\text{E}$.
    - p-value: $\mathbb{P}(T \geq t_\text{C})$ where $T \sim F(p-1, N-3p+2)$.
  - We test: $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = \cdots = \gamma_{p-1} = 0$ using $t_{\text{B}_1} = \text{MS}_{\text{B}_1}/\text{MS}_\text{E}$.
    - p-value: $\mathbb{P}\left(T \geq t_{\text{B}_1}\right)$ where $T \sim F(p-1, N-3p+2)$.
  - We test: $\mathbf{H}_0$: $\delta_1 = \delta_2 = \cdots = \delta_{p-1} = 0$ using $t_{\text{B}_2} = \text{MS}_{\text{B}_2}/\text{MS}_\text{E}$.
    - p-value: $\mathbb{P}\left(T \geq t_{\text{B}_2}\right)$ where $T \sim F(p-1, N-3p+2)$.

4.3.2 Example: Netflix Latency

Consider the latency experiment described at the beginning of the chapter in which Netflix is experimenting with server-side modifications to improve (reduce) the latency of netflix.com. In particular, they have four different experimental conditions (A, B, C, D) that are intended to reduce average latency (in milliseconds). Two nuisance factors that may also influence latency are browser (Chrome, Microsoft Edge, Firefox, Safari), and time of day ([00:01,06:00], [06:01,12:00], [12:01,18:00], [18:01,00:00]). The design of the experiment is the 4×4 Latin square shown in Table 4.11. In order to determine whether the expected latency in each condition differs significantly, we randomize $n = 500$ users to each of the $p^2 = 16$ blocks.

Table 4.11: 4×4 Latin Square Design for the Netflix Experiment

		Browser			
		Chrome	Edge	Firefox	Safari
Time	[00:01,06:00]	A	B	C	D
	[06:01,12:00]	D	A	B	C
	[12:01,18:00]	C	D	A	B
	[18:01,00:00]	B	C	D	A

We analyze the data with the following linear regression model:

$$Y_i = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \delta_2 w_{i2} + \delta_3 w_{i3} + \delta_4 w_{i4} + \varepsilon_i$$

- x_{i2}, x_{i3}, x_{i4} are indicators for conditions B, C, D, where A is the baseline.
- z_{i1}, z_{i2}, z_{i3} are browser indicators for Microsoft Edge, Firefox, Safari, where Chrome is the baseline.
- w_{i2}, w_{i3}, w_{i4} are time indicators for time periods:
  [06:01,12:00], [12:01,18:00], [18:01,00:00], where [00:01,06:00] is the baseline

The ANOVA table associated with this model is Table 4.12.

Table 4.12: Netflix Latin Square ANOVA Table

Source	SS	d.f.	MS	Test Statistic
Condition	203903.38	3	67967.79	679.14
Browser	32.95	3	10.98	0.1097
Time	333242.01	3	111080.67	1109.92
Error	799636.18	7990	100.08	
Total	1336815	7999		

In all cases, $T \sim F(3, 7990)$.

- $\mathbf{H}_0$: $\beta_2 = \beta_3 = \beta_4 = 0$.
  - p-value $= \mathbb{P}(T \geq t_{\mathrm{C}}) = \mathbb{P}(T \geq 679.14) \approx 0$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the design factor significantly influences the response and hence the expected response is not the same in each condition.
- $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = \gamma_3 = 0$.
  - p-value $= \mathbb{P}\left(T \geq t_{\mathrm{B}_1}\right) = \mathbb{P}(T \geq 0.1097) = 0.9545$.
  - Therefore, we do not reject $\mathbf{H}_0$ and conclude that "browser" does not significantly influence average latency, and so blocking by browser was probably not necessary.
- $\mathbf{H}_0$: $\delta_2 = \delta_3 = \delta_4 = 0$.
  - p-value $= \mathbb{P}\left(T \geq t_{\mathrm{B}_2}\right) = \mathbb{P}(T \geq 1109.92) \approx 0$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the time of day significantly influences average latency and so blocking by it was sensible.

[R Code] `Latin_square_means`

4.3.3 Latin Squares to Compare Proportions

- Here we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factors):
$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_p \text{ versus } \mathbf{H}_\text{A}\text{: } \pi_j \neq \pi_{j'} \text{ for some } j \neq j'$$
where π_j is the expected response in condition $j = 1, 2, \ldots, p$.
- We do this by testing:
$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j$$
with a likelihood ratio test (LRT) in the context of the following logistic regression model:
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{p-1}\beta_j x_{ij} + \sum_{k=1}^{p-1}\gamma_k z_{ik} + \sum_{\ell=1}^{p-1}\delta_\ell w_{i\ell}$$
  - $Y_i = 1$ if unit i performs some action of interest, and $Y_i = 0$ otherwise.
  - $\pi_i = \mathbb{E}[Y_i]$ = expected response of unit i.
  - The likelihood ratio test compares the full model to the one without the x's.
- Similarly, we test:
$$\mathbf{H}_0\text{: } \gamma_1 = \gamma_2 = \cdots = \gamma_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \gamma_k \neq 0 \text{ for some } k$$
with a LRT that compares the full model to the reduced one without the z's.
- And we test:
$$\mathbf{H}_0\text{: } \delta_1 = \delta_2 = \cdots = \delta_{p-1} = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \delta_\ell \neq 0 \text{ for some } \ell$$
with a LRT that compares the full model to the reduced one without the w's.
- The observed test statistic for all of these tests is:
$$\begin{aligned} t &= 2\log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2\Big[\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}\Big] \end{aligned}$$
which, if $\mathbf{H}_0$ is true, follows an approximate $\chi^2(p-1)$.
- p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(p-1)$.

4.3.4 Example: Uber Weekend Promos

EXAMPLE 4.3.1: Uber Weekend Promos

Consider an experiment in which Uber is investigating the influence of three different promotional offers on ride-booking-rate (RBR).

- Promo A: None.
- Promo B: One free ride today.
- Promo C: Book a ride today and get 50% off your next 2 rides.

The experimenters would like to control for a possible day-of-week effect, and so they want to block by day. They would also like to control for possible city-to-city differences, and so they also want to block by city. To do so they run a 3×3 Latin square design as illustrated in Table 4.13. Interest lies in determining whether the different promotions perform similarly with respect to RBR — and they wish to determine which one maximizes RBR — while controlling for the effects of day and city. In order to do this they randomize $n = 1000$ users to each of the $p^2 = 9$ blocks.

Table 4.13: 3×3 Latin Square Design for the Uber Experiment

		City		
		Toronto	Vancouver	Montreal
Day	Friday	A	B	C
	Saturday	C	A	B
	Sunday	B	C	A

We analyze the data with the following logistic regression model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \delta_1 w_{i1} + \delta_2 w_{i2}$$

- x_{i2}, x_{i3}, are condition indicators for promotions B, C, where A is the baseline.
- z_{i1}, z_{i2}, are day indicators for Saturday, Sunday, where Friday is the baseline.
- w_{i1}, w_{i2}, are city indicators for Toronto, Vancouver, where Montreal is the baseline.
- $\mathbf{H}_0$: $\beta_2 = \beta_3 = 0$.
  - p-value $= \mathbb{P}(T \geq t_{\text{C}}) = \mathbb{P}(T \geq 16.648) = 0.00024$ where $T \sim \chi^2(2)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the booking rate is not the same for each promotional offer.
- $\mathbf{H}_0$: $\gamma_1 = \gamma_2 = 0$.
  - p-value $= \mathbb{P}\left(T \geq t_{\text{B}_1}\right) = \mathbb{P}(T \geq 8.9107) = 0.01162$ where $T \sim \chi^2(2)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the day-of-week significantly influences booking, and so it is good that we blocked by this factor.
- $\mathbf{H}_0$: $\delta_1 = \delta_2 = 0$.
  - p-value $= \mathbb{P}\left(T \geq t_{\text{B}_2}\right) = \mathbb{P}(T \geq 2.1193) = 0.3466$ where $T \sim \chi^2(2)$.
  - Therefore, we do not reject $\mathbf{H}_0$ and conclude that "city" does not significantly influence booking rate, and so blocking by city may have not been necessary.

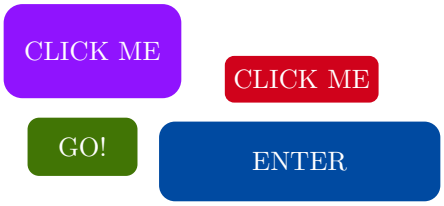
[R Code] `Latin_square_proportions`

Chapter 5

EXPERIMENTS WITH MULTIPLE DESIGN FACTORS

Week 7

- We now consider the design and analysis of experiments consisting of multiple conditions arising from multiple design factors.
  - This is often colloquially referred to as "multivariate testing" (MVT).
- Canonical MVT button test:



  Many things influence click-through rate:
  - Colour.
  - Size.
  - Position.
  - Phrase.
- Some additional, more tangible, examples:
  - Etsy.
  - Netflix.
  - Airbnb.
- This week we describe how to design and analyze experiments that efficiently investigate multiple design factors.
- **Goals**:
  1. Determine which condition is optimal with respect to some metric of interest.
     - But now a condition is defined by a specific combination of the levels of *multiple* design factors.
  2. Determine *which* factors are influential and understand *how* the factors influence the response.

5.1 The Factorial Approach

- The key to multi-factor experiments is to *efficiently* investigate different combinations of the factor levels.
- **One-factor-at-a-time approach** (OFAT):
  - Sequence of experiments, each with just one factor being varied.
  - The winning level of this factor is retained.
  - Follow-up experiments manipulate some other factor, while the previous ones are held fixed at their optimal levels.

EXAMPLE 5.1.1: Button

1. Experiment with colour, and purple wins.
2. Experiment with phrase, and "submit" wins.
3. Experiment with size, and "large" wins.

EXAMPLE 5.1.2: Twitter ♡ versus ⋆

- In 2015 Twitter changed "favouriting" a tweet (expressed as stars) to "liking" a tweet (expressed as hearts) and the internet was pissed.
- In line with Twitter's "test everything" motto, this decision came about as a result of experimentation.
- A hypothetical experiment that could have lead to this decision might involve two factors at two levels.
  - DF1 = Shape $\rightarrow \{\text{Star}, \text{Heart}\}$.
  - DF2 = Colour $\rightarrow \{\text{Yellow}, \text{Red}\}$.
- A one-factor-at-a-time approach might look like this:
  - Experiment 1: ⋆ versus ♡.
  - Experiment 2: ⋆ versus ♡.
- Suppose they conclude that ♡ is the best, but what about ⋆? The problem with OFAT experiments is that we may never observe the truly optimal combination of factor levels.

EXAMPLE 5.1.3: Etsy Search Bar

- Check it out.

- **The Factorial approach**:
  - Experimental conditions are defined as every unique combination of the design factors' levels.
  - In the **Twitter** example, the factorial experiment would have looked like this: ⋆⋆♡♡.
  - In the **Etsy** example, the factorial experiment would have looked like this:
    Small Square Box, *Small Rounded Box*, *Long Square Box*, *Long Rounded Box*.
  - **Advantage**: it explores every possible condition.
    1. We don't miss the optimal condition.

2. Our understanding of the relationship between the response and design factors is better with a factorial experiment than with an OFAT experiment.

– **Disadvantage**: it explores every possible condition.

∗ With many design factors at many levels, the number of unique combinations might be unmanageably large.

– As long as we choose our factors and their levels thoughtfully, the advantages outweigh the disadvantages.

- **Main effects**: The main effect of factor A, represents the change in the response variable produced by a change in that factor. In our **Twitter** example:
  - Main effect of shape: what happens to the response when we change the shape from ⋆ to ♡.
  - Main effect of colour: what happens to the response when we change the colour from yellow to red?
- **Interaction effects**: If the main effect of factor A depends on the level of some other factor B, we say that factors A and B interact. In our **Twitter** example:
  - Is there an interaction between shape and colour?

  $\iff$ Does the effect of going from ⋆ to ♡ depend on colour?

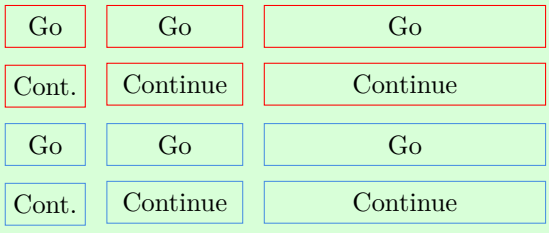
  $\iff$ Does the effect of going from yellow to red depend on shape?
- From a practical perspective it is critical to quantify both types of effects.
  - The only type of design that allows us to observe and estimate both main and interaction effects is the factorial design. The OFAT cannot do this.

5.2 Designing a Factorial Experiment

- Conceptually, the design of a factorial experiment is simple.
  1. Pick your metric of interest and define the corresponding response variable.
  2. Pick your design factors.
  3. Pick their levels.
  4. Define your experimental conditions (all possible combinations of our design factors' levels).
  5. Determine your sample sizes.

EXAMPLE 5.2.1: Button

Suppose you have $K = 3$ factors, *colour* (red, blue), *phrase* ("Continue", "Go"), and *size* (small, medium, large). These factors therefore have $m_1 = 2$, $m_2 = 2$, and $m_3 = 3$ levels respectively. Therefore, we have $m = m_1 m_2 m_3 = (2)(2)(3) = 12$ experimental conditions.



- In general, a factorial experiment with K factors requires $m = m_1 m_2 \cdots m_K$ conditions, where m_k is the number of levels of design factor k.

- As the number of factors and levels increase, the size of the experiment can get unmanageably large.
  - As such, we want to pick our factors and levels thoughtfully. Keep it simple.
    1. Don't investigate factors that are highly correlated.
    2. Don't choose levels that are very similar.
    3. Don't choose factors that are hard to manipulate outside an experiment.
- Once the factors, levels, and hence experimental conditions have been established, experimental units must be randomized to each of the m conditions.
  - The number of experimental units assigned to each condition n_j, $j = 1, 2, \ldots, m$, can be determined by sample size calculations associated with two-sample tests.
  - Make sure to account for the multiple comparison problem.

5.3 Analyzing a Factorial Experiment

- In order to determine which condition is optimal we use pairwise tests.
- In order to determine which factors are influential, and to quantify this influence we use regression.
- Whether it's a linear or logistic regression, we use a linear predictor which contains the following terms:
  - An intercept.
  - Main effect terms. We represent a factor with m_k levels using $m_k - 1$ variables.
  - Two-factor interaction terms, three-factor interaction terms,… K-factor interaction terms.
  - In general, an h-factor interaction is represented by h-way products of the main effect indicators for the factors involved.

EXAMPLE 5.3.1: Button

With $K = 3$ factors with $m_1 = 2$, $m_2 = 2$, and $m_3 = 3$ levels, the required linear predictor will contain:

- Main effect terms, and
  - Let x_1 be an indicator for "colour."
  - Let x_2 be an indicator for "phrase."
  - Let x_3 and x_4 be indicators for "size."
- Two-factor interaction terms, and pairwise products of indicators for each pair of factors.
- Three-factor interaction terms. Three-way products of the indicators for all three factors.
- The linear predictor is given by:

$$\beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4}_{\text{main effects}} + \underbrace{\beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \beta_9 x_2 x_4}_{\text{two-factor interactions}} + \underbrace{\beta_{10} x_1 x_2 x_3 + \beta_{11} x_1 x_2 x_4}_{\text{three-factor interactions}}$$

- Hypotheses concerning the main effects will therefore involve $\beta_1, \beta_2, \beta_3, \beta_4$.
- Hypotheses concerning the two-factor interactions will involve $\beta_5, \beta_6, \beta_7, \beta_8, \beta_9$.
- Hypotheses concerning the three-factor interactions will involve β_{10}, β_{11}.

- In general, hypotheses of these sort will be performed by comparing full versus reduced models via partial F-tests (in the case of linear regression) and likelihood ratio tests (in the case of logistic regression).

5.3.1 Continuous Response — The Instagram Example

- We illustrate the topics discussed in this section in the context of an **Instagram Ad** example.
- Suppose that you are a data scientist at Instagram, and you are interested in running an experiment to learn about how ad frequency and ad type influences user engagement.
- Suppose that ad frequency has levels $\{9{:}1, 7{:}1, 4{:}1, 1{:}1\}$ corresponding to ad frequencies of 1 in 10, 1 in 8, 1 in 5, and every other.
- Suppose that ad type is a second design factor with levels $\{\text{photo}, \text{video}\}$.
- We will consider here the factorial experiment that considers every combination of these two factors' levels. Therefore, $m = m_1 m_2 = (4)(2) = 8$ conditions.
- Assume $n = 1000$ users are randomly assigned to each of these $m = 8$ conditions, and on each user we measure the length of time they engage with the app (in minutes).
- We use the resulting data to create the following **main effect plots** (i.e., the plots of MOI versus the levels of each factor) in Figure 5.1.

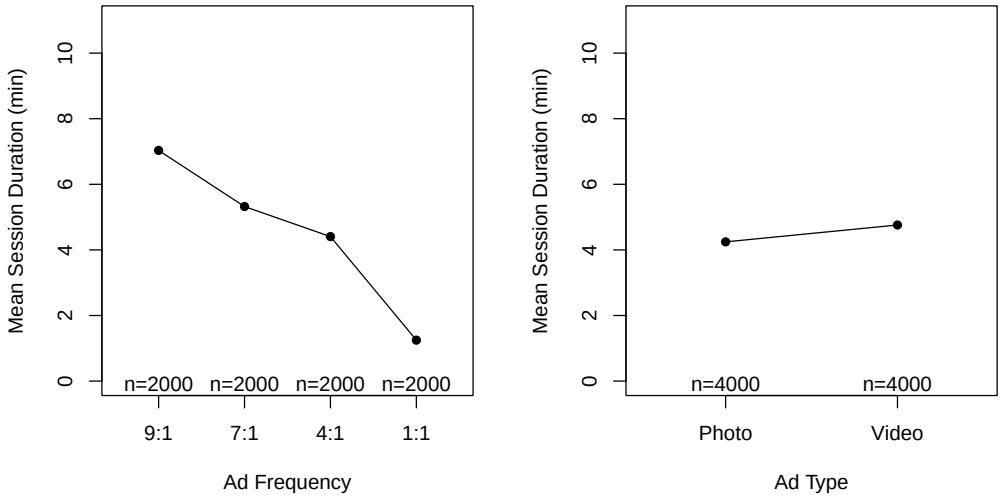


Figure 5.1: Left: Main Effect Plot for Ad Frequency; Right: Main Effect Plot for Ad Type

  - As ad frequency increases, we see that average session duration increases.
  - Average session duration increases with video ads relative to photo ads.
  - Ad frequency appears to be more influential than ad type since the change in ASD produced by changes in frequency is large (in magnitude) than those produced by changes in ad type.

- **Important**:
  - Discussing main effects can be uninformative and potentially misleading if there is a significant interaction between the factors
  - In the presence of a significant interaction effect, it no longer makes sense to discuss the main effect of a factor in isolation, because doing so ignores the fact that this effect changes depending on the level of another factor.

- We can evaluate the presence of such interaction by studying **interaction effect plots** (i.e., plots of the MOI at each level of DF1 with different line types distinguishing the levels of DF2) in Figure 5.2.

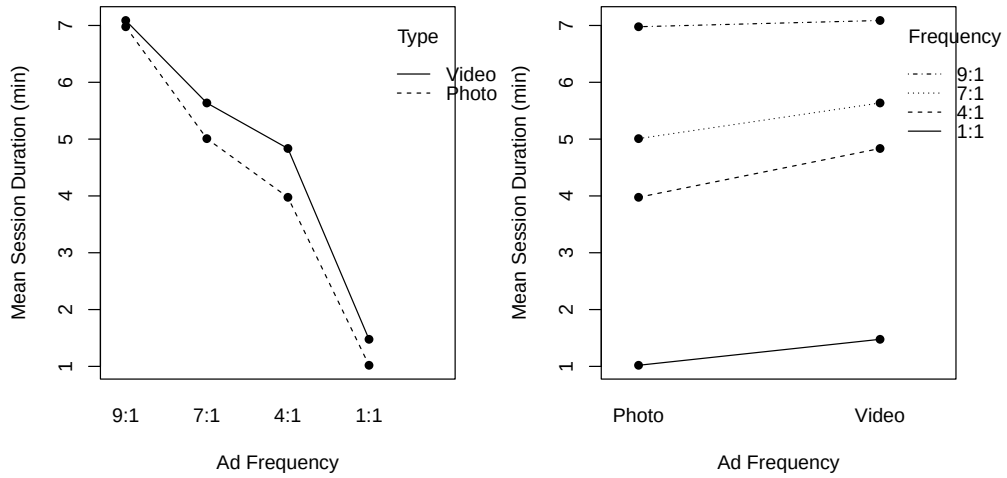


Figure 5.2: Interaction Plot for Ad Frequency and Ad Type.

- Non-parallel line segments on these plots would indicate the presence of an interaction since this would correspond to the main effect of one factor depending on the levels of the other factor.
  - The line segments in these plots are not perfectly parallel and so an interaction appears to exist.
  - However, the departure from parallelism is not drastic, and so this interaction is perhaps not strong.
- To formally evaluate whether these main effects and interaction effects are significant we fit the following linear regression model:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{main effects of frequency}} + \underbrace{\beta_4 x_{i4}}_{\text{main effects of type}} + \underbrace{\beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4}}_{\text{two-factor interaction}} + \varepsilon_i$$

  where the x's are indicator variables.
  - $x_{i1} = 1$ if unit i is in a condition with the 7:1 ad frequency.
  - $x_{i2} = 1$ if unit i is in a condition with the 4:1 ad frequency.
  - $x_{i3} = 1$ if unit i is in a condition with the 1:1 ad frequency.
  - $x_{i4} = 1$ if unit i is in a condition with video ads.
- The expected response in each condition, according to this model is Table 5.1.

		Ad Type	
		Photo	Video
Freq.	9:1	$\mathbb{E}[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$\mathbb{E}[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$\mathbb{E}[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$\mathbb{E}[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4 + \beta_5$
	4:1	$\mathbb{E}[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$\mathbb{E}[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4 + \beta_6$
	1:1	$\mathbb{E}[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$\mathbb{E}[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4 + \beta_7$

Table 5.1: Expected Response in Each Ad Frequency-Type Condition

- Clearly, a formal test of:

$$\mathbf{H}_0\text{: } \beta_5 = \beta_6 = \beta_7 = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0$$

  for $j = 5, 6, 7$ would evaluate the significance of the interaction effect.
  - If we reject $\mathbf{H}_0$, any conclusions regarding the effect of one factor must be made in the context of the levels of the other factor.

– If we do not reject $\mathbf{H}_0$, the interaction terms can be removed from the model yielding the following simplified **main effects** model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

which can be used to evaluate the significance of the main effect of each factor.

– The expected response in each condition, according to the main effects model is Table 5.2.

		Ad Type	
		Photo	Video
Freq.	9:1	$\mathbb{E}[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$\mathbb{E}[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$\mathbb{E}[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$\mathbb{E}[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4$
	4:1	$\mathbb{E}[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$\mathbb{E}[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4$
	1:1	$\mathbb{E}[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$\mathbb{E}[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4$

Table 5.2: Expected Response in Each Ad Frequency-Type Condition, Based on the Main Effects Model

– The hypothesis:

$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \beta_3 = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0$$

for $j = 1, 2, 3$ tests whether ad frequency is a significant factor.

– The hypothesis:

$$\mathbf{H}_0\text{: } \beta_4 = 0 \text{ versus } \mathbf{H}_\text{A}\text{: } \beta_4 \neq 0$$

tests whether ad type is a significant factor.

– **But remember**: these tests and the interpretation of main effects are only appropriate in the absence of interaction.

- Each of these null hypotheses generates a **reduced model** with fewer terms relative to a **full model** with all terms — we compare them using partial F-tests associated with an analysis of variance.
- Output from the relevant partial F-tests is shown below:

```
Model 1: Time ~ Frequency + Type
Model 2: Time ~ Frequency * Type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7995 6522.2
2   7992 6372.9  3    149.27 62.398 < 2.2e-16 ***

Model 1: Time ~ Frequency
Model 2: Time ~ Frequency + Type
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   7996 7049.5
2   7995 6522.2  1    527.34 646.43 < 2.2e-16 ***

Model 1: Time ~ Type
Model 2: Time ~ Frequency + Type
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1   7998 41875
2   7995  6522  3     35353 14445 < 2.2e-16 ***
```

- Conclusions:

– The p-value associated with $\mathbf{H}_0$: $\beta_5 = \beta_6 = \beta_7 = 0$ is very small, suggesting that the interaction between ad frequency and ad type is significant.

– The p-value associated with $\mathbf{H}_0$: $\beta_4 = 0$ is very small, suggesting that the main effect of ad type is significant.

  - The p-value associated with $\mathbf{H}_0$: $\beta_1 = \beta_2 = \beta_3 = 0$ is very small, suggesting that the main effect of ad frequency is significant.
    * Strictly speaking, the last two conclusions are irrelevant because we know the interaction is significant. Although, I include this for instructional purposes.

- [R Code] `Factorial_example_means`

5.3.2 Binary Response — The TinyCo Example

- The informal and formal evaluation of main and interaction effects can be performed in the context of a binary response variable as well.
  - Main effect and interaction effect plots are based on observed proportions.
  - Logistic regression is used instead of ordinary linear regression.
- The structure of the linear predictor is identical to what we have discussed in general.
- For instance, if the Instagram experiment from the previous section had a binary response instead, the relevant logistic regression model would be

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4}$$

  where the x's are the indicator variables defined previously.
- Interest lies in determining whether subsets of the β's are equal to zero to evaluate the significance of various main and interaction effects.
  - We use **likelihood ratio tests** for the comparison of full and reduced logistic regression models.
  - The test statistic for the LRT is:

$$\begin{aligned} t &= 2\log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2\Big[\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}\Big] \end{aligned}$$

  - If $\mathbf{H}_0$ is true, then t should look like it comes from a $\chi^2(\nu)$ distribution, where

$$\nu = (\#\text{ coefficients in full model}) - (\#\text{ coefficients in reduced model})$$

  - p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(\nu)$.

Week 8

Pit Stop: *Effects* vs. *Terms* in a Linear Predictor

Example: Suppose Factor A has $m_1 = 5$ levels, Factor B has $m_2 = 2$ levels, and Factor C has $m_3 = 3$ levels.

- The main effect for a given factor is represented by indicator variables corresponding to the levels of that factor.
  - **Factor A** will be represented in a regression model by $m_1 - 1 = 4$ indicator variables: x_1, x_2, x_3, x_4, and $m_1 - 1 = 4$ corresponding β's: $\beta_1 = \beta_2 = \beta_3 = \beta_4$.
  - Thus, the *main effect* of Factor A is composed of 4 *terms* in the model.
  - To determine the significance of the main effect of Factor A, we test:

$$\mathbf{H}_0\text{: } \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

  - **Factor B** will be represented in a regression model by $m_2 - 1 = 1$ indicator variable: x_5, and $m_2 - 1 = 1$ corresponding β: β_5.
  - Thus, the *main effect* of Factor B is composed of 1 *term* in the model.
  - To determine the significance of the main effect of Factor B, we test:
$$\mathbf{H_0}\text{: } \beta_5 = 0.$$
  - **Factor C** will be represented in a regression model by $m_3 - 1 = 2$ indicator variables: x_6, x_7, and $m_3 - 1 = 2$ corresponding β's: β_6, β_7.
  - Thus, the *main effect* of Factor C is composed of 2 *terms* in the model.
  - To determine the significance of the main effect of Factor C, we test:
$$\mathbf{H_0}\text{: } \beta_6 = \beta_7 = 0.$$
  - **Note**: These three hypotheses are relevant only in the context of the *main effect* model which has linear predictor:
$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

- The interaction effect between *two* factors is represented by two-way products of the indicator variables corresponding to the main effects of the two factors.
  - The **A:B interaction** *effect* is composed of the $(m_1 - 1) \times (m_2 - 1) = 4 \times 1 = 4$ *terms* resulting from the two-way products between Factor A's and Factor B's indicator variables: $x_1x_5, x_2x_5, x_3x_5, x_4x_5$, with corresponding β's: $\beta_8, \beta_9, \beta_{10}, \beta_{11}$.
  - The significance of the A:B interaction effect is determined by testing:
$$\mathbf{H_0}\text{: } \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0.$$
  - The **A:C interaction** *effect* is composed of the $(m_1 - 1) \times (m_3 - 1) = 4 \times 2 = 8$ *terms* resulting from the two-way products between Factor A's and Factor C's indicator variables:
$$x_1x_6, x_2x_6, x_3x_6, x_4x_6, x_1x_7, x_2x_7, x_3x_7, x_4x_7,$$
    with corresponding β's: $\beta_{12}, \beta_{13}, \ldots, \beta_{19}$.
  - The significance of the A:C interaction effect is determined by testing:
$$\mathbf{H_0}\text{: } \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = 0.$$
  - The **B:C interaction** *effect* is composed of the $(m_2 - 1) \times (m_3 - 1) = 1 \times 2 = 2$ *terms* resulting from the two-way products between Factor B's and Factor C's indicator variables: x_5x_6, x_5x_7, with corresponding β's: β_{20}, β_{21}.
  - The significance of the B:C interaction effect is determined by testing:
$$\mathbf{H_0}\text{: } \beta_{20} = \beta_{21} = 0.$$

- The interaction effect between *three* factors is represented by three-way products of the indicator variables corresponding to the main effects of the three factors.
  - The **A:B:C interaction** *effect* is composed of the $(m_1 - 1) \times (m_2 - 1) \times (m_3 - 1) = 4 \times 1 \times 2 = 8$ *terms* resulting from the three-way products between Factor A's, Factor B's, and Factor C's indicator variables:
$$x_1x_5x_6, x_1x_5x_7, x_2x_5x_6, x_2x_5x_7, x_3x_5x_6, x_3x_5x_7, x_4x_5x_6, x_4x_5x_7,$$
    with corresponding β's: $\beta_{22}, \ldots, \beta_{29}$.
  - The significance of the A:B:C interaction effect is determined by testing:
$$\mathbf{H_0}\text{: } \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = \beta_{27} = \beta_{28} = \beta_{29} = 0.$$

– **Note**: The hypotheses concerning the significance of interaction effects are relevant only in the context of the *full model* which has linear predictor:

$$\begin{aligned}
&\text{Intercept} \to \beta_0 + \\
&\quad \text{ME's} \to \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \\
&\quad \text{2FI's} \to \begin{cases} \beta_8 x_1 x_5 + \beta_9 x_2 x_5 + \beta_{10} x_3 x_5 + \beta_{11} x_4 x_5 + \beta_{12} x_1 x_6 + \beta_{13} x_2 x_6 + \beta_{14} x_3 x_6 + \\ \beta_{15} x_4 x_6 + \beta_{16} x_1 x_7 + \beta_{17} x_2 x_7 + \beta_{18} x_3 x_7 + \beta_{19} x_4 x_7 + \beta_{20} x_5 x_6 + \beta_{21} x_5 x_7 + \end{cases} \\
&\quad \text{3FI's} \to \begin{cases} \beta_{22} x_1 x_5 x_6 + \beta_{23} x_1 x_5 x_7 + \beta_{24} x_2 x_5 x_6 + \beta_{25} x_2 x_5 x_7 + \\ \beta_{26} x_3 x_5 x_6 + \beta_{27} x_3 x_5 x_7 + \beta_{28} x_4 x_5 x_6 + \beta_{29} x_4 x_5 x_7 \end{cases}
\end{aligned}$$

- **ALL** the tests discussed here are carried by either a **partial *F*-test** or a **likelihood ratio test**.
  - Partial F-test: p-value $= \mathbb{P}(T \geq t)$ where $T \sim F(\nu, h)$.
  - Likelihood ratio test: p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(\nu)$.
  - $\nu =$ (# β's in "full" model) $-$ (# β's in "reduced" model).
  - $h =$ error degrees of freedom in the "full" model $= N -$ # β's in "full" model.

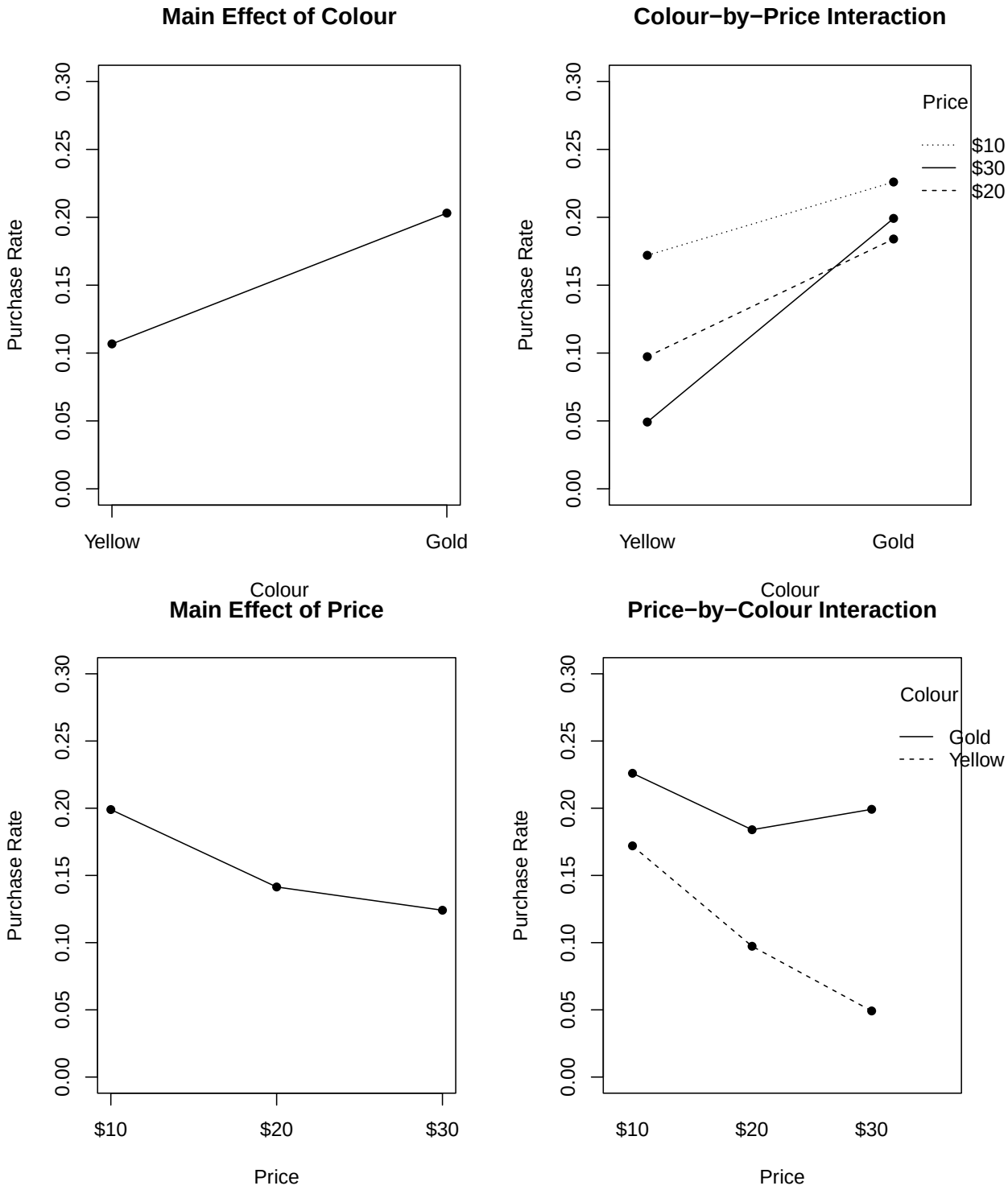
TinyCo is a mobile video game studio that develops the Tiny Zoo game. In this game users own zoos and collect animals to put in their zoos. An experiment is performed in which a new animal, the "bananimal," is released for purchase as a part of the Super Sweet Series. Interest lies in understanding the relationship between conversion (purchase rate) and two factors: the bananimal's colour (yellow or gold) and the bananimal's price ($10, $20, or $30 of in-game currency). A factorial experiment with 6 conditions was performed to investigate these relationships. A summary of the data resulting from this experiment is shown below.

Condition	Sample Size	Purchase Rate
\$10 + Yellow	500	0.1720
\$20 + Yellow	483	0.0973
\$30 + Yellow	488	0.0492
\$10 + Gold	500	0.2260
\$20 + Gold	500	0.1840
\$30 + Gold	487	0.1992

- What does the main effect plots tell us?
  - ME of colour: gold bananimals are purchased more frequently than yellow.
  - ME of price: we expect purchase rate to decrease as bananimal price increases.
  - However, we should not stop here because an interaction exists.
- What does the interaction effect plots tell us?
  - A price-colour interaction exists.
  - We see that increasing price from $20 to $30 for gold bananimals increases purchase rate, whereas the same increase for yellow bananimals decreases purchase rate.
- To formally analyze the data, we fit the *full* logistic regression model with linear predictor:

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3}$$

  - $x_{i1} = 1$ if unit i is in a gold bananimal condition.
  - $x_{i2} = 1$ if unit i is in a $20 bananimal condition.
  - $x_{i3} = 1$ if unit i is in a $30 bananimal condition.


Main Effect of Colour
Purchase Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Yellow
Gold
Colour
Colour−by−Price Interaction
Price
$10
$30
$20
Main Effect of Price
$10
$20
$30
Price
Price−by−Colour Interaction
Colour
Gold
Yellow

- We test the significance of the interaction effects via the hypothesis:
$$\mathbf{H}_0\text{: } \beta_4 = \beta_5 = 0 \text{ vs. } \mathbf{H}_\text{A}\text{: } \beta_j \neq 0 \text{ for some } j = 4, 5.$$
- This involves a comparison between the full model and the reduced *main effects* model with linear predictor:
$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$
  - p-value $= \mathbb{P}(T \geq 19.918) = 4.731 \times 10^{-5}$ where $T \sim \chi^2(2)$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that the price-colour interaction is significant.
- We can also test the main effect of colour with $\mathbf{H}_0\text{: } \beta_1 = 0$ in the context of the main effects model:
  - $t = 53.757$.
  - p-value $= 2.269 \times 10^{-13}$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that colour does significantly influence purchase rate.
- We can also test the main effect of price with $\mathbf{H}_0\text{: } \beta_2 = \beta_3 = 0$ in the context of the main effects model:
  - $t = 23.324$.
  - p-value $= 8.614 \times 10^{-6}$.
  - Therefore, we reject $\mathbf{H}_0$ and conclude that price does significantly influence purchase rate.
- So what have we learned about the influence of these factors?
  - Colour and price both significantly influence purchase rate.
  - Gold bananimals tend to be purchased more often than yellow.
  - Increasing price from \$10 to \$20 decreases purchase rate (for both colours) and increasing price from \$20 to \$30 increase purchase rate for gold but not yellow bananimal.
- And which condition was optimal?
  - It turns out that the purchase rate is not statistically significantly different in the three gold bananimal conditions. So, \$30 gold bananimals seem like a good choice for TinyCo.
- [R Code] `Factorial_example_proportions`

5.4 Two-Level Factorial Experiments

- Factorial experiments are the most informative means of exploring several design factors.
- But this may require a larger number of experimental conditions than is practically feasible.
- As a compromise, we might consider **two-level factorial experiments**.
  - This is a factorial experiment where each factor is experimented at just two levels.
- Such an experiment is typically used for **factor screening**.
  - Among a larger number of factors, we want to determine which significantly influence the response.
- Factoring screening is predicated on the **Pareto Principle**.
  - Only a "vital few" factors will be important relative to the "trivial many."
- We will discuss two types of two-level factorial experiments:
  - $\mathbf{2^K}$ **factorial designs**: investigates K design factors in 2^K conditions (i.e., all possible combinations of the factors' levels).
  - $\mathbf{2^{K-p}}$ **factorial designs**: investigates K design factors in 2^{K-p} conditions (i.e., just a fraction of all possible combinations of the factors' levels).

Chapter 6

2^K FACTORIAL EXPERIMENTS

6.1 Designing 2^K Factorial Experiments

Week 9

* 2^K factorial experiments involve K design factors, each at two levels.

- These experiments are typically used for factor screening.
  - → **Primary Goal**: Determine which among the K factors significantly influence the response variable.
  - → **Secondary Goal**: Determine which combination of levels is optimal.
    - ↪ This is really only relevant if the levels experimented with are the only ones of interest.

→ The design of the experiment involves:

1. Choose the MOI and response variables.
   - ↪ Dictated by the "question."
2. Choose the design factors.
   - ↪ Choose K factors that may influence the response and that you want to learn about.
3. Choose the levels of the design factors.
   - ↪ With the goal of factor screening we want to give influential factors as fair an opportunity as possible to show themselves as being influential.
   - ↪ Pick levels that are quite different. For example, colour and discount amount.
4. Define experimental conditions.
   - ↪ These are the 2^K unique combinations of the K factors' levels.
5. Assign n experimental units to each condition.
   - ↪ Balance is not necessary, it's just notationally convenient.
   - ↪ Overall sample size: $N = n2^K$.

- In two-level experiments we regard the two levels of a factor as *low* and *high* values of that factor.
  - * If a factor is categorical, then "low" vs. "high" labelling is arbitrary.

- We represent each factor by a binary variable:

$$x = \begin{cases} -1 & \text{if the factor is at its "low" level} \\ 1 & \text{if the factor is at its "high" level} \end{cases}$$

* We could alternatively code each x as an indicator variable, but the ± 1 coding gives rise to some convenient statistical properties.

→ With the factor levels coded in this way, each experimental condition can be identified by a unique combination of plus and minus ones.

- The experimental design can be completely summarized by the **design matrix**.
  - 2^K rows (conditions) and K columns (factors) of plus and minus ones.

  → The ± 1 entries are organized such that each row corresponds to a unique condition and the columns correspond to each of the factors.

  → The design matrix provides a prescription for running the 2^K factorial experiment.

EXAMPLE 6.1.1: 2^1 Design Matrix

$$\begin{matrix}\text{C1} \rightarrow \\ \text{C2} \rightarrow\end{matrix}\begin{bmatrix}-1 \\ +1\end{bmatrix} = [\boldsymbol{x}_1]$$

EXAMPLE 6.1.2: 2^2 Design Matrix

$$\begin{matrix}\text{C1} \rightarrow \\ \text{C2} \rightarrow \\ \text{C3} \rightarrow \\ \text{C4} \rightarrow\end{matrix}\begin{bmatrix}-1 & -1 \\ +1 & -1 \\ -1 & +1 \\ +1 & +1\end{bmatrix} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2]$$

EXAMPLE 6.1.3: 2^3 Design Matrix

$$\begin{matrix}\text{C1} \rightarrow \\ \text{C2} \rightarrow \\ \text{C3} \rightarrow \\ \text{C4} \rightarrow \\ \text{C5} \rightarrow \\ \text{C6} \rightarrow \\ \text{C7} \rightarrow \\ \text{C8} \rightarrow\end{matrix}\begin{bmatrix}-1 & -1 & -1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ +1 & +1 & -1 \\ -1 & -1 & +1 \\ +1 & -1 & +1 \\ -1 & +1 & +1 \\ +1 & +1 & +1\end{bmatrix} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \boldsymbol{x}_3]$$

→ 2^K experiments may also be visualized geometrically as K-dimensional hypercubes. See Figure 6.1.

- Vertices correspond to the unique configurations of the K factors' levels, and hence experimental conditions.

  Design Space: the space of all possible combinations of the design factors' values.

6.2 Analyzing 2^K Factorial Experiments

- Primary goal of a 2^K factorial experiment is factor screening.
  - Interest lies primarily in estimation of main and interaction effects.

* The **main effect** of a factor is defined as the expected change in response produced by changing that factor from its low to its high level.

* The **interaction effect** between two factors quantifies the difference between the main effect of one factor at the two levels of the other.

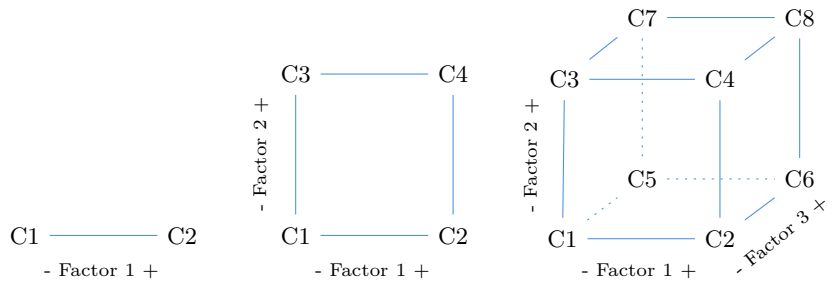


Figure 6.1: Cuboidal representation of 2^1 (left), 2^2 (middle), and 2^3 (right) factorial designs.

6.2.1 An Intuition-Based Analysis

EXAMPLE 6.2.1: Toy Example

Factors A and B are investigated in a 2^2 factorial experiment with $n = 3$.

Condition	Factor A	Factor B	Response (y)	Average Response ($\bar{y}$)
1	-1	-1	$\{1, 1, 2\}$	$4/3$
2	$+1$	-1	$\{3, 4, 5\}$	$12/3$
3	-1	$+1$	$\{2, 1, 3\}$	$6/3$
4	$+1$	$+1$	$\{1, 2, 5\}$	$8/3$

- Intuitive estimate of the main effect of A:

$$\begin{aligned}\widehat{\text{ME}}_{\text{A}} = \bar{y}_{\text{A}^+} - \bar{y}_{\text{A}^-} &= \frac{\bar{y}_{\text{A}^+\cap\text{B}^-} + \bar{y}_{\text{A}^+\cap\text{B}^+}}{2} - \frac{\bar{y}_{\text{A}^-\cap\text{B}^-} + \bar{y}_{\text{A}^-\cap\text{B}^+}}{2} \\ &= \frac{(12/3) + (8/3)}{2} - \frac{(4/3) + (6/3)}{2} \\ &= 10/6\end{aligned}$$

  Therefore, we expect the average response to go up by $10/6$ when A is moved from its low to high level.

- Intuitive estimate of the main effect of B:

$$\begin{aligned}\widehat{\text{ME}}_{\text{B}} = \bar{y}_{\text{B}^+} - \bar{y}_{\text{B}^-} &= \frac{\bar{y}_{\text{A}^-\cap\text{B}^+} + \bar{y}_{\text{A}^+\cap\text{B}^+}}{2} - \frac{\bar{y}_{\text{A}^-\cap\text{B}^-} + \bar{y}_{\text{A}^+\cap\text{B}^-}}{2} \\ &= \frac{(6/3) + (8/3)}{2} - \frac{(4/3) + (12/3)}{2} \\ &= -1/3\end{aligned}$$

  Therefore, we expect the average response to go down by $1/3$ when B is moved from its low to high level.

- To evaluate whether factors A and B interact, we should compare the main effect of A when B is at its high level to the main effect of A when B is at its low level.
  - Conditional ME of A when B is high:

$$\widehat{\text{ME}}_{\text{A}|\text{B}^+} = \bar{y}_{\text{A}^+\cap\text{B}^+} - \bar{y}_{\text{A}^-\cap\text{B}^+} = \frac{8}{3} - \frac{6}{3} = \frac{2}{3}$$

  - Conditional ME of A when B is low:

$$\widehat{\text{ME}}_{\text{A}|\text{B}^-} = \bar{y}_{\text{A}^+\cap\text{B}^-} - \bar{y}_{\text{A}^-\cap\text{B}^-} = \frac{12}{3} - \frac{4}{3} = \frac{8}{3}$$

Therefore, because $\widehat{\text{ME}}_{\text{A}|\text{B}^+} \neq \widehat{\text{ME}}_{\text{A}|\text{B}^-}$ we know there exists an A:B interaction.

- The interaction effect is defined as the average difference between the conditional main effects:

$$\begin{aligned}
\widehat{\text{IE}}_{\text{AB}} &= \frac{\widehat{\text{ME}}_{\text{A}|\text{B}^+}}{2} - \frac{\widehat{\text{ME}}_{\text{A}|\text{B}^-}}{2} \\
&= \frac{\widehat{\text{ME}}_{\text{B}|\text{A}^+}}{2} - \frac{\widehat{\text{ME}}_{\text{B}|\text{A}^-}}{2} \\
&= \frac{\bar{y}_{\text{A}^+\cap\text{B}^+} + \bar{y}_{\text{A}^-\cap\text{B}^-}}{2} - \frac{\bar{y}_{\text{A}^+\cap\text{B}^-} + \bar{y}_{\text{A}^-\cap\text{B}^+}}{2} \\
&= \frac{2}{6} - \frac{8}{6} \\
&= -1
\end{aligned}$$

- If a third factor C were involved, we may define the three-way ABC interaction as:

$$\begin{aligned}
\widehat{\text{IE}}_{\text{ABC}} &= \frac{\widehat{\text{IE}}_{\text{AB}|\text{C}^+}}{2} - \frac{\widehat{IE}_{\text{AB}|\text{C}^-}}{2} \\
&= \frac{\widehat{\text{IE}}_{\text{AC}|\text{B}^+}}{2} - \frac{\widehat{IE}_{\text{AC}|\text{B}^-}}{2} \\
&= \frac{\widehat{\text{IE}}_{\text{BC}|\text{A}^+}}{2} - \frac{\widehat{IE}_{\text{BC}|\text{A}^-}}{2}
\end{aligned}$$

- So what actually happened here? See Figure 6.2.

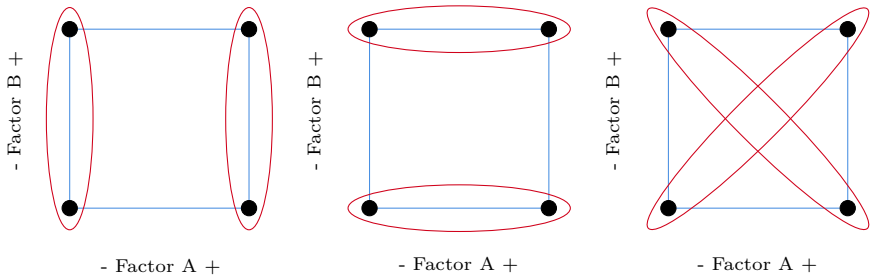


Figure 6.2: Visualization of main and interaction effects in a 2^2 factorial experiment.

  - ME of A: average response in the rightmost corners, minus the average response in the leftmost corners.
  - ME of B: average response in the topmost corners, minus the average response in the bottommost corners.
  - IE of AB: difference of the average response in ellipses joining opposing corners.

- These intuitive comparisons are still relevant when the response variable is binary:

$$\widehat{\text{ME}}_{\text{A}} = \sqrt{\frac{\bar{y}_{\text{A}^+\cap\text{B}^-}}{1-\bar{y}_{\text{A}^+\cap\text{B}^-}} \times \frac{\bar{y}_{\text{A}^+\cap\text{B}^+}}{1-\bar{y}_{\text{A}^+\cap\text{B}^+}}} \div \sqrt{\frac{\bar{y}_{\text{A}^-\cap\text{B}^-}}{1-\bar{y}_{\text{A}^-\cap\text{B}^-}} \times \frac{\bar{y}_{\text{A}^-\cap\text{B}^+}}{1-\bar{y}_{\text{A}^-\cap\text{B}^+}}}$$

$$\widehat{\text{ME}}_{\text{B}} = \sqrt{\frac{\bar{y}_{\text{A}^-\cap\text{B}^+}}{1-\bar{y}_{\text{A}^-\cap\text{B}^+}} \times \frac{\bar{y}_{\text{A}^+\cap\text{B}^+}}{1-\bar{y}_{\text{A}^+\cap\text{B}^+}}} \div \sqrt{\frac{\bar{y}_{\text{A}^+\cap\text{B}^-}}{1-\bar{y}_{\text{A}^+\cap\text{B}^-}} \times \frac{\bar{y}_{\text{A}^-\cap\text{B}^-}}{1-\bar{y}_{\text{A}^-\cap\text{B}^-}}}$$

$$\widehat{\text{IE}}_{\text{AB}} = \sqrt{\frac{\bar{y}_{\text{A}^+\cap\text{B}^+}}{1-\bar{y}_{\text{A}^+\cap\text{B}^+}} \times \frac{\bar{y}_{\text{A}^-\cap\text{B}^-}}{1-\bar{y}_{\text{A}^-\cap\text{B}^-}}} \div \sqrt{\frac{\bar{y}_{\text{A}^+\cap\text{B}^-}}{1-\bar{y}_{\text{A}^+\cap\text{B}^-}} \times \frac{\bar{y}_{\text{A}^-\cap\text{B}^+}}{1-\bar{y}_{\text{A}^-\cap\text{B}^+}}}$$

↪ Where do these come from?

* Calculate the odds that $Y = 1$ in each corner (condition).
* Compare the corners in one red ellipse to the other.
* This comparison is based on a ratio of geometric means (as opposed to a difference of arithmetic means like in the non-binary case).

6.2.2 A Regression-Based Analysis

The Model

- Fitted regression models provide an estimate of the **response surface**.

  ↪ **Response surface**: functional relationship between the response and design factors.

- Each of the K factors is represented by the binary variables:

$$x_j = \begin{cases} -1 & \text{if the factor is at its "low" level} \\ 1 & \text{if the factor is at its "high" level} \end{cases} \qquad \text{for } j = 1, 2, \ldots, K$$

- Since each factor is represented by a single term, the linear predictor contains:
  - An intercept: β_0.
  - K main effect terms corresponding to $x_1, x_2, \ldots, x_K$.
  - $\binom{K}{2}$ two-factor interaction terms corresponding to $x_1x_2, x_1x_3, x_1x_4, \ldots$.
  - $\binom{K}{3}$ three-factor interaction terms corresponding to $x_1x_2x_3, x_1x_2x_4, \ldots$.
  - $\vdots$
  - $\binom{K}{K} = 1$ K-factor interaction term corresponding to $x_1x_2 \cdots x_K$.

  In total, there are $\sum_{j=0}^{K} \binom{k}{j} = 2^K$ terms.

EXAMPLE 6.2.2: 2^1 Example

$$\beta_0 + \beta_1 x_1$$

EXAMPLE 6.2.3: 2^2 Example

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

EXAMPLE 6.2.4: 2^3 Example

$$\beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\text{main effects}} + \underbrace{\beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3}_{\text{two-factor interactions}} + \underbrace{\beta_{123} x_1 x_2 x_3}_{\text{three-factor interaction}}$$

Estimation

- Estimation of the β's is carried out by:

  ↪ Ordinary least squares (in the case of linear regression).

  ↪ Maximum likelihood (in the case of logistic regression).

- In both cases there is a one-to-one connection between the β estimates and the expressions for the main and interaction effects. Note that both $\widehat{\text{Effect}}$'s below are calculated using the "intuitive" formulas described above.
  - Continuous response:
    $$\widehat{\text{Effect}} = 2\hat{\beta}$$
  - Binary response:
    $$\widehat{\text{Effect}} = e^{2\hat{\beta}}$$
    where β is the regression coefficient corresponding to the effect of interest.
- Recall the **Toy Example**:
  - The linear predictor for that experiment is:
    $$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$
    where x_1 and x_2 correspond to factors A and B respectively.
  - The linear regression model is:
    $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \quad \text{for } i = 1, 2, \dots, N = n2^K$$
    which can be written in matrix-vector notation as:
    $$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
    where
    $$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 2 \\ 1 \\ 3 \\ 1 \\ 2 \\ 5 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_{12} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{12} \end{bmatrix}$$
    - The columns of X are orthogonal!
      - This is why we code x's using ± 1's.
  - The least squares estimate of $\boldsymbol{\beta}$ is:
    $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}$$
    Therefore,
    $$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 12 \end{bmatrix} \rightarrow (\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{12}\mathbf{I}_4$$
    $$\mathbf{X}^\top \boldsymbol{Y} = \begin{bmatrix} \mathbf{1}^\top \boldsymbol{Y} \\ \boldsymbol{x}_1^\top \boldsymbol{Y} \\ \boldsymbol{x}_2^\top \boldsymbol{Y} \\ \boldsymbol{x}_3^\top \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i:\text{A}^+} Y_i - \sum_{i:\text{A}^-} Y_i \\ \sum_{i:\text{B}^+} Y_i - \sum_{i:\text{B}^-} Y_i \\ \sum_{i:\text{A}^+\cap\text{B}^+} Y_i + \sum_{i:\text{A}^-\cap\text{B}^-} Y_i - \sum_{i:\text{A}^-\cap\text{B}^+} Y_i - \sum_{i:\text{A}^+\cap\text{B}^-} Y_i \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{Y} = \begin{bmatrix} 5/2 \\ 10/12 \\ -1/6 \\ -1/2 \end{bmatrix}$$

- Notice that:

$$2\hat{\boldsymbol{\beta}} = \begin{bmatrix} 5 \\ 10/6 \\ -1/3 \\ -1 \end{bmatrix} = \begin{bmatrix} 2\bar{y} \\ \widehat{\text{ME}}_{\text{A}} \\ \widehat{\text{ME}}_{\text{B}} \\ \widehat{\text{IE}}_{\text{AB}} \end{bmatrix}$$

  This is the same as what we calculated using the "intuitive" formulas. This is not a coincidence!

- In general:
  - $\boldsymbol{Y}$ is an $N \times 1$ vector of response observations.
  - $\boldsymbol{\varepsilon}$ is an $N \times 1$ random vector of error terms.
  - $\boldsymbol{\beta}$ is a $2^K \times 1$ vector of regression coefficients.
  - $\mathbf{X}$ is the $N \times 2^K$ **model matrix** containing plus and minus ones.
    - Each column represents a different effect (i.e., term in the linear predictor).
    - Interaction columns are obtained from element-wise multiplication of the main effects columns involved in the interaction.
    - ± 1's in the rows are defined in terms of the design matrix (i.e., which condition the response observation was observed in).
    - The columns of the model matrix are always orthogonal $\rightarrow \mathbf{X}^\top \mathbf{X} = N\mathbf{I}_{2^K} \rightarrow (\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{N}\mathbf{I}_{2^K}$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix.
  - Due to the orthogonality of the model matrix, any effect (whether main or interaction) is estimated as:

$$\widehat{\text{Effect}} = 2\hat{\beta} = \frac{\boldsymbol{x}^\top \boldsymbol{Y}}{n2^{K-1}}$$

    where $\boldsymbol{x}$ is the column of $\mathbf{X}$ corresponding to the effect of interest, and β is the corresponding regression coefficient.
    - This should make sense: β's in ordinary regression are interpreted as the expected change in response resulting from a unit increase in x. Here, we care about two-unit increases (i.e., low $\rightarrow$ high, $-1 \rightarrow +1$).

Hypothesis Testing

- The significance of main and interaction effects is determined by testing hypotheses that set the relevant β's equal to 0.

* But now, because each effect is represented by just a single term, the hypotheses of interest involve just a single β.

- In the **Toy Example**, if we wanted to determine the significance of factor A, we simply test

$$\mathbf{H}_0\text{: } \beta_1 = 0$$

  or if we want to determine whether the A:B interaction is significant, we test

$$\mathbf{H}_0\text{: } \beta_{12} = 0.$$

- Hypotheses like these are tested with ordinary significance tests for individual regression coefficients.
  - t-tests in the case of linear regression.
  - Z-tests in the case of logistic regression.

  * All tests can be done in the full model (linear predictor with 2^K terms).

→ But if for some reason we still want to test hypotheses about several β's simultaneously, we can compare full and reduced models with the usual:

- Partial F-tests in the case of linear regression.
- Likelihood ratio tests in the case of logistic regression.

6.2.3 The Credit Card Example

- To illustrate a complete analysis of a 2^K factorial experiment, we consider an example from Montgomery (2019) in which an experiment was performed to test new ideas to improve the conversion rate of credit card offers. For this example, the response is binary — indicating whether an individual signed up for a credit card as a result of the offer — and so an analysis based on logistic regression is performed.
- A 2^4 factorial experiment was carried out to investigate four factors and their influence on credit card sign-ups. The four factors and each of their levels are summarized in Table 6.1.

Table 6.1: Factors and levels for the credit card example.

Factor	Low (−)	High (+)
Annual Fee (x_1)	Current	Lower
Account-Opening Fee (x_2)	No	Yes
Initial Interest Rate (x_3)	Current	Lower
Long-term Interest Rate (x_4)	Low	High

- The $2^4 = 16$ unique combinations of these factor levels produced 16 experimental conditions, each of which was assigned $n = 7500$ units. Practically speaking, 16 credit card offers were devised (one corresponding to each condition) and each was mailed to 7500 customers. The design matrix and a summary of the conversion rates are provided in Table 6.2.

Table 6.2: Design matrix and response summary for the 2^4 factorial credit card experiment.

Condition	Factor 1	Factor 2	Factor 3	Factor 4	Sign-ups	Conversion Rate
1	−1	−1	−1	−1	184	2.45%
2	+1	−1	−1	−1	252	3.36%
3	−1	+1	−1	−1	162	2.16%
4	+1	+1	−1	−1	172	2.29%
5	−1	−1	+1	−1	187	2.49%
6	+1	−1	+1	−1	254	3.39%
7	−1	+1	+1	−1	174	2.32%
8	+1	+1	+1	−1	183	2.44%
9	−1	−1	−1	+1	138	1.84%
10	+1	−1	−1	+1	168	2.24%
11	−1	+1	−1	+1	127	1.69%
12	+1	+1	−1	+1	140	1.87%
13	−1	−1	+1	+1	172	2.29%
14	+1	−1	+1	+1	219	2.92%
15	−1	+1	+1	+1	153	2.04%
16	+1	+1	+1	+1	152	2.03%

- Using this data we fit a logistic regression model with the following linear predictor:

$$\begin{aligned}
\text{Intercept} &\rightarrow \beta_0 + \\
\text{ME's} &\rightarrow \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{14}x_1x_4 + \beta_{23}x_2x_3 + \beta_{24}x_2x_4 + \beta_{34}x_3x_4 + \\
\text{2FI's} &\rightarrow \beta_8x_1x_5 + \beta_9x_2x_5 + \beta_{10}x_3x_5 + \beta_{11}x_4x_5 + \beta_{12}x_1x_6 + \beta_{13}x_2x_6 + \beta_{14}x_3x_6 + \\
\text{3FI's} &\rightarrow \beta_{123}x_1x_2x_3 + \beta_{124}x_1x_2x_4 + \beta_{134}x_1x_3x_4 + \beta_{234}x_2x_3x_4 + \\
\text{4FI} &\rightarrow \beta_{1234}x_1x_2x_3x_4
\end{aligned}$$

- The regression output associated with this model is:

```
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.739697   0.019342 -193.347  < 2e-16 ***
x1           0.080845   0.019342    4.180 2.92e-05 ***
x2          -0.106211   0.019342   -5.491 3.99e-08 ***
x3           0.058248   0.019342    3.011  0.00260 **
x4          -0.108086   0.019342   -5.588 2.29e-08 ***
x1:x2       -0.055164   0.019342   -2.852  0.00434 **
x1:x3       -0.004794   0.019342   -0.248  0.80426
x2:x3       -0.006967   0.019342   -0.360  0.71868
x1:x4       -0.013178   0.019342   -0.681  0.49566
x2:x4        0.010625   0.019342    0.549  0.58280
x3:x4        0.038079   0.019342    1.969  0.04899 *
x1:x2:x3    -0.009646   0.019342   -0.499  0.61799
x1:x2:x4     0.010629   0.019342    0.550  0.58265
x1:x3:x4    -0.002543   0.019342   -0.131  0.89539
x2:x3:x4    -0.020946   0.019342   -1.083  0.27885
x1:x2:x3:x4 -0.009496   0.019342   -0.491  0.62347
```

  - The p-value for $\mathbf{H}_0$: $\beta_1 = 0$ is 2.92×10^{-5}.
  - The p-value for $\mathbf{H}_0$: $\beta_2 = 0$ is 3.99×10^{-8}.
  - The p-value for $\mathbf{H}_0$: $\beta_3 = 0$ is 0.00260.
  - The p-value for $\mathbf{H}_0$: $\beta_4 = 0$ is 2.29×10^{-8}.
  - The p-value for $\mathbf{H}_0$: $\beta_{12} = 0$ is 0.00434.
  - The p-value for $\mathbf{H}_0$: $\beta_{34} = 0$ is 0.04899.

- We now know which main and interaction effects are significant (i.e., all main effects, x_1:x_2, and x_3:x_4).
  - Let's use main and interaction effect plots to help us interpret these effects.
    - In Figure 6.3, the conversion rate is maximized when the annual fee is low, no account opening fee, and when initial and long-term interest rates are low.
    - In Figure 6.4:
      - When an account opening fee exists, the effect of the annual fee is less than if there wasn't an account opening fee.
      - As long as long-term interest is low, the effect of initial interest rate is not large.

- [R Code] `2^4_Factorial_Example`

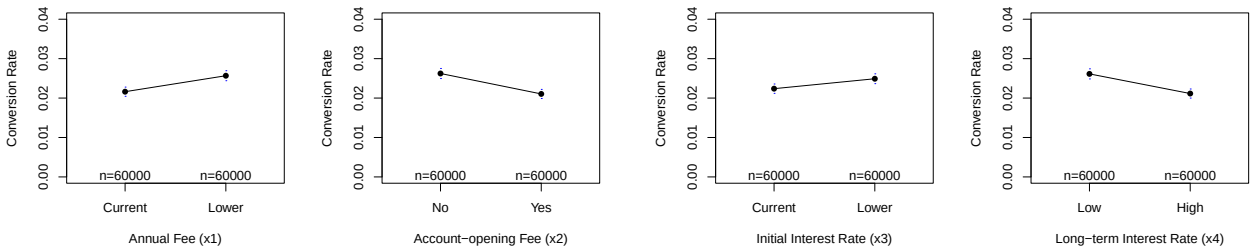


Figure 6.3: Main effect plots for the credit card example.

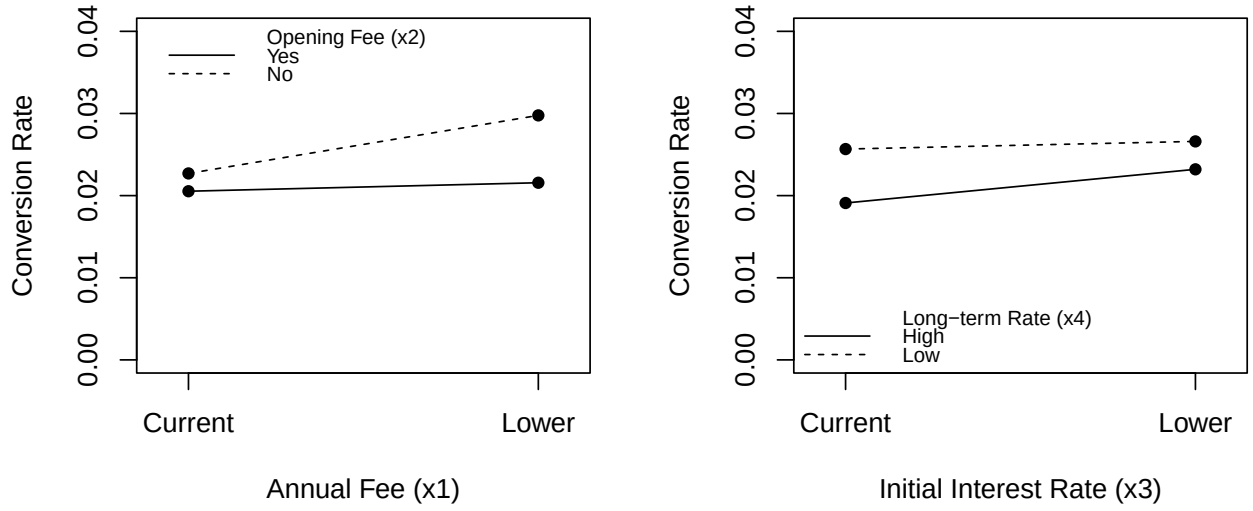


Figure 6.4: Interaction effect plots for the credit card example.

Chapter 7

2^{K-p} FRACTIONAL FACTORIAL EXPERIMENTS

WEEK 10

Let $p \in \mathbf{Z}^+$, $1 \leq p < K$, and $2^{K-p} > K$.

* A 2^K factorial experiment is an economical special case of a general factorial experiment.

  → It minimizes the number of levels being investigated.

  → Thus, it reduces the overall number of experimental conditions.

- However, 2^K can still be a very large number of conditions even for moderate K.

  EXAMPLE 7.0.1

  If $K = 8$, then $2^K = 256 = m$.

* In a 2^{K-p} fractional factorial experiment we also investigate K factors but in just a fraction of the conditions.

  Specifically, $(1/2)^p$ as many since $m = 2^{K-p}$.

→ Rather than experimenting with all 2^K conditions, we specially select 2^{K-p} of them.

  – When $p = 1$, we investigate K factors in half as many conditions (i.e., "one-half fraction").

  – When $p = 2$, we investigate K factors in a quarter of the conditions (i.e., "one-quarter fraction").

- The value p dictates the degree of *fractioning* and is typically chosen to:

  → Minimize the number of experimental conditions m, given a fixed number of design factors K, or

  → Maximize the number of design factors K, given a fixed number of conditions m.

  * Goal: explore as many factors as possible in as few conditions as possible.

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

  → $\sim 40\%$ of ME's were significant.

  → $\sim 10\%$ of 2FI's were significant.

  → $\sim 5\%$ of 3+FI's were significant.

- But consider the linear predictor from the full 2^K factorial experiment. There are:
  - An intercept: β_0.
  - K main effect terms.
  - $\binom{K}{2}$ two-factor interaction terms.
  - $\binom{K}{3}$ three-factor interaction terms.
  - $\vdots$
  - $\binom{K}{K} = 1$ K-factor interaction term.

  This is a total of $\sum_{k=1}^{K} \binom{K}{k} = 2^K - 1$ estimated effects and just $K + \binom{K}{2}$ of these are main effects and two-factor interactions.

EXAMPLE 7.0.2

If $K = 8$, then $\binom{K}{2} = 28$, $2^K - 1 = 255$, and so $255 - 28 - 8 = 219$ is the number of 3+FI's.

- In light of effect sparsity, it is a waste of resources to estimate higher order interaction terms.
  * It would be a better use of resources to estimate the main effects and low-order interactions of a larger number of factors.
- So how do we choose *which* 2^{K-p} conditions to run?
- Consider the following three examples as motivation:

EXAMPLE 7.0.3: The 2^{3-1} Example

In this example we consider a one-half fraction of the 2^3 design which explores $K = 3$ factors (A, B, C) in $m = 4$ conditions rather than 8. The design matrix associated with a full 2^3 design and a visualization of the full 2^3 design are shown below. The question of primary interest is: *which* $m = 4$ conditions do we choose for the 2^{3-1} experiment?

Condition	Factor A	Factor B	Factor C
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

EXAMPLE 7.0.4: The 2^{4-1} Example

In this example we consider a one-half fraction of the 2^4 design which explores $K = 4$ factors (A, B, C, D) in $m = 8$ conditions rather than 16. The design matrix associated with a full 2^4 design and a visualization of the full 2^4 design are shown below. Similar to the 2^{3-1} example, the

question of primary interest is: *which* $m = 8$ conditions do we choose for the 2^{4-1} experiment?

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

EXAMPLE 7.0.5: The 2^{5-2} Example

In this example we consider a one-quarter fraction of the 2^5 design which explores $K = 5$ factors (A, B, C, D, E) in $m = 8$ conditions rather than 32. The design matrix associated with a full 2^5 design and a visualization of the full 2^5 design are shown below. Similar to the previous two examples, the question of primary interest is: *which* $m = 8$ conditions do we choose for the 2^{5-2}

experiment?

Condition	Factor A	Factor B	Factor C	Factor D	Factor 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
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

7.1 Designing 2^{K-p} Fractional Factorial Experiments

Given 2^K conditions to choose from, how do we choose which 2^{K-p} conditions to experiment with?

7.1.1 Aliasing

- The first step in constructing a 2^{K-p} fractional factorial experiment is to write out the model matrix (when $n = 1$) for a *full* 2^{K-p} design.

EXAMPLE 7.1.1: 2^{3-1} Example

The model matrix (when $n = 1$) for a full 2^2 design with factors A and B is shown below:

Condition	I	A	B	AB = C
1	+1	−1	−1	+1
2	+1	+1	−1	−1
3	+1	−1	+1	−1
4	+1	+1	+1	+1

  - Rather than asking "which 4 conditions from a full 2^3 design do I run?" we now ask "in which of the four conditions in a full ^{2} design should I run factor C at its low versus high levels?"

- We use the ± 1's in the AB interaction column to dictate, for a given condition, whether to run factor C at its low or high levels.
- Conditions 1 and 4 have AB = +1, so C will run at its high level.
- Conditions 2 and 3 have AB = −1, so C will be run at its low level.
- What results is a prescription for experimenting with $K = 3$ factors in $2^{3-1} = 4$ conditions?
- This is a 2^{3-1} fractional factorial design. We visualize it as follows:
- **Principal fraction**: The conditions selected by associating the levels of C with the ± 1's in the AB column.

  Red points.
- **Complementary fraction**: The conditions selected by associating the levels of C with −AB.

  Green points — this is also a 2^{3-1} fractional factorial design.
- What we did there is called **aliasing**: associate the main effect of a new factor with an existing condition. We aliased the main effect of C with the AB interaction. Notation: C = AB.
- We call C = AB the **design generator**.
- When we do this, we **confound** the interaction effect with the main effect of the new factor.

  ↪ These effects cannot be separately estimated.
- In an ordinary 2^2 experiment with factors A and B, the AB column of the model matrix is used to estimate IE_{AB}.
  - But do to the C = AB, the AB column now jointly quantifies the main effect of C *and* the AB interaction effect.

$$\widehat{\text{IE}}_{\text{AB}} = \frac{\bar{y}_{\text{A}^+\cap\text{B}^+} + \bar{y}_{\text{A}^-\cap\text{B}^-}}{2} - \frac{\bar{y}_{\text{A}^-\cap\text{B}^+} + \bar{y}_{\text{A}^+\cap\text{B}^-}}{2}$$

$$\begin{aligned}\widehat{\text{ME}}_{\text{C}} &= \bar{y}_{\text{C}^+} - \bar{y}_{\text{C}^-} \\ &= \frac{\bar{y}_{\text{A}^+\cap\text{B}^+} + \bar{y}_{\text{A}^-\cap\text{B}^-}}{2} - \frac{\bar{y}_{\text{A}^-\cap\text{B}^+} + \bar{y}_{\text{A}^+\cap\text{B}^-}}{2} \\ &= \widehat{\text{IE}}_{\text{AB}}\end{aligned}$$

  This calculation now estimates both the main effect of C and the AB interaction effect simultaneously. We can't separate them!
- This is the price we pay for using fewer conditions than what is prescribed by the full 2^K design.

  ↪ We cannot separately estimate confounded/aliased effects. It turns out this problem doesn't only impact C and AB.

7.1.2 The Defining Relation

- In the 2^{3-1} example, we aliased C with the AB interaction.
  - We saw that this means the main effect of C and the AB interaction effect are confounded.
  - However, the aliasing (and hence confounding) doesn't stop there.

* Upon closer inspection we find that the main effect of A and B are now also aliased with interaction effects.

- This becomes evident when we consider the **defining relation**:

$$\text{Design Generator} \rightarrow \text{C} = \text{AB} \rightarrow \text{C} \times \text{C} = \text{AB} \times \text{C}$$
$$\text{I} = \text{ABC}$$

- This may be used to uncover all aliases by multiplying it by any effect:

$$\text{A} \times \text{I} = \text{A}^2\text{BC} \quad \text{B} \times \text{I} = \text{AB}^2\text{C} \quad \text{C} = \text{AB}$$
$$\text{A} = \text{IBC} \qquad \text{B} = \text{AC}$$
$$\text{A} = \text{BC}$$

- Every main effect is aliased with a two factor interaction.

Introducing aliasing anywhere causes confounding everywhere.

EXAMPLE 7.1.2: 2^{4-1} Example

- To construct this factorial design we consider the model matrix (when $n = 1$) associated with a full 2^3 design:

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

- We need to choose one interaction column to alias a new factor D with.

  ↪ This tell us when to run factor D at low vs. high.

  * We could choose AB, AC, BC, or ABC. Which one is the *right* choice?
    · We choose D = ABC because the effect sparsity principle tells us that high order interactions are less likely to be significant.
  * The complete aliasing structure is:

$$\text{Defining relation} \rightarrow \text{I} = \text{ABCD}$$
$$\text{A} = \text{BCD}$$
$$\text{B} = \text{ACD}$$
$$\text{C} = \text{ABD}$$

$$D = ABC$$

$$AB = CD$$

$$AC = BD$$

$$BC = AD$$

- What would have happened if we had chosen D = AB or D = AC or D = BC as design generators instead of D = ABC?
- Which one of these designs is the best?

  ↪ We'll come back to this.

EXAMPLE 7.1.3: 2^{5-2} Example

- In addition to choosing an alias for factor D like we just did with the 2^{4-1} design, we also need to choose an alias for factor E.

  We now have $p = 2$ design generators D = ABC, E = BC.

  * The 2^{5-2} fractional factorial design that results from these choices is visualized below:
  * In general, the number of design generators will always equal p.
- These design generators give rise to the following defining relation:

  $$\left\{\begin{array}{c} D = ABC \rightarrow I = ABCD \\ E = BC \rightarrow I = BCE \end{array}\right\} \rightarrow I = ABCD = BCE = ABCD \times BCE = AB^2C^2DE = ADE$$

  Therefore, I = ABCD = BCE = ADE.
- As usual, this may be used to determine the complete aliasing structure:

  $$A = BCD = ABCE = DE$$

  $$B = ACD = CE = ABDE$$

  $$C = ABD = BE = ACDE$$

  $$D = ABC = BCDE = AE$$

  $$E = ABCDE = BC = AD$$

  $$AB = CD = ACE = BDE$$

  $$AC = BD = ABE = CDE$$

  * Every effect is aliased (i.e., confounded) with 3 other effects.
  * In general, the number of effects aliased with a given effect is $2^p - 1$.
  * Thus, in a 2^{K-p} fractional factorial design, every effect estimate actually jointly quantifies 2^p effects.

- **SUMMARY**: To design a 2^{K-p} fractional factorial experiment, you must:
  * Look at the model matrix (with $n = 1$) for a full 2^{K-p} design with $K - p$ factors.
  * Choose p interaction columns to alias an additional p factors with.

* Use the ± 1's in these columns to dictate, for each condition, whether the p additional factors are run at their low or high level.

But how do we know *which* interactions to choose?

7.1.3 Resolution

* Due to the confounding that results from aliasing a new main effect with an existing interaction, 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 because separately estimating significant effects is desirable.
  - High order interaction terms (that are unlikely to be significant) are good choices for aliases.

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

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

  What is the smallest order interaction your main effects are aliased with? Resolution is that number +1.

- The easiest way to determine R is by looking at the defining relation.
  - Each of the terms in the equivalence is referred to as a *word*.

  → The length of the shortest word is the resolution of the design.

  * The defining relations for 2^{3-1}, 2^{4-1}, and 2^{5-2} designs are:

$$I = ABC$$

$$I = ABCD$$

$$I = ABCD = BCE = ADE$$

  For 2^{3-1} and 2^{5-2} designs: shortest word has length 3. Therefore, it's a Resolution III design.
  For 2^{4-1} design: shortest word has length 4. Therefore, it's a Resolution IV design.
  These designs are described succinctly as:

$$2^{3-1}_{\text{III}}, 2^{4-1}_{\text{IV}}, 2^{5-2}_{\text{III}}$$

- General notation: 2^{K-p}_{R} where
  - 2: number of levels.
  - K: number of factors.
  - p: degree of fractioning.
  - R: resolution.

* In general, higher resolution designs are to be preferred over lower resolution designs.
  - Resolution IV and V designs are to be preferred over a resolution III design.
    ↪ Because the resolution IV and V designs do not alias main effects with two-factor interactions.

- The resolution of a fractional factorial experiment is determined by two things:
  1. The degree of fractioning desired (i.e., the size of p relative to K).

2. The design generators chosen for aliasing.

* Given K and p, we should choose design generators that *maximize resolution.*

- Let us return to the 2^{4-1} example.

Design Generator	Defining Relation
D = ABC	I = ABCD
D = AB	I = ABD
D = AC	I = ACD
D = BC	I = BCD

  * The generator D = ABC is the best because it gives rise to a resolution IV design.

- Another way to justify the maximum resolution criterion is by the **projective property** of fractional factorial designs.

  * A resolution R fractional factorial design can be projected into a full factorial design on *any subset* of $R-1$ factors.

  – Let's visualize this with the 2^{3-1} design:

  – This property can be exploited when analyzing the experimental data.

    ↪ If $R-1$ (or fewer) factors have significant main effects, they can be analyzed as full factorial designs without confounding.

* Maximizing R maximizes the size of the projected full factorial design.

7.1.4 Minimum Aberration

- The maximum resolution criterion is one way to choose design generators.

- But what if several choices lead to the same resolution? Then how do we choose?
  Minimum Aberration Criterion.

- Consider a 2^{7-2}_{IV} design which is resolution IV and explores $K = 7$ factors in $m = 32$ conditions.

  – Three design generator configurations that all give rise to a 2^{7-2}_{IV} design are shown below:

Design	Design Generators	Defining Relation
1	F = ABC, G = ABD	I = ABCF = ABDG = CDFG
2	F = ABC, G = CDE	I = ABCF = CDEG = ABDEFG
3	F = ABCD, G = ABCE	I = ABCDF = ABCEG = DEFG

    The shortest word length is 4, therefore $R = 4$.

  – How should we choose among these? Is one better than the others?

    * We can compare these designs on the basis of how many words of length 4 appear in the defining relation. Word lengths: $(4, 4, 4)$, $(4, 4, 6)$, $(5, 5, 4)$.

    * Design 3 minimizes this number, and hence minimizes the number of main effects aliased with the lowest-order interactions.

  * In general, for a given resolution R the **minimum aberration design** is one which minimizes the number of minimum-length words in the defining relation.

  → These designs are preferred since they minimize the number times main effects are aliased with the lowest order ($(R-1)$-factor) interactions.

7.2 Analyzing 2^{K-p} Fractional Factorial Experiments

* We have seen that 2^{K-p} fractional factorial designs are a clever alternative to full 2^K designs for purposes of factor screening.
  - They still explore K factors, but in just a *fraction* of the conditions required by a full 2^K design.
    ↪ $(1/2)^p$ as many.
  - This is made possible by *aliasing* and reliance on the *principle of effect sparsity*.
  - However, this aliasing causes *confounding* which can complicate conclusions.
    ↪ Can't separately estimate confounded effects.
  - We try to mitigate the negative side effects of confounding by choosing designs with *maximum resolution* and *minimum aberration*.

- It turns out that the analysis of a 2^{K-p} fractional factorial design is not very different from the analysis of a full 2^K factorial design.
  - We visually summarize effects of interest via main and interaction effect plots.
  - Regression models are used to test hypotheses of the form (to determine whether a given effect is significantly different from zero):
    $$\mathbf{H}_0\text{: } \beta = 0.$$
    → t-tests in linear regression.
    → Z-tests in logistic regression.

- Now we have to deal with confounding. Recall: two effects that are **confounded** cannot be separately estimated.

  → Just 2^{K-p} effects (and hence β's) can be estimated. The number of β's estimable is the number of conditions. However, there are 2^K effects, so we're not estimating all of them.

  → Each of these β's jointly quantifies 2^p different effects.

  - It is therefore important to know the complete aliasing structure of the design to be fully aware of *which* effects are confounded.

- Accounting for this confounding is particularly important when interpreting effect estimates and evaluating their significance.

EXAMPLE 7.2.1: The 2^{5-2}_{III} Example

Suppose we find that the main effect of factor A is significant. What can we conclude?

$$\text{A} = \text{BCD} = \text{ABCE} = \text{DE}$$

We can't be 100% certain the significance of the effect is solely due to the main effect of A.

- It could be that ME_{A} is significant.
- It could be that IE_{BCD} is significant.
- It could be that IE_{ABCE} is significant.
- It could be that IE_{DE} is significant.
- Or they could all be significant.
- Or individually none of these are significant, but in aggregate they are.

* The uncertainty surrounding this interpretation motivates why we avoid confounding effects that are likely to be significant with other ones that are also likely to be significant.

  ↪ Maximizing resolution and minimizing aberration should help here.

* Next week, in an illustrative example of a 2^{8-4} fractional factorial experiment, we will demonstrate how to estimate and carefully interpret the effects of the factors involved.

WEEK 11

7.2.1 The Chehalem Example

- Here we consider an example from Montgomery (2019) in which a 2^{8-4} fractional factorial experiment was used in the production of wine to study the influence of a variety of factors on a particular vintage of Pinot Noir.
- In this experiment $K = 8$ factors were investigated each at two levels (the factors and their levels are shown in Table 7.1) which, if a full factorial experiment was used, would have required 256 conditions.

Table 7.1: Factors and levels for the wine example.

Factor	Low (−)	High (+)
Pinot Noir clone (A)	Pommard	Wädenswil
Oak type (B)	Allier	Tronçais
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)

- To keep the experiment as small as possible a 2^{8-4}_{IV} fractional factorial experiment was performed that required only 16 conditions (i.e., 16 different wines).
- The response variable in this case is the rating of the wine as determined by 5 raters.
- Thus, 16 different wines were produced (based on the 16 unique combinations of these factors' levels) and $n = 5$ raters tasted and rated each of them (low scores are good, large scores are bad). The design matrix and a summary of the response is provided in Table 7.2.
- Because the response variable in this setting is continuous, we use linear regression to analyze the data from this experiment.
- 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 would be the model with 4 main effects, 6 two-factor interactions, 4 three-factor interactions and 1 four-factor interaction:

```
            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 ***
A:D          -0.8750     0.2658  -3.293 0.001619 **
```

Table 7.2: Design matrix and response summary for the 2^{8-4} fractional factorial wine experiment.

Condition	A	B	C	D	E = BCD	F = ACD	G = ABC	H = ABD	Average Rating = $\bar{y}$
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

```
B:C           0.4500     0.2658   1.693 0.095261 .
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 .
```

- But this output does not involve the factors E, F, G or H — it only directly references factors A, B, C and D.
- This is because of confounding.
  - The BCD interaction estimate also corresponds to the main effect of E.
  - The ACD interaction estimate also corresponds to the main effect of F.
  - The ABC interaction estimate also corresponds to the main effect of G.
  - The ABD interaction estimate also corresponds to the main effect of H.

* While we cannot technically separate these effects, we assume that the three-factor interactions are negligible, and hence any significant effect observed is due to the aliased main effect.

- The same model summary from above is shown in below, but this time with factors E, F, G and H referenced instead of the three-factor interactions:

```
            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 ***
```

* All main effects are significant except for factor H (fermentation temperature).
* Also, the AC, AD, and AH interactions are significant. However, factors D, E, F, G, are most influential, so it's more likely that the significance of these two-factor interactions is driven by aliased interactions involving D, E, F, G.

$$\text{AC} = \text{DF}, \qquad \text{AD} = \text{EG}, \qquad \text{AH} = \text{FG}$$

* We will therefore speculate that it's the DF, EG, and FG interactions that are important.

- Figure 7.1 depicts the main effect plots for all eight factors.

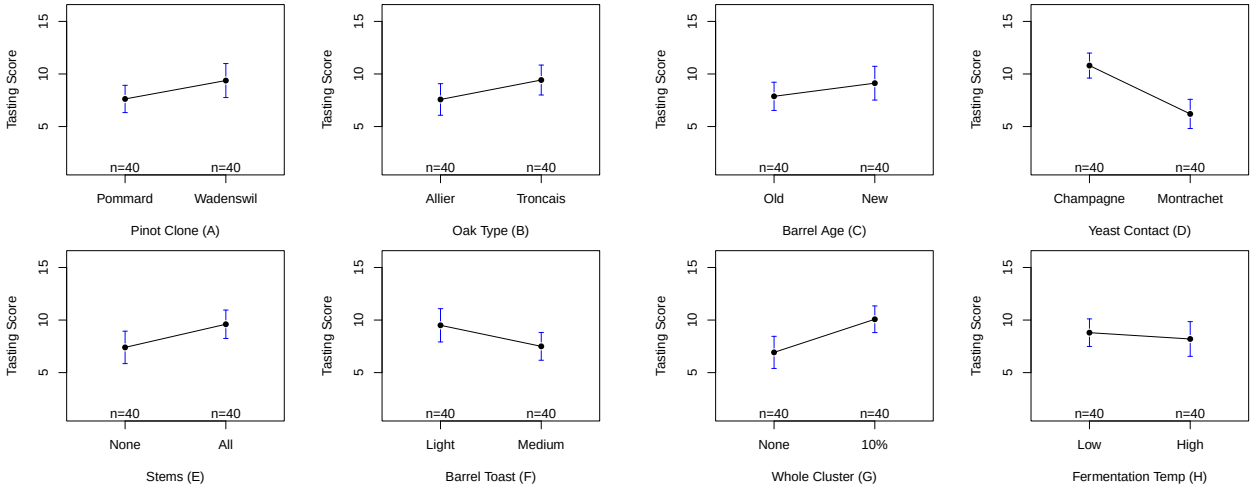


Figure 7.1: Main effect plots for the wine example.

- Figure 7.2 depicts the interaction effect plots for the three significant interactions.

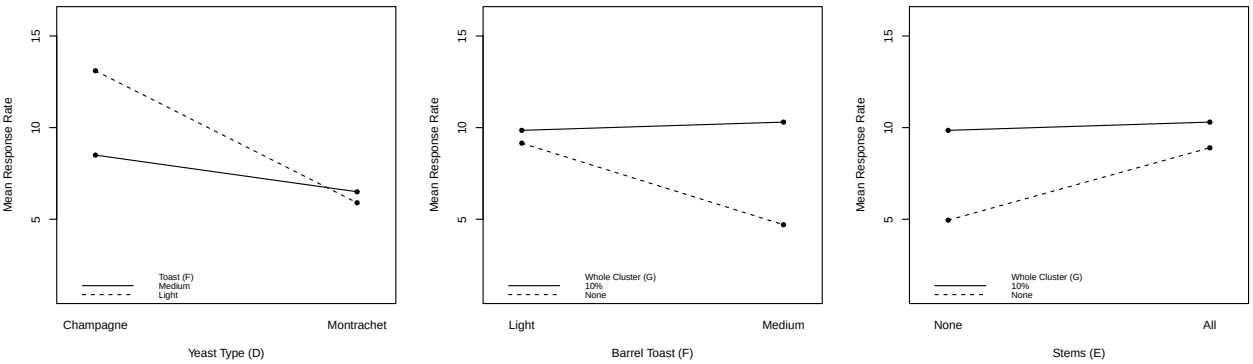


Figure 7.2: Interaction effect plots for the wine example.

  - If yeast type is Montrachet, then the effect of Barrel toast is minimal.
  - The effect of whole clusters is minimal if Barrel toast is light.
  - Effect of the whole cluster is minimal when all stems are used.

- [R Code] `Fractional_Factorial_Example`

Chapter 8

RESPONSE SURFACE METHODOLOGY

* Effective experimentation is sequential
  - Information gained in one experiment can help to inform future experiments.
  - This is the philosophy of **response surface methodology**.

- We have seen that the primary purpose of screening experiments is to identify which among numerous factors are the ones that significantly influence the response variable.

  ↪ Phase 1: Factor screening with two-level designs.

* Now we discuss how screening experiments may be followed-up by further experiments whose primary purpose is response optimization.

  ↪ We use the method of **steepest ascent/descent** (phase 2) and **response surface designs** (phase 3) to locate optimal settings of the factors that were identified as significant in the screening phase.

  * Phase 4: Confirmation.

8.1 Overview of Response Optimization

Coded Factors

- Here we consider $K' \leq K$ design factors which are a subset of the K factors investigated during the screening phase.
- The set of possible values these factors can take on is referred to as the **region of operability**.

  * It is this region that we explore and in which we run our experiments to determine the *optimal* operating condition.

  EXAMPLE 8.1.1

  If the design factor is discount amount, then the region of operability is $[0, 100]$.

- Although this region specifies acceptable factor values in their natural units (such as dollars, minutes, percent, etc.), we typically work on a transformed scale.
- Just like in the regression models used in the experiments, we represent each factor by a coded variable x that takes on the values -1 and $+1$ when the factor is at its *low* and *high* levels.

  * When the factor is categorical this coding is arbitrary.
  - When the factor is numeric the coding arises through the following transformation:

$$x = \frac{U - (U_{\mathbf{H}} + U_{\mathbf{L}})/2}{(U_{\mathbf{H}} - U_{\mathbf{L}})/2}$$

    * U is any value of the factor in natural units.
      · $U_{\mathbf{H}}$ and $U_{\mathbf{L}}$ are "high" and "low" levels of the factor recorded in natural units.
    * x is the transformed version of U in coded units.
      · If $U = U_{\mathbf{H}}$, then $x = +1$, and if $U = U_{\mathbf{L}}$, then $x = -1$.

EXAMPLE 8.1.2

Assume we're experimenting with discount amount where $U_{\mathbf{L}} = 20\%$ and $U_{\mathbf{H}} = 50\%$. Then $x = (U - 35)/15$ may be used to convert from natural units to coded units.

* If $U = 50$, then $x = +1$.
* If $U = 20$, then $x = -1$.
* If $U = 40$, then $x = 1/3$.
* If $U = 60$, then $x = 5/3$.

  - This equation may also be inverted allowing for conversion from the coded units back to the natural units as follows:

$$U = x \times \frac{U_{\mathbf{H}} - U_{\mathbf{L}}}{2} + \frac{U_{\mathbf{H}} + U_{\mathbf{L}}}{2}$$

    * Translating from coded to natural units is especially useful when translating the location of the optimum from coded units to natural units.

EXAMPLE 8.1.3

Optimal $x = 1.45 \rightarrow U = 1.45(15) + 35 = 56.75\%$.

- Adopting this notation, the objective of response optimization may be stated as determining the value of $\boldsymbol{x} = (x_1, x_2, \ldots, x_{K'})^\top$ (and hence $\boldsymbol{U} = (U_1, U_2, \ldots, U_{K'})^\top$) at which we expect the response to be optimized.

The Models

- The goal of response optimization may be achieved via **response surface experimentation** where one seeks to characterize the relationship between the expected response $\mathbb{E}[Y]$ and the K' design factors.
- In the case of a continuous response, we may write this relationship generally as:

$$\mathbb{E}[Y] = f(x_1, x_2, \ldots, x_{K'})$$

  and in the case of a binary response:

$$\log\left(\frac{\mathbb{E}[Y]}{1 - \mathbb{E}[Y]}\right) = f(x_1, x_2, \ldots, x_{K'})$$

- In both cases, the function $f(x_1, x_2, \ldots, x_{K'})$ represents the *true* but *unknown* **response surface**.
  - If we knew $f(\,\cdot\,)$, we could easily find the values of $x_1, \ldots, x_{K'}$ that optimize it.

- Because $f(\,\cdot\,)$ is unknown, we must fit models that approximate this surface. As usual, we use linear and logistic regression.
- Although many models may be used to approximate the response surface we exploit Taylor's Theorem and use *low-order polynomials*:
  - First-order model:
$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{K'} x_{K'}$$
    ↪ Main effects only.
  - First-order + Interaction model:
$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<\ell} \beta_{j\ell} x_j x_\ell$$
    ↪ Main effect + two-factor interactions.
  - Second-order model:
$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<\ell} \beta_{j\ell} x_j x_\ell + \sum_{j=1}^{K'} \beta_{jj} x_j^2$$
    ↪ Quadratic effects in addition to main effects and two-factor interactions.
- Examples of such response surfaces (for $K' = 2$) are visualized in Figure 8.1:

* We must acknowledge that the approximation of $f(x_1, x_2, \ldots, x_{K'})$ by η (regardless of whether η is first-order or second-order) is likely to be poor when considered across the entire x-space.

→ However, in the small localized region of an experiment, such low-order polynomials should well-approximate $f(\,\cdot\,)$.

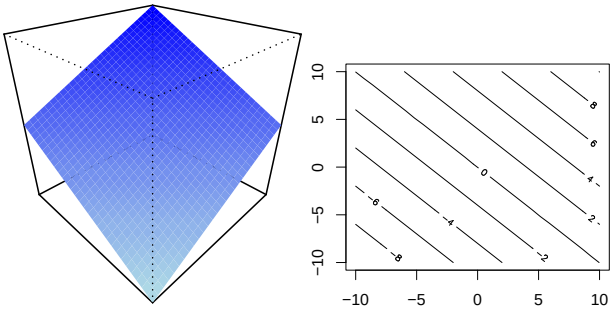
→ Which model is appropriate is dictated by the goal of the experiment.

- In the context of factor screening we saw that first-order and first-order-plus-interaction models suited our needs.
- But in order to identify maxima/minima we require the second-order model as it is capable of modelling concavity/convexity.
- Therefore, second-order models are used for response surface optimization.

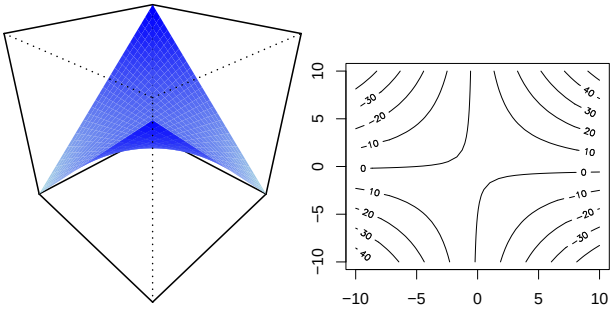
8.2 Method of Steepest Ascent/Descent

- We use the method of steepest of ascent/descent to determine *roughly* where in the x-space the optimum lies.
  - Hence, this tells us where a response surface design and a second-order model would be most useful.
  - * We want to find the "vicinity" of the optimum.

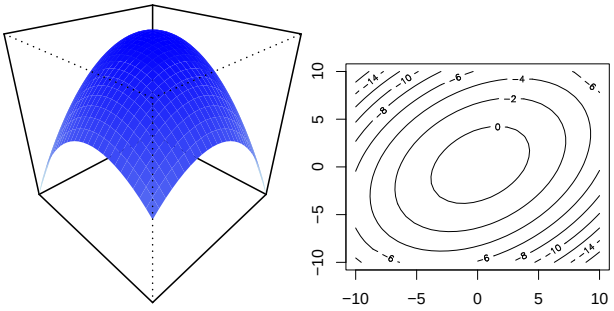
* The method is gradient-based and designed to identify the direction that when traversed moves you toward the optimum as quickly as possible.



(a) $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (Plane)



(b) $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$ (Twisted Plane)



(c) $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$

Figure 8.1: Example 3D surface and 2D contour plots of first-order (Figure 8.1a), first-order-plus-interaction (Figure 8.1b) and second-order (Figure 8.1c) response surfaces.

8.2.1 The Path of Steepest Ascent/Descent

- We use a $2^{K'}$ (or $2^{K'-p}$) factorial experiment to estimate a *first-order response surface*:

$$\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{K'} x_{K'}$$

- The gradient of this surface is then calculated:

$$\boldsymbol{g} = \nabla \hat{\eta} = \left(\frac{\partial \hat{\eta}}{\partial x_1}, \frac{\partial \hat{\eta}}{\partial x_2}, \ldots, \frac{\partial \hat{\eta}}{\partial x_{K'}} \right)^\top$$

  - This gradient defines the **path of steepest ascent/descent** (i.e., the direction of steepest increase/decrease on the fitted surface).
  - If maximizing the response is of interest, then we should ascend the surface by moving in the direction of $+\boldsymbol{g}$:

$$\boldsymbol{x}' = \boldsymbol{x} + \lambda \boldsymbol{g} \tag{1}$$

  - If minimizing the response is of interest, then we should descend the surface by moving in the direction of $-\boldsymbol{g}$:

$$\boldsymbol{x}' = \boldsymbol{x} - \lambda \boldsymbol{g} \tag{2}$$

  - With a fixed step size λ we move from $\boldsymbol{x}$ to $\boldsymbol{x}'$.
  - We typically define the step size as:

$$\lambda = \frac{\Delta x_j}{|\hat{\beta}_j|}$$

    * Pick factor j, the one you know the most about, or the one that is hardest to manipulate.
    * Δx_j is the step size of factor j in coded units.
    * $\hat{\beta}_j$ is the estimated coefficient corresponding to factor j in the estimated first-order response model.

Steepest Ascent/Descent Algorithm

1. The first condition along the path of steepest ascent/descent is at the origin of the x-space $\boldsymbol{x}_0 = (0, 0)^\top$ (i.e., the centre of the $2^{K'}$ factorial design that was used to fit $\hat{\eta}$). Data is collected and the metric of interest is calculated.
2. Then the step size λ is determined.
3. The location of the next condition is determined by formula (1) in the case of maximization and (2) in the case of minimization. Data is collected and the metric of interest is calculated.
4. Repeat Step 3 until incremental improvements in the MOI cease.
5. Return to the location of the best MOI value and *test for curvature*.

   → If the test for curvature suggests that you are not yet in the vicinity of the optimum, fit a new first-order model and repeat Steps 1 to 4.

   → If the test for curvature suggests that you are in the vicinity of the optimum, use a response surface design to fit a full second-order model and hence precisely identify the coordinates of the optimum.

8.2.2 Checking for Curvature

- A test for quadratic curvature is an important component of the method of steepest ascent/descent.
  * The presence of quadratic curvature signifies that you are in the vicinity of the optimum.
- Such a test is possible when a $2^{K'}$ factorial experiment is augmented with a **centre point** condition.
  - The centre point condition is defined (in coded units) as $x_1 = x_2 = \cdots = x_{K'} = 0$.
  - Located at the centre of the cuboidal region defined by the $2^{K'}$ factorial conditions.

* The data arising from a $2^{K'}$ factorial design is insufficient to estimate a second-order linear predictor.
  * We are able to estimate the main effects and the two-factor interaction effects, but *not* the quadratic effects.

- With the addition of the centre point condition, one *additional* effect may be estimated: the **pure quadratic effect**.

$$\beta_{\mathsf{PQ}} = \sum_{j=1}^{K'} \beta_{jj}$$

- A test of $\mathbf{H}_0$: $\beta_{\mathsf{PQ}} = 0$ is a test of *overall curvature*.

EXAMPLE 8.2.1: $K' = 2$

- When $K' = 2$, the second-order linear predictor is:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

- In a 2^2 factorial design plus a centre point, we have the following five unique experimental conditions: $(x_1, x_2) \in \{(-1,-1), (+1,-1), (-1,+1), (+1,+1), (0,0)\}$ which respectively give rise to five unique variants of the linear predictor, which we define as:

$$\eta_{\mathsf{LL}} = \beta_0 - \beta_1 - \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{\mathsf{HL}} = \beta_0 + \beta_1 - \beta_2 - \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{\mathsf{LH}} = \beta_0 - \beta_1 + \beta_2 - \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{\mathsf{HH}} = \beta_0 + \beta_1 + \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{\mathsf{C}} = \beta_0$$

* With only these five conditions, we cannot separately estimate β_{11} and β_{22}, but we *can* estimate $\beta_{\mathsf{PQ}} = \beta_{11} + \beta_{22}$.

- Notice that:

$$\beta_{\mathsf{PQ}} = \frac{\eta_{\mathsf{LL}} + \eta_{\mathsf{HL}} + \eta_{\mathsf{LH}} + \eta_{\mathsf{HH}}}{4} - \eta_{\mathsf{C}}$$

- The estimate is therefore:

$$\hat{\beta}_{\mathsf{PQ}} = \frac{\hat{\eta}_{\mathsf{LL}} + \hat{\eta}_{\mathsf{HL}} + \hat{\eta}_{\mathsf{LH}} + \hat{\eta}_{\mathsf{HH}}}{4} - \hat{\eta}_{\mathsf{C}}$$

  * If this difference, and hence $\hat{\beta}_{\mathsf{PQ}}$, is small then it suggests that the response values observed in the factorial conditions are similar to those observed in the centre point condition and hence that there isn't significant curvature in the response surface.
  * If $\hat{\beta}_{\mathsf{PQ}}$ is very different from zero, it suggests that there is significant quadratic curvature.

- We formally test $\mathbf{H}_0$: $\hat{\beta}_{\mathbf{PQ}}$ using t-tests (or Z-tests) in a linear (or logistic) regression model that has linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$$

where

$$x_{\mathbf{PQ}} = \begin{cases} 1 & (x_1, x_2) \in \{(-1,-1), (+1,-1), (-1,+1), (+1,+1)\} \\ 0 & (x_1, x_2) = (0,0) \end{cases}$$

which indicates whether a response observation came from a factorial condition or the centre point condition.

- If $\beta_{\mathbf{PQ}}$ is significantly different from 0 then it suggests that *both* β_{11} and β_{22} are significantly non-zero, and therefore that there is significant quadratic curve.

- For general K', we conduct a $2^{K'}$ factorial experiment with a centre point and then test for curvature using a regression model with linear predictor:

$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<\ell} \beta_{j\ell} x_j x_\ell + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$$

where now

$$\beta_{\mathbf{PQ}} = \sum_{j=1}^{K'} \beta_{jj}$$

and $x_{\mathbf{PQ}}$ is again a binary indicator indicating whether a response value was observed in a factorial condition or the centre point condition.

* No matter the value of K', the pure quadratic effect is always represented by a single term in the model.

  – As such, the test for curvature is always a test of $\mathbf{H}_0$: $\beta_{\mathbf{PQ}} = 0$ and is carried out with ordinary t-tests in a linear regression and Z-tests in a logistic regression.

- The intuitive estimate for $\beta_{\mathbf{PQ}}$ in the $K' = 2$ case also generalizes:

$$\hat{\beta}_{\mathbf{PQ}} = \hat{\bar{\eta}}_{\mathbf{F}} - \hat{\eta}_{\mathbf{C}}$$

where

  → $\hat{\bar{\eta}}_{\mathbf{F}}$ is the average if the estimated linear predictor values in the factorial conditions.
  → $\hat{\eta}_{\mathbf{C}}$ is the estimated linear predictor value at the centre point.

- **IMPORTANT**: This test assumes that all the β_{jj}'s, $j = 1, 2, \ldots, K'$, have the same sign.

  – If they didn't, then it's possible that significantly large β_{jj}'s could cancel each other out, making $\beta_{\mathbf{PQ}} = \sum_{j=1}^{K'} \beta_{jj}$ close to zero.

    ↪ We are misled into thinking that we are not in the vicinity of the quadratic curvature, even when we are.

  * This assumption is fine as long as the experiment is not conducted near a saddle point on the response surface.

    ↪ This problem can only be identified by separately estimating each of the quadratic effects.

8.2.3 The Netflix Example

- Here we illustrate the *method of steepest descent* using a modified version of the hypothetical Netflix experiment from your final project.
- We focus on the Preview Length factor (defined analogously as in your project) and a Preview Size factor (which corresponds to the size of the enlarged window a preview is played in).
- We begin with a 2^2 factorial experiment with a centre point condition. The factor levels in coded and natural units are shown in Table 8.1.

Table 8.1: Average browsing time by condition in the $2^2 + \mathbf{CP}$ Netflix experiment.

Condition	Preview Length (s)	x_1	Preview Size	x_2	Average Browsing Time (min)
1	90	-1	0.2	-1	22.163
2	120	$+1$	0.2	-1	22.197
3	90	-1	0.5	$+1$	20.223
4	120	$+1$	0.5	$+1$	21.982
5	105	0	0.35	0	22.046

- Prior to embarking down the path of steepest descent, a curvature test was performed to determine whether this experimental region was already in the vicinity of the optimum.
- The linear regression model with linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$$

was fit. The resulting output is shown in Table 8.2.

Table 8.2: Summary of $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$.

	Estimate	Std. Error	t-value	$\Pr(>\lvert t\rvert)$
(Intercept)	22.046	0.195	112.979	$<2.222 \times 10^{-16}$
x_1	0.448	0.098	4.595	4.777×10^{-6}
x_2	-0.539	0.098	-5.524	4.036×10^{-8}
$x_{\mathbf{PQ}}$	-0.405	0.218	-1.855	6.386×10^{-2}
$x_1{:}x_2$	0.431	0.098	4.419	1.076×10^{-5}

- To begin the method of steepest descent procedure, we use the aforementioned data to fit the first order regression model with linear predictor:

$$\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

The model summary is shown in Table 8.3.

Table 8.3: Summary of $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$.

	Estimate	Std. Error	t-value	$\Pr(>\lvert t\rvert)$
(Intercept)	21.722	0.088	246.852	$<2.222 \times 10^{-16}$
x_1	0.448	0.098	4.556	5.714×10^{-6}
x_2	-0.539	0.098	-5.478	5.202×10^{-8}

- Figure 8.2 depicts the contours of the estimated first-order response surface.

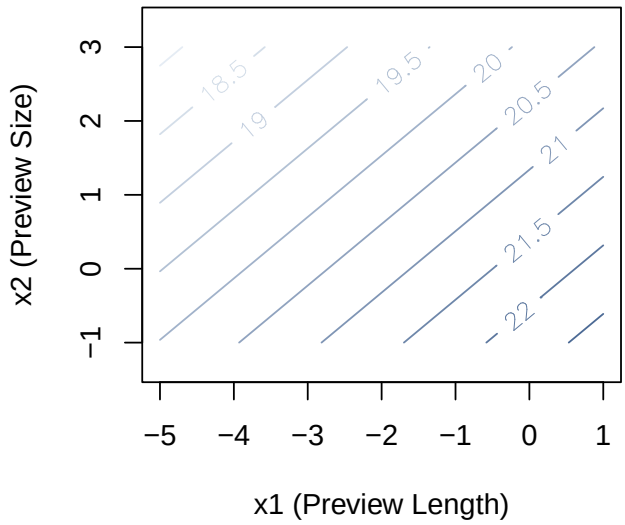


Figure 8.2: Contour plot for the estimated first-order response surface for the Netflix experiment.

- We calculate the gradient:

$$\boldsymbol{g} = (\hat{\beta}_1, \hat{\beta}_2)^\top = \left(\frac{\partial \hat{\eta}}{\partial x_1}, \frac{\partial \hat{\eta}}{\partial x_2} \right)^\top = (0.448, -0.539)^\top$$

* This path of steepest descent is depicted by the dashed black line in Figure 8.3. The red dots signify the experimental conditions conducted along this path, beginning from the centre point $(x_1, x_2) = (0, 0)$.

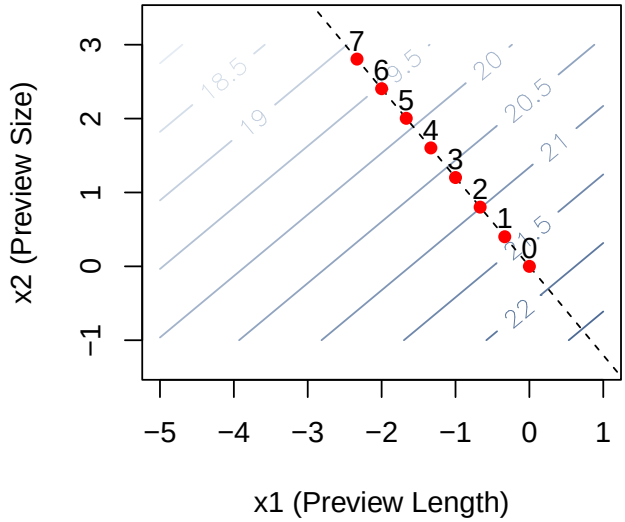


Figure 8.3: Contour plot for the path of steepest descent for the Netflix experiment.

- The locations in coded and natural units for each of these conditions are provided in Table 8.4.
  - Note that a step size of:

$$\lambda = \frac{\Delta x_1}{|\hat{\beta}_1|} = \frac{1/3}{|0.448|}$$

    was used, where the value $1/3$ was chosen to ensure steps of 5 seconds in Preview Lengths.
- The average browsing time in each condition is reported in Table 8.4 and visualized in Figure 8.4.

Table 8.4: Average browsing time along the path of steepest descent.

Condition	Preview Length (s)	x_1	Preview Size	x_2	Average Browsing Time (min)
0	105	0	0.350	0	21.998
1	100	−0.333	0.410	0.401	21.672
2	95	−0.667	0.470	0.801	21.258
3	90	−1.000	0.530	1.202	19.105
4	85	−1.333	0.590	1.603	18.245
5	80	−1.667	0.651	2.004	15.944
6	75	−2.000	0.711	2.404	14.889
7	70	−2.333	0.771	2.805	17.160

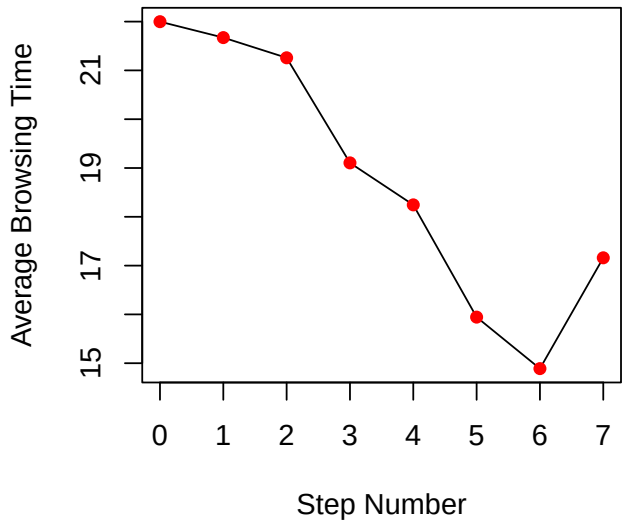


Figure 8.4: Average browsing time along the path of steepest descent.

* Clearly that Step 6 corresponded to the lowest observed average browsing time, and so we should perform another test of curvature in this region to determine whether we've reached the vicinity of the optimum.

- In order to do so, another 2^2 factorial experiment with a centre point needs to be run. The factor levels in coded and natural units for this next experiment are shown in Table 8.5.

Table 8.5: Average browsing time by condition in the second $2^2 + \mathbf{CP}$ Netflix experiment.

Condition	Preview Length (s)	x_1	Preview Size	x_2	Average Browsing Time (min)
1	60	−1	0.6	−1	14.571
2	90	+1	0.6	−1	18.173
3	60	−1	0.8	+1	18.220
4	90	+1	0.8	+1	18.655
5	75	0	0.7	0	14.831

- Once again we fit a linear regression model with linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$$

- The resulting output is shown in Table 8.6.

Table 8.6: Summary of $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{\mathbf{PQ}} x_{\mathbf{PQ}}$.

	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
(Intercept)	14.831	0.189	78.397	$<2.22 \times 10^{-16}$
x_1	1.009	0.095	10.668	$<2.22 \times 10^{-16}$
x_2	1.033	0.095	10.920	$<2.22 \times 10^{-16}$
$x_{\mathbf{PQ}}$	2.573	0.212	12.167	$<2.22 \times 10^{-16}$
$x_1{:}x_2$	−0.792	0.095	−8.372	$<2.22 \times 10^{-16}$

  - Reject $\mathbf{H}_0$: $\beta_{\mathbf{PQ}} = 0$, therefore we conclude there is a quadratic curve.

- [R Code] `PSTD_example`

WEEK 12

8.3 Response Surface Experiments

* Effective experimentation is sequential: information gained in one experiment can help to inform future experiments.

- Screening experiments are used to identify which among numerous factors are the ones that significantly influence the response variable.

- We follow these up with further experimentation where the goal is **response optimization**.
  - **Method of Steepest Ascent/Descent**.
  - **Response Surface Designs**.

- In these investigations, response optimization requires investigating and characterizing response surfaces of the form:

$$\mathbb{E}[Y] = f(x_1, x_2, \ldots, x_{K'})$$

(for a continuous response) and

$$\log\left(\frac{\mathbb{E}[Y]}{1-\mathbb{E}[Y]}\right) = f(x_1, x_2, \ldots, x_{K'})$$

(for a binary response).

Finding the Optimum

- Supposing that sufficient data is collected and the second-order model may be fitted, we obtain the estimated response surface.

$$\hat{\eta} = \hat{\beta}_0 + \sum_{j=1}^{K'} \hat{\beta}_j x_j + \sum_{j<\ell} \hat{\beta}_{j\ell} x_j x_\ell + \sum_{j=1}^{K'} \hat{\beta}_{jj} x_j^2$$

- This expression may be re-written in vector-matrix notation as:

$$\hat{\eta} = \hat{\beta}_0 + \boldsymbol{x}^\top \boldsymbol{b} + \boldsymbol{x}^\top \mathbf{B} \boldsymbol{x}$$

where

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K'} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_{K'} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \hat{\beta}_{11} & \frac{1}{2}\hat{\beta}_{12} & \cdots & \frac{1}{2}\hat{\beta}_{1K'} \\ \frac{1}{2}\hat{\beta}_{12} & \hat{\beta}_{22} & \cdots & \frac{1}{2}\hat{\beta}_{2K'} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\hat{\beta}_{1K'} & \frac{1}{2}\hat{\beta}_{2K'} & \cdots & \hat{\beta}_{K'K'} \end{bmatrix}$$

  - $\boldsymbol{x}$ is a $K' \times 1$ vector of specific factor values.
  - $\boldsymbol{b}$ is a $K' \times 1$ vector of the estimates of the main effect coefficients.
  - $\mathbf{B}$ is a $K' \times K'$ symmetric matrix of second-order effect estimates (i.e., the second-order interactions and quadratic effects).

- In order to find the value of $\boldsymbol{x} = (x_1, x_2, \ldots, x_{K'})^\top$ that maximizes/minimizes the expected response, we must find the **stationary point** of the estimated response surface.

- The stationary point is:

$$\boldsymbol{x}_{\mathbf{s}} = -\frac{1}{2}\mathbf{B}^{-1}\boldsymbol{b}$$

which is found by solving:

$$\frac{\partial \hat{\eta}}{\partial \boldsymbol{x}} = \boldsymbol{b} + 2\mathbf{B}\boldsymbol{x} = \mathbf{0}$$

- The optimal expected response is:

$$\widehat{\mathbb{E}[Y]} = \hat{\eta}_{\mathbf{s}} = \hat{\beta}_0 + \frac{1}{2}\boldsymbol{x}_{\mathbf{s}}^\top \boldsymbol{b}$$

in the case of linear regression and

$$\widehat{\mathbb{E}[Y]} = \frac{\exp[\hat{\eta}_{\mathbf{s}}]}{1+\exp[\hat{\eta}_{\mathbf{s}}]} = \frac{\exp[\hat{\beta}_0 + \frac{1}{2}\boldsymbol{x}_{\mathbf{s}}^\top \boldsymbol{b}]}{1+\exp[\hat{\beta}_0 + \frac{1}{2}\boldsymbol{x}_{\mathbf{s}}^\top \boldsymbol{b}]}$$

in the case of logistic regression.

- For practical implementation of this solution, the stationary point $\boldsymbol{x}_{\mathbf{s}}$ must be translated into optimal operating conditions in natural units $\boldsymbol{U}_{\mathbf{s}}$ using the following conversion formula:

$$U = x \times \frac{U_{\mathbf{H}} - U_{\mathbf{L}}}{2} + \frac{U_{\mathbf{H}} + U_{\mathbf{L}}}{2}$$

* However, for us to be confident that $\boldsymbol{x}_{\mathbf{s}}$ indeed optimizes $f(\,\cdot\,)$, we must be confident that η and, in particular, that $\hat{\eta}$ adequately represents $f(\,\cdot\,)$.
    * Since we only expect the second-order approximation to be adequate in a small localized region, it is important that this small localized region contains the true optimum.
    - It is quite unlikely that the values of $x_1, x_2, \ldots, x_{K'}$ considered in the screening phase are close to the optimum.
    * This is why we needed the method of steepest ascent/descent.
        * This intermediate phase of experimentation helped us determine roughly where the region of the optimum lies.

8.3.1 The Central Composite Design

* The goal of a response surface experiment is to be able to fit a full second-order response surface model.
    - This requires estimating $(K'+1)(K'+2)/2$ coefficients.

- Several such designs exist (i.e., "response surface"), but here we study one in particular: the **central composite design** (CCD).
- A CCD is typified by three different types of experimental conditions:
    i. **two-level** factorial conditions,
    ii. a **centre point** condition, and
    iii. **axial**, or *star*, conditions.
- In other words,
    i. The factorial conditions constitute a full $2^{K'}$ factorial design.
    ii. The centre point condition sits at $x_1 = x_2 = \cdots = x_{K'} = 0$ in the centre of the factorial ones.
    iii. The axial conditions sit 'outside' of the factorial ones at $\pm a$ on each of the K' factors' axes. Note that a is defined in coded units.
- When investigating K' factors the central composite design therefore requires $2^{K'} + 2K' + 1$ distinct experimental conditions.
- These designs may be visualized geometrically as we see in the figures below, for $K' = 1, 2, 3$.
- The design matrices that give rise to these designs (for $K' = 1, 2, 3$) are shown in Table 8.7.
- **Choosing** a:
    - The value of a is determined by the experimenter, and may be chosen to balance both practical and statistical concerns.
    - The experimenter must be mindful of the constraints imposed by the region of operability and whether the natural-unit counterpart to a is something inconvenient/infeasible.
    - Barring practical constraints, two common choices for a are $a = 1$ and $a = \sqrt{K'}$.
- $a = 1$:
    - The CCD reduces to a $3^{K'}$ design.
    - It is referred to as *face-centred central composite design*.
    - A benefit is that it requires just 3 (not 5) levels for every factor.
    - Another benefit is that it is a cuboidal design and so it inherits some usual conveniences associated with orthogonal cuboidal designs.

Table 8.7: Design matrices associated with central composite designs on $K' = 1$ (left), $K' = 2$ (middle) and $K' = 3$ (right) factors.

Condition	x_1
1	-1
2	$+1$
3	$-a$
4	$+a$
5	0

Condition	x_1	x_2
1	-1	-1
2	$+1$	-1
3	-1	$+1$
4	$+1$	$+1$
5	$-a$	0
6	$+a$	0
7	0	$-a$
8	0	$+a$
9	0	0

Condition	x_1	x_2	x_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	$+1$	$+1$	$+1$
9	$-a$	0	0
10	$+a$	0	0
11	0	$-a$	0
12	0	$+a$	0
13	0	0	$-a$
14	0	0	$+a$
15	0	0	0

  - You might choose $a = 1$ if the region of the optimum is in a corner of the region of operability, and hence $a = 1$ keeps the experimental conditions inside the region of operability.

- $a = \sqrt{K'}$:
  - In this design the axial conditions are at an equal distance from the centre point as the factorial conditions.
  - Such a design is referred to as *spherical* since it places all axial and factorial conditions on a "hyper" sphere of radius $\sqrt{K'}$.
  - The benefit of such equal spacing is that it ensures that the estimate of the response surface at each condition is equally precise.
    ↪ Designs with this property are called rotatable.

* No matter the choice of $a > 0$, the CCD facilitates estimation of the full second-order response surface model, and hence identification of the optimum.

8.3.2 The Lyft Example

- We illustrate the design and analysis of a central composite experiment in the context of a common ride-sharing problem.
- Suppose that Lyft is interested in designing a promotional offer that maximizes ride-bookings during an experimental period.

→ Previous screening experiments evaluated the influence of discount amount, discount duration, ride type, time-of-day, and the method of dissemination. It was found that the most important factors were discount amount (x_1) and discount duration (x_2).

- A previous steepest ascent exercise also suggested that the optimal discount duration is somewhere in the vicinity of 4.5 d and the optimal discount amount is somewhere in the vicinity of 50 %.
- To find optimal values of these factors a follow-up two-factor central composite design was run in order to fit a second-order response surface model.
- The experimental conditions (in both coded and natural units) are shown in Table 8.8.

Table 8.8: Booking rate by condition in the Lyft experiment.

Condition	Discount Amount (%)	x_1	Discount Duration (d)	x_2	Booking Rate
1	25	-1	2	-1	0.71
2	75	$+1$	2	-1	0.32
3	25	-1	7	$+1$	0.71
4	75	$+1$	7	$+1$	0.35
5	85	$+1.4$	4.5	0	0.53
6	15	-1.4	4.5	0	0.50
7	50	0	8	$+1.4$	0.26
8	50	0	1	-1.4	0.78
9	50	0	4.5	0	0.72

- **NOTE**: that the experimenters had intended to perform axial conditions with $a = \sqrt{2}$, but the corresponding discount amounts and discount durations were:

$$(14.645\,\%, 85.355\,\%) \text{ and } (0.964\,\text{d}, 8.036\,\text{d})$$

  In the interest of defining experimental conditions with practically convenient levels they opted for $a = 1.4$ yielding the discount amounts and durations shown in Table 8.8.

- $n = 500$ users were then randomized into each of these $m = 9$ conditions and for each user, whether they booked a ride in the experimentation period was recorded.
  - The booking rates in each condition are also shown in Table 8.8.

- The output from the fitted second-order logistic regression model is shown in Table 8.9.

Table 8.9: Summary of second-order logistic regression model.

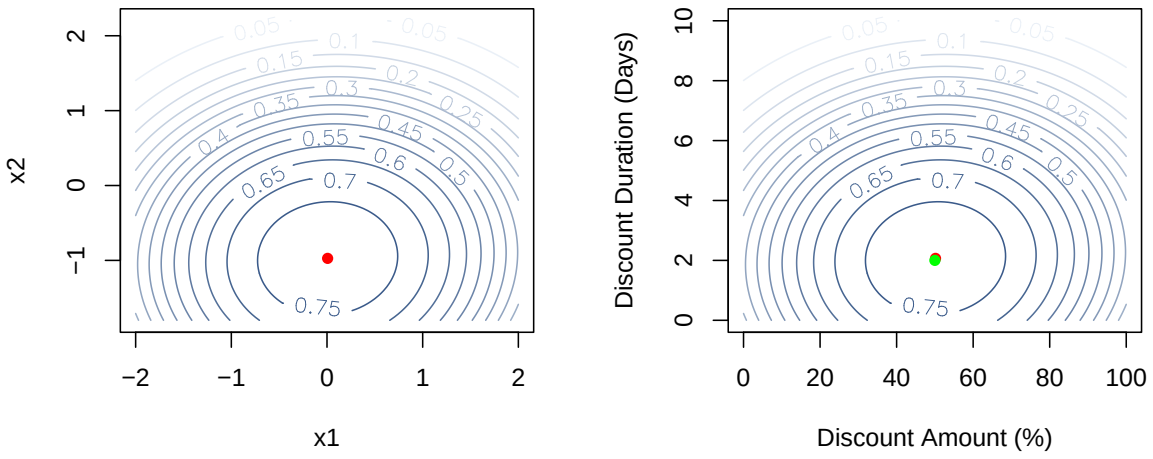
	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
(Intercept)	0.943	0.100	9.474	$<2.222 \times 10^{-16}$
x_1	0.039	0.033	1.174	2.406×10^{-1}
x_2	-0.807	0.036	-22.612	$<2.222 \times 10^{-16}$
$\mathrm{I}(x_1^2)$	-0.442	0.058	-7.637	2.221×10^{-14}
$\mathrm{I}(x_2^2)$	-0.414	0.059	-6.989	2.770×10^{-12}
$x_1{:}x_2$	0.034	0.048	0.700	4.840×10^{-1}

- Contour plots of the fitted response surface are shown in Figure 8.5.

- The stationary point for this second-order model is located (in coded units) at $x_1 = 0.007$, $x_2 = -0.973$ using $\boldsymbol{x_s} = -\frac{1}{2}\mathbf{B}^{-1}\boldsymbol{b}$.
  - In the natural units this corresponds to a discount rate of 50.164 % that lasts for 2.067 d.
  - The predicted booking rate at this point is 0.792, with a 95 % prediction interval given by $(0.769, 0.814)$.

- A slightly less optimal but more practically feasible promotion would be a 50 % discount lasting 2 d (this is what Lyft should move forward with into a confirmation phase).
  - This achieves a booking rate of 0.792 with a 95 % prediction interval of $(0.769, 0.814)$.

- [R Code] `CCD_example`



(a) (b)

Figure 8.5: 2D contour plots of the second-order Lyft model. Figure 8.5a: Coded-Unit Factor Space. Figure 8.5b: Natural-Unit Factor Space.

8.4 RSM with Qualitative Factors

- What do you do if you have ≥ 1 categorical factors in addition to our numeric factor(s)?

* Everything that has been discussed thus far with respect to central composite designs and response surface optimization has assumed that the factors under experimentation are quantitative (i.e., the factors have numeric levels).

$\rightarrow$ In the presence of one or more categorical factors we need to take additional care.

- When categorical factors are present, we can think of there being different response surfaces that relate the response to the quantitative factors at each of the factorial combinations of the categorical factors' levels.
- Thus, the general strategy is to enumerate all factorial combinations of the categorical factors' levels and employ the methods of response surface methodology independently within each.
  - Perform the method of steepest ascent/descent independently on each surface
  - Perform CCDs independently on each surface
  - Independently fit second-order models for each surface
  - Independently identify the stationary point on each surface

* Among all the candidate surfaces, the one with the most optimal optimum is the 'winner.'

  - The factor levels (numeric and categorical) that gave rise to it should be defined as the optimal operating conditions.

* This investigation should now be followed up by a response surface experiment so that a full second-order model may be fit and the optimum identified.

EXAMPLE 8.4.1

uppose we have two numeric factors x_1 and x_2, and two categorical factors x_3 [3 levels ($\mathbf{L}, \mathbf{M}, \mathbf{H}$)] and x_4 [2 levels ($\mathbf{L}, \mathbf{H}$)]. There are six combinations of the categorical factors levels, and then at each one of those six configurations, you can imagine a response surface that relates the expected response to x_1 and x_2 holding at x_3 and x_4 fixed at that particular specification.

1. (x_3, x_4) at $(\mathbf{L}, \mathbf{L})$, then do RSM on x_1 and x_2.
2. (x_3, x_4) at $(\mathbf{M}, \mathbf{L})$, then do RSM on x_1 and x_2.
3. (x_3, x_4) at $(\mathbf{H}, \mathbf{L})$, then do RSM on x_1 and x_2.
4. (x_3, x_4) at $(\mathbf{L}, \mathbf{H})$, then do RSM on x_1 and x_2.
5. (x_3, x_4) at $(\mathbf{M}, \mathbf{H})$, then do RSM on x_1 and x_2.
6. (x_3, x_4) at $(\mathbf{H}, \mathbf{H})$, then do RSM on x_1 and x_2.

Optional: [R Code] `Visualizing_response_surfaces`