Experimental Design

STAT 430

Spring 2021 (1215)

LaTeXer: *Cameron Roopnarine*
Instructor: *Nathaniel Stevens*

May 31, 2021

Contents

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 the experiment is meant to investigate.

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

The factors are thought to 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 the response variable is measured.

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 composed of 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 the factors are purposefully controlled in order 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, data is collected 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 subjects are forced 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

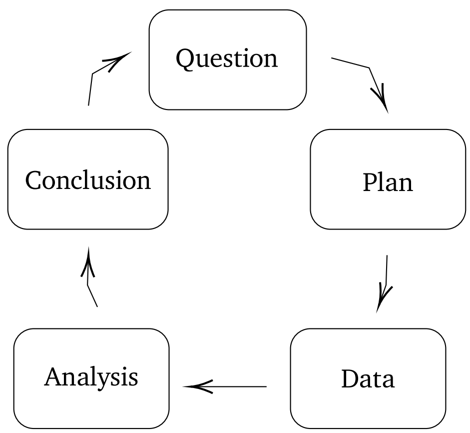

Figure 1.1: QPDAC Cycle

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, the experiment is designed 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 the response variable is measured 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, the data are collected according to the **Plan**. It is extremely important that this step be done correctly; the suitability and effectiveness of the analysis relies on the data being collected correctly. 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: units are assigned 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.
  - *Simple Ratio Mismatch Test*: check whether the observed sample ratios match what would be expected if assignment was truly done at random.
    - ⋆ 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, the **Data** are statistically analyzed 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 the data was collected correctly, this step should be straightforward.

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

- These conclusions should then be clearly communicated 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 experimental units are *selected for inclusion* in the experiment and the manner in which they are *assigned 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 more than one unit is assigned 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 is needed?
  - How big a sample size is needed?
  - Power analysis + sample size calculations will help answer this.

DEFINITION 1.4.4: Blocking

Blocking is the mechanism by which the nuisance factors are controlled for.

- 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

  Consider an email promotion 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 that it is known that the ‘open rate’ is also influenced by the time of the day and the day of the week that the email is sent.

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:

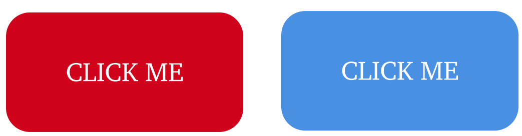

Figure 2.1: Canonical Button Colour Test.

    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, such a question is phrased as a statistical hypothesis that we test using the data collected from the experiment.
  - $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 \neq \theta_2$ (two-sided).
  - $\mathbf{H}_0$: $\theta_1 \leq \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 > \theta_2$ (one-sided).
  - $\mathbf{H}_0$: $\theta_1 \geq \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 < \theta_2$ (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 conclusion in the context of a hypothesis test.
- Recall that by design, either $\mathbf{H}_0$ or $\mathbf{H}_\text{A}$ is true. Thus 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 $\mathbf{H}_\text{A}$: 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 the defendant is assumed innocent until proven guilty. Formally,
$\mathbf{H}_0$: the defendant is innocent versus $\mathbf{H}_\text{A}$: 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 defined as $\alpha = \mathbb{P}(\text{Type I Error})$.

DEFINITION 2.0.7: Power

The **power** of a test is defined as $1-\beta$ 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 the response variable of interest is measured on a continuous scale.
- We assume that the response observations collected in the two conditions follow normal distributions, and in particular

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

  - 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*:

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

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

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

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 $\mathbf{H}_A$: $\sigma_1^2 \neq \sigma_2^2$.
  - $\mathbf{H}_0$: $\sigma_1^2/\sigma_2^2 = 1$ versus $\mathbf{H}_A$: $\sigma_1^2/\sigma_2^2 \neq 1$.
- *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 how user engagement is influenced by ad frequency.
- 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 being tested here is:

$$\mathbf{H}_0\text{: } \mu_1 \leq \mu_2 \text{ versus } \mathbf{H}_A\text{: } \mu_1 > \mu_2$$

- The data summaries are:
  - $n_1 = 500$, $\hat{\mu}_1 = \bar{y}_1 = 4.92$, $\hat{\sigma}_1 = s_1 = 0.96$.
  - $n_2 = 500$, $\hat{\mu}_2 = \bar{y}_2 = 3.05$, $\hat{\sigma}_2 = s_2 = 0.99$.

F-test:

- $t = \hat{\sigma}_1^2/\hat{\sigma}_2^2 = 0.96^2/0.99^2 = 0.938$.
- p-value $= \mathbb{P}(T \leq 0.938) + \mathbb{P}(T \geq 1/0.938) = 0.4720$ 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:

- $\hat{\sigma}^2 = \dfrac{499(0.96)^2 + 499(0.99)^2}{998} = 0.9793^2$.
- $t = \dfrac{4.92 - 3.05}{0.9793\sqrt{\frac{1}{500} + \frac{1}{500}}} = 30.1$.
- p-value $= \mathbb{P}(T \geq 30.1) = 1.84 \times 10^{-142} \approx 0$ 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

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

- 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

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

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

$$\mathbf{H}_0\colon \pi_1 \geq \pi_2 \text{ versus } \mathbf{H}_A\colon \pi_1 < \pi_2$$

- The data from this experiment may be summarized in a 2×2 contingency table:

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

- $\hat{\pi}_1 = 280/8872 = 0.0316$ and $\hat{\pi}_2 = 399/8642 = 0.0462$. Thus,

$$\hat{\pi} = \frac{8872(0.0316) + 8642(0.0462)}{17514}$$

$$t = \frac{0.0316 - 0.0462}{\sqrt{(0.0388)(1 - 0.0388)(1/8872 + 1/8642)}} = -5.002$$

- p-value $= \mathbb{P}(T \leq -5.002) = 2.84 \times 10^{-7} \approx 0$ 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:

$$\mathbf{H}_0\colon \theta_1 = \theta_2 \text{ versus } \mathbf{H}_A\colon \theta_1 \neq \theta_2$$

- 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 $\mathbf{H}_0$:

$$\mathcal{R} = \{t \mid \mathbf{H}_0 \text{ is rejected}\}$$

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

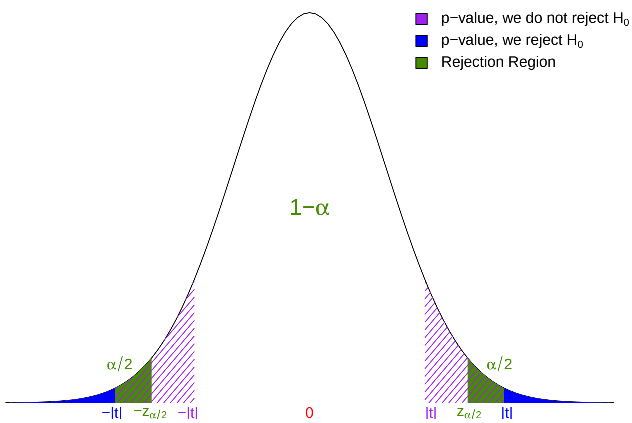

Figure 2.2: $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_A$: $\theta_1 \neq \theta_2$
$\mathcal{R} = \{t \mid t \leq -z_{\alpha/2} \text{ or } t \geq z_{\alpha/2}\}$

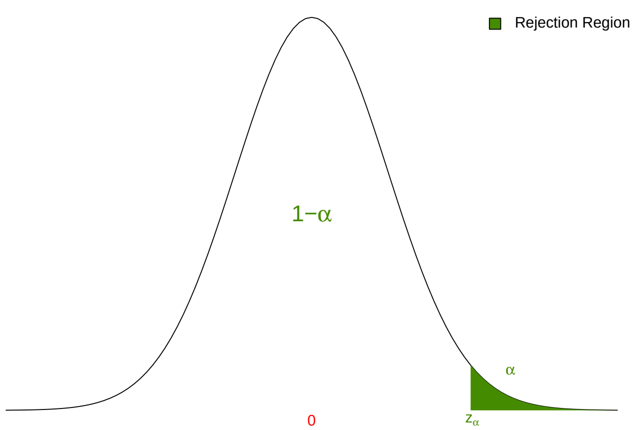

Figure 2.3: $\mathbf{H}_0$: $\theta_1 \leq \theta_2$ versus $\mathbf{H}_A$: $\theta_1 > \theta_2$
$\mathcal{R} = \{t \mid t \geq z_\alpha\}$

- Defining Type I and Type II error rates in terms of a rejection region is also useful:
  - $\alpha = \mathbb{P}(\text{Type I Error}) = \mathbb{P}(\text{Reject } \mathbf{H}_0 \mid \mathbf{H}_0 \text{ is true}) = \mathbb{P}(T \in \mathcal{R} \mid \mathbf{H}_0 \text{ is true})$.

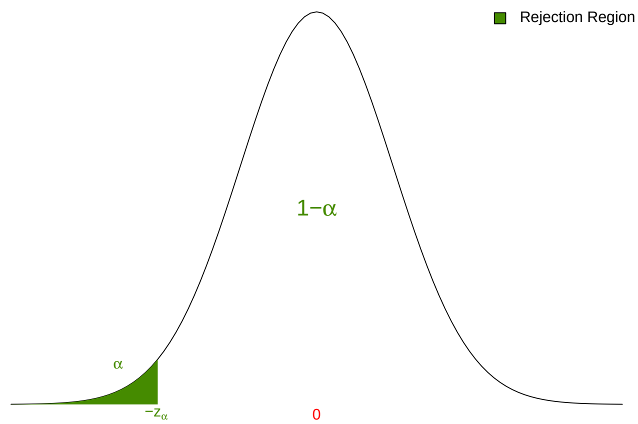

Figure 2.4: $\mathbf{H}_0$: $\theta_1 \geq \theta_2$ versus $\mathbf{H}_A$: $\theta_1 < \theta_2$
$\mathcal{R} = \{t \mid t \leq -z_\alpha\}$

– $\beta = \mathbb{P}(\text{Type II Error}) = \mathbb{P}(\text{Do Not Reject } \mathbf{H}_0 \mid \mathbf{H}_0 \text{ is false}) = \mathbb{P}(T \in \mathcal{R}^c \mid \mathbf{H}_0 \text{ is false})$.

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

Assuming $\mathbf{H}_0$ is true, $\theta_1 - \theta_2 = 0$ and $\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \sim \mathcal{N}(0,1)$. 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

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

- $\mathbb{V}(Y_1)$ 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 sampling.
  - 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 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.

    $$n_1 = n_2 = 50 \implies \binom{n_1 + n_2}{n_1} = \binom{100}{50} = 1.0089 \times 10^{29}$$

- 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.
- A **randomization test** is carried out similarly, except that 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."

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

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

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:

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

- $\mathbf{H}_0$: $\theta_1 = \theta_2$ versus $\mathbf{H}_\mathrm{A}$: $\theta_1 \neq \theta_2$. 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 $\mathbf{H}_\mathrm{A}$: $\theta_1 < \theta_2$. 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 $\mathbf{H}_\mathrm{A}$: $\theta_1 > \theta_2$. 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 $n_1 = n_2 = n_3 = 100$ users are randomized to each condition.
- For each user, the amount of real money (in USD) they spend in the 30 days following the experiment is recorded.
- The data summaries are:
  - $\bar{y}_1 = \$10.74$, $Q_{y_1}(0.5) = \$9$.
  - $\bar{y}_2 = \$9.53$, $Q_{y_2}(0.5) = \$8$.
  - $\bar{y}_3 = \$13.41$, $Q_{y_3}(0.5) = \$10$.

  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:

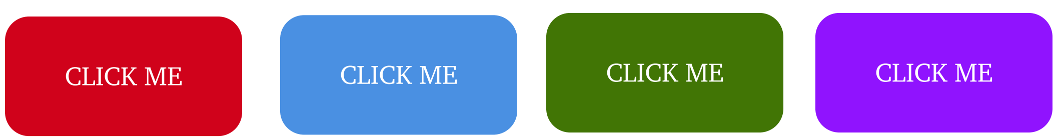

Figure 3.1: Canonical Button Colour Test.

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

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

- Determining which conditions are 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, such a question is phrased as the following statistical hypothesis:
$$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \dots = \theta_m \text{ versus } \mathbf{H}_\text{A}\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k$$
In the case of means:
$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \dots = \mu_m \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$
In the case of proportions:
$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \dots = \pi_m \text{ versus } \mathbf{H}_\text{A}\text{: } \pi_j \neq \pi_k \text{ for some } j \neq k$$

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 the measurements were taken, and that the variance is the same across all conditions.

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

- The "gatekeeper" test for means is tested using an F-test:
$$\mathbf{H}_0\text{: } \mu_1 = \mu_2 = \dots = \mu_m \text{ versus } \mathbf{H}_\text{A}\text{: } \mu_j \neq \mu_k \text{ for some } j \neq k$$

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:

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

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

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

    * Y_i = response observation for unit $i = 1, 2, \dots, N = \sum_{j=1}^{m} n_j$.
    * ε_i = random error term which we assume follows a $\mathcal{N}(0, \sigma^2)$ distribution.
    * 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 may be interpreted 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:

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

⋆ 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 $j = 1, 2, \dots, m-1$.

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

- 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:

  $\mathbf{H}_0$: $\beta_1 = \beta_2 = \cdots = \beta_{m-1}$ versus $\mathbf{H}_\text{A}$: $\beta_j \neq 0$ 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 defined to be 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 *__between__* conditions: $\text{SS}_\text{C} = \displaystyle\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}\right)^2$.
  - Quantifies variability *__within__* conditions: $\text{SS}_\text{E} = \displaystyle\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_{\bullet j}\right)^2$.
  - Quantifies *__overall__* variability: $\text{SS}_\text{T} = \displaystyle\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2 = \text{SS}_\text{C} + \text{SS}_\text{E}$.
- The null distribution for this test is $F(m-1, N-m)$.
- The p-value for this test is calculated by p-value $= \mathbb{P}(T \geq t)$ where $T \sim F(m-1, N-m)$.

3.1.2 Example: Candy Crush Boosters

Table 3.1: ANOVA Table

Source	SS	d.f.	MS	Test Stat.
Condition	SS_C	$m-1$	$\text{SS}_\text{C}/(m-1)$	$t = \text{MS}_\text{C}/\text{MS}_\text{E}$
Error	SS_E	$N-m$	$\text{SS}_\text{E}/(N-m)$	
Total	SS_T	$N-1$		

EXAMPLE 3.1.1: 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.
- Users are randomized 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 a particular value of the response was observed 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:

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

  was $t = 851.8947$ and the null distribution was $T \sim F(2, 370)$. The corresponding p-value was:

$$p\text{-value} = \mathbb{P}(T \geq 851.8947) = 3.28 \times 10^{-139}$$

- 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:

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

- $Y_{ij} \sim \text{Binomial}(1, \pi_j)$ where π_j is the probability of a unit in condition j performing the action.
- The "gatekeeper" test for proportions is tested using the **chi-squared test of independence** (also known as Pearson χ^2-test).

$$\mathbf{H}_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m$$

3.2.1 The Chi-squared Test of Independence

- The chi-squared test of independence is typically used 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.
  - 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 $\hat{\pi} = 679/17514 = 0.0388$.
  - Let X = number of conversions in a condition with n units, therefore $X \sim$ Binomial(n, π) where $\mathbb{E}[X] = n\pi$.
  - Therefore, we would expect $n_1\hat{\pi} = 8872(0.0388) = 343.96$ conversions in condition 1, and $n_2\hat{\pi} = 8642(0.0388) = 335.04$ 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 below.

		Condition		
		1	2	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	O_2
		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, 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
$$T = \sum_{\ell=0}^{1}\sum_{j=1}^{2} \frac{(O_{\ell_j} - E_{\ell,j})^2}{E_{\ell,j}} \sim \chi^2(1)$$
  - The p-value for this test is calculated by p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(1)$.
  - Returning to the Optimizely example, the *expected* table is:

		Condition		
		1	2	
Conversion	Yes	343.96	335.04	679
	No	8528.04	8306.96	16835
		8872	8642	17514

- And the resultant test statistic and p-value are:

$$t = \frac{(280 - 343.96)^2}{343.96} + \frac{(399 - 335.04)^2}{335.04} + \frac{(8592 - 8528.04)^2}{8528.04} + \frac{(8243 - 8306.96)^2}{8306.96} = 25.075$$

$$p\text{-value} = \mathbb{P}(T \geq 25.075) = 5.52 \times 10^{-7}$$

- 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:

		Condition				
		1	2	⋯	m	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	⋯	$O_{1,m}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	⋯	$O_{0,m}$	O_2
		n_1	n_2	⋯	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:

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

  - The p-value associated with this test is calculated as p-value $= \mathbb{P}(T \geq t)$ where $T \sim \chi^2(m-1)$.

3.2.2 Example: Nike SB Ads

EXAMPLE 3.2.1: Nike SB 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. $Y = 1$ if ad was viewed and 0 otherwise.
- Each of these 5 videos is shown to $n_1 = 5014$, $n_2 = 4971$, $n_3 = 5030$, $n_4 = 5007$, and $n_5 = 4980$ users, and the results are summarized in the table below.

		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:

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

- The expected cell frequencies are found by multiplying n_j by $\hat{\pi}$ and $(1 - \hat{\pi})$ for $j = 1, 2, 3, 4, 5$.

		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:
  $$\mathbf{H}_0\text{: } \theta_1 = \theta_2 = \cdots = \theta_m \text{ versus } \mathbf{H}_\mathrm{A}\text{: } \theta_j \neq \theta_k \text{ for some } j \neq k$$
  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.
  - $M = M_0 + M_A$.
- The outcomes of these M decisions are nicely summarized in Table 3.2.

		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

Table 3.2: Outcomes from M simultaneous hypothesis tests

  - R and $M - R$ are observable.
  - M_0, M_A, V, U, S, T are random variables; that is, their values are determined by the random process of collecting data and testing the M hypotheses. 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 defined as 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 each of the M tests are carried out with a significance level α, 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, $\lim_{M \to \infty}$ 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: $\mathbf{H}_{0,1}, \mathbf{H}_{0,2}, \dots, \mathbf{H}_{0,M}$.
- Denote their corresponding p-values as: $p_1, p_2, \dots, p_M$.

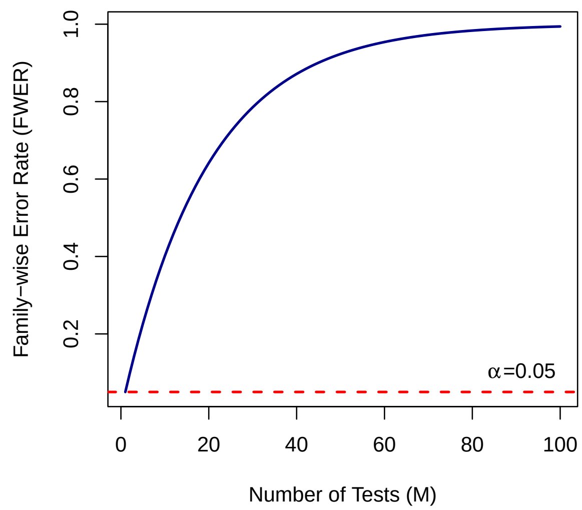

Figure 3.2: Family-wise error rate versus the number of hypothesis tests, M.

EXAMPLE 3.3.5

Suppose $M = 4$ hypotheses are tested 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

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

  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

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

- When independence is assumed 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.

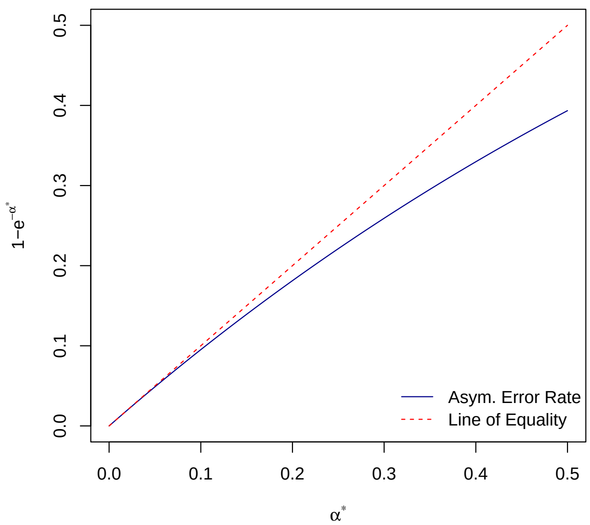

Figure 3.3: Illustration of the Bonferroni correction for asymptotically large M.

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

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

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,

$$\frac{\alpha^\star}{M} = 0.005 \quad \text{and} \quad 1 - (1-\alpha^\star)^{1/M} = 0.005116$$

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:

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

   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 $\mathbf{H}_{0,(1)}, \dots, \mathbf{H}_{0,(k^\star-1)}$ and do not reject $\mathbf{H}_{0,(k^\star)}, \dots, \mathbf{H}_{0,(M)}$.

- 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.

Proof of Theorem 3.3.10 †

- We need to show that $\text{FWER} = \mathbb{P}(V \geq 1) \leq \alpha^\star \in [0,1]$ 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

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

Also, 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,

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(V \geq 1) \\
&= \mathbb{P}(\text{At least one Type I Error in } M \text{ tests}) \\
&= \mathbb{P}(\text{Reject at least one true } \mathbf{H}_0) \\
&= \mathbb{P}\left(\exists\, k \in K_0 \text{ such that } p_k \leq \frac{\alpha^\star}{M_0}\right) \\
&= \mathbb{P}\left(\bigcup_{k \in K_0} p_k \leq \frac{\alpha^\star}{M_0}\right) \\
&\leq \sum_{k \in K_0} \mathbb{P}\left(p_k \leq \frac{\alpha^\star}{M_0}\right) && \text{by Boole's Inequality} \\
&= \sum_{k \in K_0} \frac{\alpha^\star}{M_0} && \text{since } p\text{-values for true null hypotheses follow a Uniform}(0,1) \text{ distribution} \\
&= M_0 \left(\frac{\alpha^\star}{M_0}\right) && \text{since the set } K_0 \text{ has cardinality } M_0 \\
&= \alpha^\star
\end{aligned}$$

Therefore, $\text{FWER} \leq \alpha^\star$ as required.

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 $\text{FWER} \leq \alpha^\star =$

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 $k^\star = 3$.

- 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.

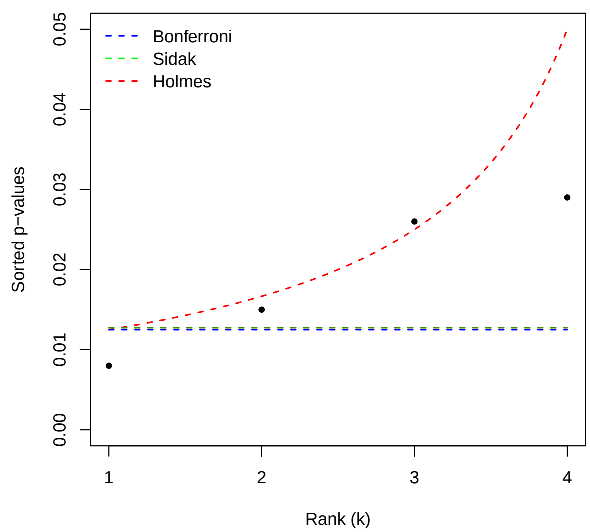

Figure 3.4: Significance thresholds for several methods of correction.

- 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, $p_1^\star = 0.0587$, $p_2^\star = 0.111$, $p_3^\star = 0.0316$, 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

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

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

- FWER methods were developed in the mid-1900's and 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 defined as

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

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 defined as

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

- 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. The main difference is the threshold we compare the ordered p-values to.
- The BH procedure, which aims to ensure FDR $\leq \alpha^\star$, may be summarized as follows:

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

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

   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 $\mathbf{H}_{0,(1)}, \dots, \mathbf{H}_{0,(k^\star)}$ and do not reject $\mathbf{H}_{0,(k^\star+1)}, \dots, \mathbf{H}_{0,(M)}$.

- The decision process associated with this procedure is best visualized with a plot of the ordered p-values $p_{(k)}$ versus their ranks $k = 1, 2, \dots, M$ 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 $\alpha^\star/M$.

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 FWER $\leq \alpha^\star = 0.05$. 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).

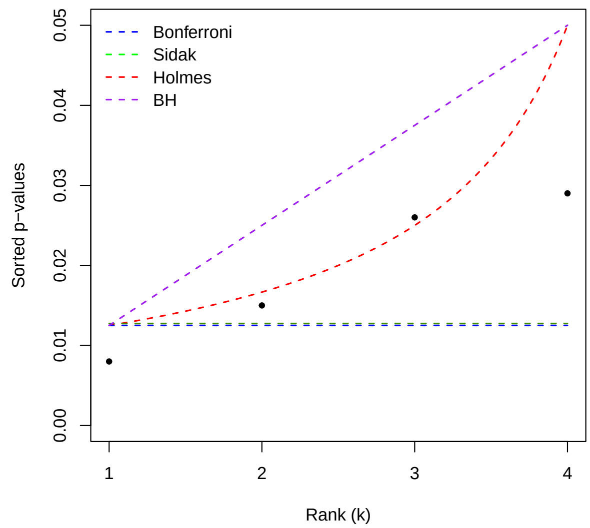

Figure 3.5: Significance thresholds for several methods of correction.

- 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

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

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	0.0032	$\min\{0.0032, 0.03, 0.035, 0.029\} = 0.029 = p^\star_{(1)}$
2	0.015	0.03	$\min\{0.03, 0.035, 0.029\} = 0.029 = p^\star_{(2)}$
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.

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.
    - ⋆ Assume $\mathcal{R} = \{t \mid t \geq z_\alpha\}$.

  - ⋆ Recall that:
    - · $\alpha = \mathbb{P}(\text{Type I Error}) = \mathbb{P}(T \in \mathcal{R} \mid \mathbf{H}_0 \text{ is true}) = \mathbb{P}(T \geq z_\alpha \mid \mathbf{H}_0 \text{ is true})$.
    - · $\beta = \mathbb{P}(\text{Type II Error}) = \mathbb{P}(T \notin \mathcal{R} \mid \mathbf{H}_\mathrm{A} \text{ is true}) = \mathbb{P}(T < z_\alpha \mid \mathbf{H}_\mathrm{A} \text{ is true})$.

- Since all of our correction methods decrease the effective significance levels, the power of such tests is negatively impacted.
- 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 whichever correction method you use.
  - This is easier to do with *some* correction methods than others.