Experimental Design

STAT 430

Spring 2021 (1215)

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

May 18, 2021

Contents

Chapter 1

Week 1 — Introduction

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

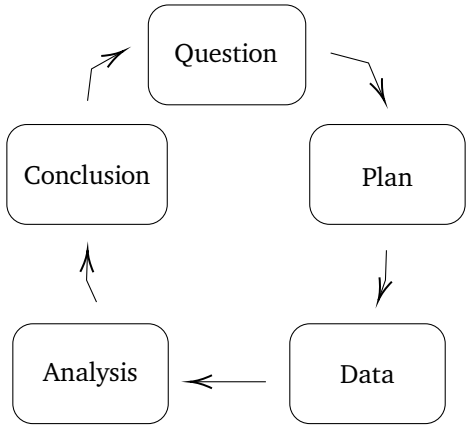

Figure 1.1: QPDAC Cycle

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

Week 2 — Experiments with Two Conditions

2.1 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: Button-Colour Experiment

    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.
  - H_0: $\theta_1 = \theta_2$ versus H_A: $\theta_1 \neq \theta_2$ (two-sided).
  - H_0: $\theta_1 \leq \theta_2$ versus H_A: $\theta_1 > \theta_2$ (one-sided).
  - H_0: $\theta_1 \geq \theta_2$ versus H_A: $\theta_1 < \theta_2$ (one-sided).
- "Absense 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* H_0 or *not reject* H_0.
- In order to draw such a conclusion, we will define a **test statistic** T, which 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 2.1.1: *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 H_0, is determined by the **significance level** of the test, denoted α.
  - If p-value $\leq \alpha$, we reject H_0.
  - If p-value $> \alpha$, we do not reject H_0.

REMARK 2.1.2

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 H_0 or H_A is true. Thus means that there are four possible outcomes when using data to decide which statement is true:
  (1) No Error: H_0 is true, and we correctly do not reject it.
  (2) Type I Error: H_0 is true, and we incorrectly reject it.
  (3) Type II Error: H_0 is false, and we incorrectly do not reject it.
  (4) No Error: 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 2.1.3: Pregnancy Test

H_0: person is not pregnant versus H_A: person is pregnant.
- Type I Error: person is pregnant (false positive).
- Type II Error: person is not pregnant (false negative).

EXAMPLE 2.1.4: Courtroom

H_0: the defendant is innocent versus H_A: the defendant is guilty.
- Type I Error: sentencing an innocent person to jail.
- Type II Error: letting a guilty person go free.

- $\alpha = \mathbb{P}(\text{Type I Error})$.
- $\beta = \mathbb{P}(\text{Type II Error})$ where $1 - \beta =$ Power.
- 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.

2.2 Comparing Two Means

- 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

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

  - $Y_{ij} =$ response observation for unit i in condition j.
- Using the observed data, we test hypotheses of the form:
  - H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$.
  - H_0: $\mu_1 \leq \mu_2$ versus H_A: $\mu_1 > \mu_2$.
  - H_0: $\mu_1 \geq \mu_2$ versus H_A: $\mu_1 < \mu_2$.

t-tests, F-tests, and an Example

STATISTICAL TEST 2.2.1: 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 H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$, we compute p-value $= \mathbb{P}(T \geq |t|) + \mathbb{P}(T \leq -|t|)$.
  - For H_0: $\mu_1 \leq \mu_2$ versus H_A: $\mu_1 > \mu_2$, we compute p-value $= \mathbb{P}(T \geq t)$.
  - For H_0: $\mu_1 \geq \mu_2$ versus H_A: $\mu_1 < \mu_2$, we compute p-value $= \mathbb{P}(T \leq t)$.

REMARK 2.2.2

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

STATISTICAL TEST 2.2.3: 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 2.2.1, but where the null distribution is $T \sim t(\nu)$.

STATISTICAL TEST 2.2.4: F-test for Variances

- *Purpose*:
  - H_0: $\sigma_1^2 = \sigma_2^2$ versus H_A: $\sigma_1^2 \neq \sigma_2^2$.
  - H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus 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 2.2.5

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

EXAMPLE 2.2.6: 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:

$$H_0\text{: } \mu_1 \leq \mu_2 \text{ versus } H_A\text{: } \mu_1 > \mu_2$$

- 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 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 H_0: $\mu_1 \leq \mu_2$, and conclude that increasing ad frequency significantly reduces average session duration.

[R Code] `Comparing_two_means`

2.3 Comparing Two Proportions

- 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, $Y_{ij} \sim \text{Bernoulli}(1, \pi_j)$ 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:
  - H_0: $\pi_1 = \pi_2$ versus H_A: $\pi_1 \neq \pi_2$.
  - H_0: $\pi_1 \leq \pi_2$ versus H_A: $\pi_1 > \pi_2$.
  - H_0: $\pi_1 \geq \pi_2$ versus H_A: $\pi_1 < \pi_2$.

Z-tests and an Example

STATISTICAL TEST 2.3.1: 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 $T \sim \mathcal{N}(0,1)$.

EXAMPLE 2.3.2: 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 $\pi =$ conversion rate (where $y = 1$ if a homepage visitor signed up and 0 otherwise).
- The hypothesis tested here is:

$$H_0\text{: } \pi_1 \geq \pi_2 \text{ versus } H_A\text{: } \pi_1 < \pi_2$$

- 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 H_0 and conclude that the redesigned homepage significantly increases conversion rate.
- [R Code] `Comparing_two_proportions`

2.4 A Power Calculation

- Used to control Type II Error.
- Power analyses help determine required sample sizes.
- Suppose, for illustration, that we are interested in testing the hypothesis:
$$H_0\text{: } \theta_1 = \theta_2 \text{ versus } H_A\text{: } \theta_1 \neq \theta_2$$
- Suppose, also for illustration, that the test statistic associated with this test has the form:
$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \overbrace{(\theta_1 - \theta_2)}^{0}}{\sqrt{\frac{\mathbb{V}(Y_1)}{n} + \frac{\mathbb{V}(Y_2)}{n}}} \sim \mathcal{N}(0, 1)$$
- It will be useful to define the notion of a **rejection region**, denoted $\mathcal{R}$, which is all the values of the observed test statistic t that would lead to the rejection of H_0:
$$\mathcal{R} = \{t \mid H_0 \text{ is rejected}\}$$
  - If $t \in \mathcal{R}$, we reject H_0.
  - If $t \in \mathcal{R}^c$, we do not reject H_0.

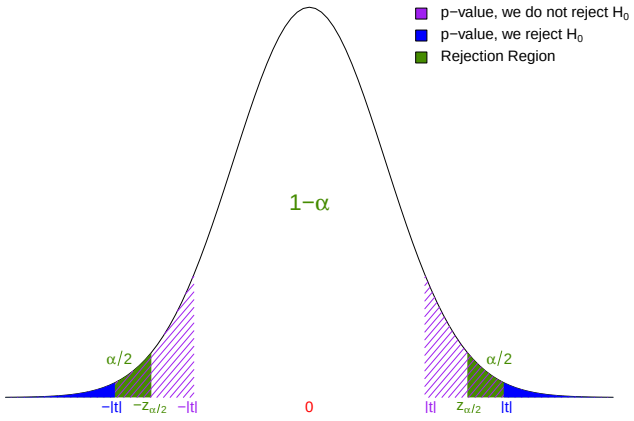

Figure 2.2: H_0: $\theta_1 = \theta_2$ versus H_A: $\theta_1 \neq \theta_2$
$\mathcal{R} = \{t \mid t \leq -z_{\alpha/2} \text{ or } t \geq z_{\alpha/2}\}$

- Defining Type I and Type II error rates in terms of a rejection region is also useful:
  - $\alpha = \mathbb{P}(\text{Type I Error}) = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ is true}) = \mathbb{P}(T \in \mathcal{R} \mid H_0 \text{ is true})$.
  - $\beta = \mathbb{P}(\text{Type II Error}) = \mathbb{P}(\text{Do Not Reject } H_0 \mid H_0 \text{ is false}) = \mathbb{P}(T \in \mathcal{R}^c \mid H_0 \text{ is false})$.

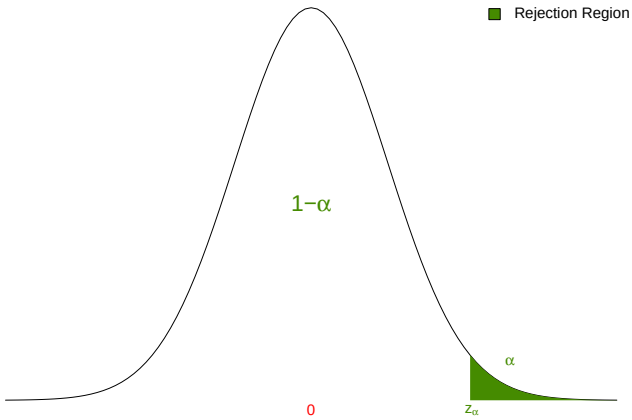

Figure 2.3: H_0: $\theta_1 \leq \theta_2$ versus H_A: $\theta_1 > \theta_2$
$\mathcal{R} = \{t \mid t \geq z_\alpha\}$

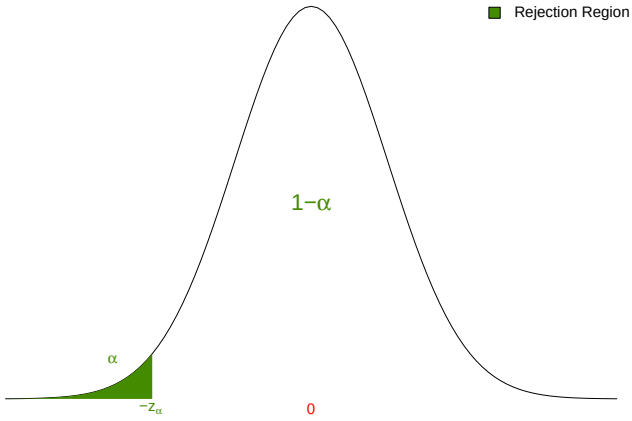

Figure 2.4: H_0: $\theta_1 \geq \theta_2$ versus H_A: $\theta_1 < \theta_2$
$\mathcal{R} = \{t \mid t \leq -z_\alpha\}$

$$
\begin{aligned}
1-\beta &= \text{Power} \\
&= 1-\mathbb{P}(\text{Type II Error}) \\
&= 1-\mathbb{P}(T \in \mathcal{R}^c \mid H_0 \text{ is false}) \\
&= \mathbb{P}(T \in \mathcal{R} \mid H_0 \text{ is false}) \\
&= \mathbb{P}\Big(T \geq z_{\alpha/2} \cup T \leq -z_{\alpha/2} \,\Big|\, H_0 \text{ is false}\Big) \\
&= \mathbb{P}\Big(T \geq z_{\alpha/2} \,\Big|\, H_0 \text{ is false}\Big) + \mathbb{P}\Big(T \leq -z_{\alpha/2} \,\Big|\, H_0 \text{ is false}\Big) \\
&= \mathbb{P}\left(\frac{\bar{Y}_1-\bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \geq z_{\alpha/2} \,\middle|\, H_0 \text{ is false}\right) + \mathbb{P}\left(\frac{\bar{Y}_1-\bar{Y}_2}{\sqrt{\frac{\mathbb{V}(Y_1)+\mathbb{V}(Y_2)}{n}}} \leq -z_{\alpha/2} \,\middle|\, H_0 \text{ is false}\right)
\end{aligned}
$$

Assuming 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, 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), and it is the smallest difference between θ_1 and θ_2 that has practical importance and that we would like to detect as being statistically significant.