A/B Testing

- A/B Testing: A technique used by eCommerce websites, SaaS web apps, and media websites to compare two versions of a web page to see which one performs better.
- For simplicity, let's just try to decide whether the two web designs are the same. Suppose we have collected the following data
  - **version A:** n_1 visitors and m_1 clicks
  - **version B:** n_2 visitors and m_2 clicks
- Of course, if the difference of their CTR (click-through-rate) is big, we could conclude that they are different. But how big is big? That is, how to choose the cut-off value C

$$\left|\frac{m_1}{n_1} - \frac{m_2}{n_2}\right| \geq C?$$

The Probabilistic Model

- Model the data as samples from two Bernoulli models

$$\begin{aligned} X_1, \ldots, X_{n_1} \quad iid \quad &\sim \quad \text{Bern}(p_1) \\ Y_1, \ldots, Y_{n_2} \quad iid \quad &\sim \quad \text{Bern}(p_2) \end{aligned}$$

$$\begin{aligned} \hat{p}_1 \quad &= \quad \bar{X}, \quad \mathbb{E}(\bar{X}) = p_1, \quad \text{Var}(\bar{X}) = p_1(1-p_1)/n_1 \\ \hat{p}_2 \quad &= \quad \bar{Y}, \quad \mathbb{E}(\bar{Y}) = p_2, \quad \text{Var}(\bar{Y}) = p_2(1-p_2)/n_2 \end{aligned}$$

- We aim to test

$$H_0 : p_1 = p_2, \quad H_a : p_1 \neq p_2.$$

- Conclude that H_a holds, if $|\hat{p}_1 - \hat{p}_2| > C$, which is often referred to as the **Rejection Region**.

Trade-off of the Two Types of Errors

- Type I error

$$\mathbb{P}(|\hat{p}_1 - \hat{p}_2| > C \mid p_1 = p_2 = p_0)$$

To reduce type I error, we need to make C large; in the extreme case, set C to be 1, then we will not make any type I error, but ....

- Type II error

$$1 - \mathbb{P}(|\hat{p}_1 - \hat{p}_2| > C \mid p_1 \neq p_2)$$

To reduce type II error, we need to make C small; in the extreme case, set C to be 0, then we will not make any type II error, but ....

- Strategy: control only the type I error. The type I error rate α is also called the significance level of the test.

  Later we will study how to derive the best test that has the smallest type II error among all level-α tests; this is related to the concept of UMP (uniformly most powerful) test.

Significant Level and Rejection Region

How to choose the rejection region such that the type I error is $\leq \alpha$, i.e., the significant level of the test is α. Focus on the null hypothesis: assume data are generated from the null model, i.e., $p_1 = p_2 = p_0$.

- $X_1, \ldots, X_{n_1}$ iid $\sim$ Bern(p_0) and $Y_1, \ldots, Y_{n_2}$ iid $\sim$ Bern(p_0).

  By CLT,

  $$\sqrt{n_1}(\bar{X} - p_0) \xrightarrow{D} \mathsf{N}(0, p_0(1-p_0)), \quad \text{i.e. } \bar{X} = \frac{m_1}{n_1} \approx \mathsf{N}\Big(p_0, \frac{p_0(1-p_0)}{n_1}\Big)$$

  $$\sqrt{n_2}(\bar{Y} - p_0) \xrightarrow{D} \mathsf{N}(0, p_0(1-p_0)), \quad \text{i.e. } \bar{Y} = \frac{m_2}{n_2} \approx \mathsf{N}\Big(p_0, \frac{p_0(1-p_0)}{n_2}\Big)$$

  and

  $$\bar{X} - \bar{Y} = \frac{m_1}{n_1} - \frac{m_2}{n_2} \approx \mathsf{N}\Big(0, p_0(1-p_0)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big).$$

- Construct a pooled estimate of p_0:

$$\hat{p}_0 = \frac{\sum_i X_i + \sum_j Y_j}{n_1 + n_2} = \frac{m_1 + m_2}{n_1 + n_2} \xrightarrow{P} p_0, \text{ under the null.}$$

- By Slutsky's Theorem, we can replace p_0 by its pooled estimate and eventually we have

$$T = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx \mathsf{N}(0,1).$$

- So we can set the Rejection Region to be

$$|T| \geq z_{\alpha/2}.$$

How to Construct a Level-α Test?

- Approximate the data by a statistical model, and state the corresponding H_0 and H_a.
- Pick a test statistic T (that you can compute from the data). Know the range of T that is against H_0 (e.g., right-tail, left-tail, or both tails), which will be the Rejection Region.
- Find the distribution or asymptotic distribution of T under H_0, then find the rejection region such that

$$\mathbb{P}(T \text{ in Rejection Region} \mid H_0) = \alpha.$$

After observing some data, we compute the observed statistic $T(\text{data}) = t$. Then how to make decision?

- If t is in the Rejection Region, reject H_0, or equivalently
- compute p-value, and if p-value is less than α, reject H_0.

Note: p-value is NOT equal to the probability that H_0 is true given the data; it is about a probability computed under the assumption H_0 is true.

Go through the Piazza notes and practice the following skills.

- Given the significant level α and test statistic T, how to determine the rejection region.
- Given the rejection region, compute the type I error (significant level).
- Given the significant level α, rejection region, and the true value of θ (which is from H_a), compute the power or type II error.
- Given the significant level α, rejection region, and the observed statistic, make your decision (rejection H_0 or not) or report p-value.
- Know the difference between one-sided test and two-sided test.
- Determine the minimal sample size to reach certain power of a level-α test.