2 Experiments with Two Conditions

2.1 Anatomy of an A/B Test

→ sometimes called "experimentation"

- 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".

  - Canonical A/B test:

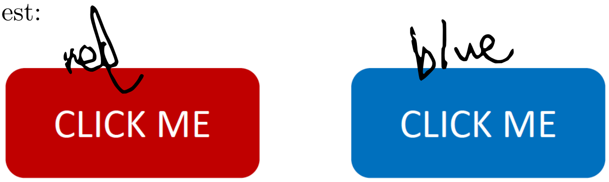


Figure 1: Button-Colour Experiment

- 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
  - proportion
  - variance
  - quantile
  - technically any statistic that can be calculated 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 this 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.

$$H_0\colon\ \theta_1 = \theta_2 \text{ vs. } H_A\colon\ \theta_1 \neq \theta_2. \qquad (1)$$

or

$$H_0\colon\ \theta_1 \leq \theta_2 \text{ vs. } H_A\colon\ \theta_1 > \theta_2 \qquad (2)$$

or

$$H_0\colon\ \theta_1 \geq \theta_2 \text{ vs. } H_A\colon\ \theta_1 < \theta_2 \qquad (3)$$

- 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 define a **test statistic** T which is a random variable that satisfies three properties:

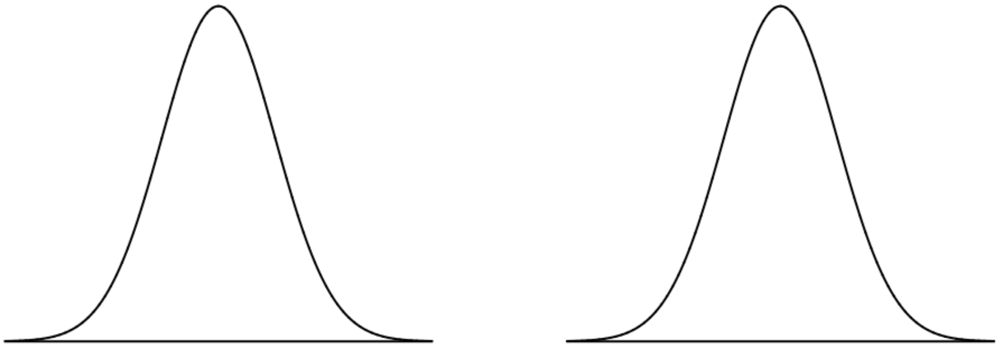
  (i) it must be a function of the observed data

  (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**.

- 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.
  - 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

- 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

- In order to choose α one needs to understand the two types of error that can be made when drawing conclusions in the context of a hypothesis test.

- Recall that by design either H_0 is true or H_A is true. This means that there are four possible outcomes when using data to decide which statement is true:

  (1) H_0 is true and we correctly do not reject it
  (2) H_0 is true and we incorrectly reject it (type I
  (3) H_0 is false and we incorrectly do not reject it (type II)
  (4) 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 to each type of error
  - So we may wish to treat them differently

- **Pregnancy Test Example**

$$H_0\text{: person is not pregnant vs. } H_A\text{: person is pregnant}$$

- **Courtoom Example**

$$H_0\text{: the defendant is innocent vs. } H_A\text{: the defendant is guilty}$$

- Fortunately it is possible to control the frequency with 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

$$Y_{i1} \sim \mathrm{N}(\mu_1, \sigma^2) \text{ and } Y_{i2} \sim \mathrm{N}(\mu_2, \sigma^2)$$

  where $i = 1, 2, \ldots, n_j$ for $j = 1, 2$.

- Using the observed data we test hypotheses of the form

$$H_0\colon \mu_1 = \mu_2 \text{ vs. } H_A\colon \mu_1 \neq \mu_2 \tag{4}$$

$$H_0\colon \mu_1 \leq \mu_2 \text{ vs. } H_A\colon \mu_1 > \mu_2 \tag{5}$$

$$H_0\colon \mu_1 \geq \mu_2 \text{ vs. } H_A\colon \mu_1 < \mu_2 \tag{6}$$

2.2.1 *t*-tests, *F*-tests, and an Example

Student's *t*-test: $(\sigma_1 = \sigma_2)$

- Purpose: compare μ_1 and μ_2
- Test Statistic:

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - \overset{\text{plugged in } 0}{\boxed{(\mu_1 - \mu_2)}}}{\boxed{\hat{\sigma}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \tag{7}$$

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

- Observed Version:

  where $\overline{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \sim N\left(\mu_j, \frac{\sigma_j^2}{n}\right)$

☆ and $\hat{\sigma}^2 = \frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}$ (sum all residual squares divided by total d.o.f.)

  and $\hat{\sigma}_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 \ (= s_j^2)$

For (4) → p-val = $P(T \geq |t|) + P(T \leq -|t|)$
$= 2P(T \geq |t|)$

$T \sim t_{(n_1+n_2-2)}$
for all cases
in Student's T-test

Hyp. (5) → p-val = $P(T \geq |t|)$

- p-value Calculation:

Hyp. (6) → p-val = $P(T \leq -|t|)$

Welch's t-test: $(\sigma_1 \neq \sigma_2)$

"wall-toh"

- Purpose: compare μ_1 and μ_2
- Test Statistic:

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \dot{\sim} t_\nu \tag{8}$$

→ approximately follows

where

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

→ but IRL, p-vals are similar if using $\min(n_1, n_2) - 1$ as the d.o.f. (approaches z-test anyway)

- Observed Version: $t = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$
- p-value Calculation:

same w Student's t-test?

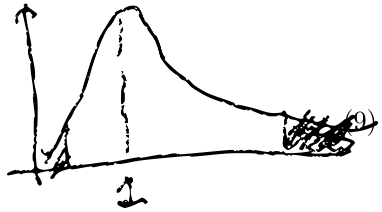
***F*-test for Variances:**

- Purpose: Test whether two (popu.) variances are ~~equal~~ leq / geq

$H_0: \sigma_1^2 = \sigma_2^2$ $\left(\sigma_1^2/\sigma_2^2 = 1\right)$; $H_A: \sigma_1^2 \neq \sigma_2^2$ $\left(\sigma_1^2/\sigma_2^2 \neq 1\right)$

- Test Statistic:

$$T = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2} \sim F_{(n_1-1, n_2-1)} \tag{9}$$



Fdist cdf

- Observed Version: $t = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$
- p-value Calculation:

If $t > 1$, p-val = $P(T \geq t) + P(T \leq \frac{1}{t})$, where $T \sim F_{(n_1-1, n_2-1)}$

If $t < 1$, p-val = $P(T \geq \frac{1}{t}) + P(T \leq t)$,

Example: 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 $\mu =$ average session time (where $y =$ the length of time a user engages with the app – in minutes).
- The hypothesis being tested here is:

$$H_0\colon \mu_1 \leq \mu_2 \text{ vs. } H_A\colon \mu_1 > \mu_2$$

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

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

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

  for $i = 1, 2, \ldots, n_j$ for $j = 1, 2$.

- Because the Y_{ij}'s are binary, it is common to assume that they follow a Bernoulli distribution:

$$Y_{ij} \sim \text{BIN}(1, \pi_j)$$

  where π_j represents the probability that $Y_{ij} = 1$

- Using the observed data we test hypotheses of the form

$$H_0\colon \pi_1 = \pi_2 \text{ vs. } H_A\colon \pi_1 \neq \pi_2 \tag{10}$$

$$H_0\colon \pi_1 \leq \pi_2 \text{ vs. } H_A\colon \pi_1 > \pi_2 \tag{11}$$

$$H_0\colon \pi_1 \geq \pi_2 \text{ vs. } H_A\colon \pi_1 < \pi_2 \tag{12}$$

2.3.1 *Z*-tests and an Example

***Z*-test for Proportions:**

- Purpose: test whether $\pi_1 = \pi_2$

due to CLT

- Test Statistic://

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\cdot}{\sim} N(0,1) \tag{13}$$

  where $\hat{\pi} = \dfrac{n_1 \hat{\pi}_1 + n_2 \hat{\pi}_2}{n_1 + n_2}$ where $\hat{\pi}_j = \bar{y}_j$

"common proportion estimate"

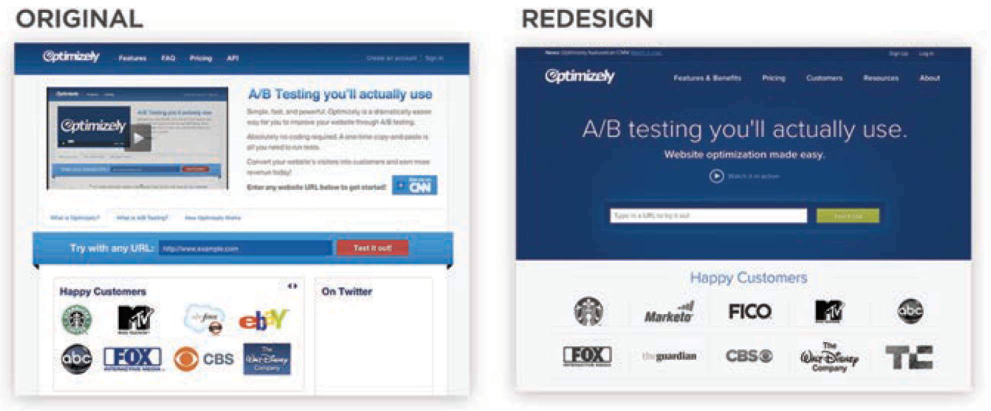
- Observed Version: $t = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

- p-value Calculation:

  Same

Example: 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 or not 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 being tested here is:

$$H_0\colon\ \pi_1 \geq \pi_2 \text{ vs. } H_A\colon\ \pi_1 < \pi_2$$

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

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

Optional Exercises:

- Calculations: 1, 2
- R Analysis: 1, 3, 4, 6 (not e), 13 (not g), 22 (not h)