

# STAT1301 Assignment 4

Caleb Yates s49886351

October 14, 2025

# 1 Question 4

Let  $X$  be the random variable for the number of people received the direct mail strategy and completed screening. Let  $Y$  be the random variable for the number of people who received the education only outreach and completed screening.

## 1.1 Part a)

The notation  $p_X$  and  $p_Y$  represent the population proportion for  $X$  and  $Y$  respectively.

The null hypothesis is that both population proportions are equal:  $H_0 : p_X = p_Y = p$ . The alternative hypothesis is therefore:  $H_1 : p_X > p_Y$ .

## 1.2 Part b)

$X \sim \text{Bin}(n_X, p_X)$  where  $n_X = 1415$  so  $X \sim \text{Bin}(1415, p_X)$ . It is (implicitly) assumed that samples  $X_i$  from  $X$  follow the distribution of  $X$  and are all independent, hence:

$$X_i \sim \text{Bin}(n_X, p_X)$$

Since  $n_X p_X = 505 \gg 5$  and  $n_X(1 - p_X) = 910 \gg 5$ , the conditions for the Central Limit Theorem (CLT) to be a good approximation are met, as well as a suitably large  $n_X$ . Hence the CLT is reasonable for the research problem. Therefore:

$$X_i \overset{\text{approx}}{\sim} N(n_X p_X, n_X p_X (1 - p_X))$$

$$\bar{X} = \hat{P}_X \overset{\text{approx}}{\sim} N(p_X, \frac{p_X(1 - p_X)}{n_X})$$

Under  $H_0$ :

$$\hat{P}_X \sim N(p, \frac{p(1 - p)}{n_X})$$

$Y \sim \text{Bin}(n_Y, p_Y)$  where  $n_Y = 1408$  so  $Y \sim \text{Bin}(1408, p_Y)$ . It is (implicitly) assumed that samples  $Y_i$  from  $Y$  follow the distribution of  $Y$  and are all independent, hence:

$$Y_i \sim \text{Bin}(n_Y, p_Y)$$

Since  $n_Y p_Y = 264 \gg 5$  and  $n_Y(1 - p_Y) = 1144 \gg 5$ , the conditions for the Central Limit Theorem (CLT) to be a good approximation are met, as well as a suitably large  $n_Y$ . Hence the CLT is reasonable for the research problem. Therefore:

$$Y_i \overset{\text{approx}}{\sim} N(n_Y p_Y, n_Y p_Y (1 - p_Y))$$

$$\bar{Y} = \hat{P}_Y \overset{\text{approx}}{\sim} N(p_Y, \frac{p_Y(1 - p_Y)}{n_Y})$$

Under  $H_0$ :

$$\hat{P}_Y \sim N(p, \frac{p(1 - p)}{n_Y})$$

It is additionally assumed that  $X$  and  $Y$  are independent from each other.

We can now give notation for the specific sample information we are given:  $\bar{x} = \hat{p}_X = \frac{505}{1415} \approx 0.3568$  and  $\bar{y} = \hat{p}_Y = \frac{264}{1408} \approx 0.1875$

### 1.3 Part c)

$$\hat{P}_X - \hat{P}_Y \sim N(p_X - p_Y, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y})$$

Under  $H_0$ , or assuming  $H_0$ :

$$\hat{P}_X - \hat{P}_Y \sim N(0, \frac{p(1-p)}{n_X} + \frac{p(1-p)}{n_Y})$$

To find a pivotal variable that doesn't depend on the unknown  $p$ , a pooled unbiased estimator  $\hat{P} = \frac{X+Y}{n_X+n_Y}$  will be used in place of  $p = \hat{P}$ . For our sample  $\hat{P} = \frac{505+264}{1415+1408} \approx 2.724$ . Rearranging gives:

$$T = \frac{\hat{P}_X - \hat{P}_Y}{\hat{P}(1-\hat{P})(\frac{1}{n_X} + \frac{1}{n_Y})} \sim N(0, 1)$$

The p-value is therefore computable relative to our specific sample:

$$\text{p-value} = P(\{\hat{P}_X - \hat{p}_X \geq \hat{P}_Y - \hat{p}_Y\})$$

In a slightly more useful arrangement:

$$\text{p-value} = P(\{\hat{P}_X - \hat{P}_Y \geq \hat{p}_X - \hat{p}_Y\})$$

$$\text{p-value} = P(\{\frac{\hat{P}_X - \hat{P}_Y}{\hat{P}(1-\hat{P})(\frac{1}{n_X} + \frac{1}{n_Y})} \geq \frac{\hat{p}_X - \hat{p}_Y}{\hat{P}(1-\hat{P})(\frac{1}{n_X} + \frac{1}{n_Y})}\})$$

Note  $\hat{p}_X - \hat{p}_Y \approx 0.16939$  and  $\hat{P}(1-\hat{P})(\frac{1}{n_X} + \frac{1}{n_Y}) \approx 0.01676$ , hence:

$$\text{p-value} = P(\{Z \geq \frac{0.16939}{0.01676}\})$$

$$\text{p-value} = P(\{Z \geq 10.1079\})$$

$$\text{p-value} = 1 - P(\{Z < 10.1079\})$$

Looking at the stats table, this is way off the charts!  $\Phi(3.69) = 0.9999$ , so

$$\text{p-value} < 1 - 0.9999$$

$$\text{p-value} < 0.0001$$

This is very strong evidence to reject the null hypothesis  $H_0$ . Therefore we can conclude we have very strong evidence that direct-mail self-sampling kits have a higher screening population proportion than education only outreach.

## 1.4 Part d)

A 97% confidence interval means  $\alpha = 3\% = 0.03$ . The formula for a 2-sample binomial confidence interval is as follows:

$$\hat{P}_X - \hat{P}_Y \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_X(1 - \hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1 - \hat{P}_Y)}{n_Y}}$$

$z_{\frac{\alpha}{2}}$  is the the solution to:

$$P(\{Z > z_{\frac{\alpha}{2}}\}) = \frac{\alpha}{2} = 0.015$$

$$P(\{Z < z_{\frac{\alpha}{2}}\}) = 0.985$$

From the stats table,  $z_{\frac{\alpha}{2}} = 2.17$ . Plugging in our specific sample notation:

$$\hat{p}_X - \hat{p}_Y \pm 2.17 \sqrt{\frac{\hat{p}_X(1 - \hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1 - \hat{p}_Y)}{n_Y}}$$

Evaluating yields the CI (0.1337, 0.2051) for  $p_X - p_Y$ . This confidence interval does not contain 0, therefore this is evidence that the proportions  $p_X$  and  $p_Y$  are not equal.