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 compelted 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: $\text{H}_0 : p_\text{X} = p_\text{Y} = p$. The alternative hypothesis is therefore: $\text{H}_1 : p_\text{X} > p_\text{Y}$.

1.2 Part b)

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

$$\text{X}_i \sim \text{Bin}(n_\text{X}, p_\text{X})$$

Since $n_\text{X} p_\text{X} = 505 \gg 5$ and $n_\text{X}(1 - p_\text{X}) = 910 \gg 5$, 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:

$$\text{X}_i \overset{\text{approx}}{\sim} N(n_\text{X} p_\text{X}, n_\text{X} p_\text{X}(1 - p_\text{X}))$$

$$\bar{\text{X}} = \hat{P}_\text{X} \overset{\text{approx}}{\sim} N(p_\text{X}, \frac{p_\text{X}(1 - p_\text{X})}{n_\text{X}})$$

Under H_0:

$$\hat{P}_\text{X} \sim N(p, \frac{p(1-p)}{n_\text{X}})$$

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

$$\text{Y}_i \sim \text{Bin}(n_\text{Y}, p_\text{Y})$$

Since $n_\text{Y} p_\text{Y} = 264 \gg 5$ and $n_\text{Y}(1 - p_\text{Y}) = 1144 \gg 5$, 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:

$$\text{Y}_i \overset{\text{approx}}{\sim} N(n_\text{Y} p_\text{Y}, n_\text{Y} p_\text{Y}(1 - p_\text{Y}))$$

$$\bar{\text{Y}} = \hat{P}_\text{Y} \overset{\text{approx}}{\sim} N(p_\text{Y}, \frac{p_\text{Y}(1 - p_\text{Y})}{n_\text{Y}})$$

Under H_0:

$$\hat{P}_\text{Y} \sim N(p, \frac{p(1-p)}{n_\text{Y}})$$

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

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

1.3 Part c)

$$\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}} \sim N(p_{\mathrm{X}} - p_{\mathrm{Y}}, \frac{p_{\mathrm{X}}(1-p_{\mathrm{X}})}{n_{\mathrm{X}}} + \frac{p_{\mathrm{Y}}(1-p_{\mathrm{Y}})}{n_{\mathrm{Y}}})$$

Under H_0, or assuming H_0:

$$\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}} \sim N(0, \frac{p(1-p)}{n_{\mathrm{X}}} + \frac{p(1-p)}{n_{\mathrm{Y}}})$$

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

$$T = \frac{\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}}}{\hat{P}(1-\hat{P})(\frac{1}{n_{\mathrm{X}}} + \frac{1}{n_{\mathrm{Y}}})} \sim N(0,1)$$

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

$$\text{p-value} = \mathrm{P}(\{\hat{P}_{\mathrm{X}} - \hat{p}_{\mathrm{X}} \geq \hat{P}_{\mathrm{Y}} - \hat{p}_{\mathrm{Y}}\})$$

In a slightly more useful arrangement:

$$\text{p-value} = \mathrm{P}(\{\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}} \geq \hat{p}_{\mathrm{X}} - \hat{p}_{\mathrm{Y}}\})$$

$$\text{p-value} = \mathrm{P}(\{\frac{\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}}}{\hat{P}(1-\hat{P})(\frac{1}{n_{\mathrm{X}}} + \frac{1}{n_{\mathrm{Y}}})} \geq \frac{\hat{p}_{\mathrm{X}} - \hat{p}_{\mathrm{Y}}}{\hat{P}(1-\hat{P})(\frac{1}{n_{\mathrm{X}}} + \frac{1}{n_{\mathrm{Y}}})}\})$$

Note $\hat{p}_{\mathrm{X}} - \hat{p}_{\mathrm{Y}} \approx 0.16939$ and $\hat{P}(1-\hat{P})(\frac{1}{n_{\mathrm{X}}} + \frac{1}{n_{\mathrm{Y}}}) \approx 0.01676$, hence:

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

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

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

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.