STAT1301 Assignment 4

Caleb Yates s49886351

October 14, 2025

1 Question 4

Let X be the random variable for the number of people who 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: $\mathrm{H}_0 : p_{\mathrm{X}} = p_{\mathrm{Y}} = p$. The alternative hypothesis is therefore: $\mathrm{H}_1 : p_{\mathrm{X}} > p_{\mathrm{Y}}$.

1.2 Part b)

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

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

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

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

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

Under H_0:

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

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

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

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

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

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

Under H_0:

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

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: $\bar{x} = \hat{p}_{\mathrm{X}} = \frac{505}{1415} \approx 0.3568$ and $\bar{y} = \hat{p}_{\mathrm{Y}} = \frac{264}{1408} \approx 0.1875$

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 $\hat{P} = \frac{505+264}{1415+1408} \approx 0.2724$. Rearranging gives:

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

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! $\Phi(3.69) = 0.9999$, and $\Phi(10.1079) > \Phi(3.69)$, 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:

$$\hat{P}_{\mathrm{X}} - \hat{P}_{\mathrm{Y}} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_{\mathrm{X}}(1-\hat{P}_{\mathrm{X}})}{n_{\mathrm{X}}} + \frac{\hat{P}_{\mathrm{Y}}(1-\hat{P}_{\mathrm{Y}})}{n_{\mathrm{Y}}}}$$

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

$$\mathrm{P}(\{\mathrm{Z} > z_{\frac{\alpha}{2}}\}) = \frac{\alpha}{2} = 0.015$$
$$\mathrm{P}(\{\mathrm{Z} < z_{\frac{\alpha}{2}}\}) = 0.985$$

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

$$\hat{p}_{\mathrm{X}} - \hat{p}_{\mathrm{Y}} \pm 2.17 \sqrt{\frac{\hat{p}_{\mathrm{X}}(1-\hat{p}_{\mathrm{X}})}{n_{\mathrm{X}}} + \frac{\hat{p}_{\mathrm{Y}}(1-\hat{p}_{\mathrm{Y}})}{n_{\mathrm{Y}}}}$$

Evaluating yields the CI $(0.1337, 0.2051)$ for $p_{\mathrm{X}} - p_{\mathrm{Y}}$. This confidence interval does not contain 0, therefore this is evidence that the proportions p_{X} and p_{Y} are not equal.

1.5 Part e)

$$\mathrm{X} \sim \mathrm{Bin}(n_{\mathrm{X}}, p_{\mathrm{X}})$$
$$\mathrm{Y} \sim \mathrm{Bin}(n_{\mathrm{Y}}, p_{\mathrm{Y}})$$

Under H_0 $p_{\mathrm{X}} = p_{\mathrm{Y}} = p$

$$\mathrm{Var}(\mathrm{X}) = n_{\mathrm{X}} p(1-p)$$
$$\mathrm{Var}(\mathrm{Y}) = n_{\mathrm{Y}} p(1-p)$$

$$\hat{P}^w = w\frac{\mathrm{X}}{n_{\mathrm{X}}} + (1-w)\frac{\mathrm{Y}}{n_{\mathrm{Y}}}$$

To show $\mathrm{E}(\hat{P}^w) = p$ under H_0:

$$\begin{aligned}
\mathrm{E}(\hat{P}^w) &= \mathrm{E}(w\frac{\mathrm{X}}{n_{\mathrm{X}}}) + \mathrm{E}((1-w)\frac{\mathrm{Y}}{n_{\mathrm{X}}}) \\
&= w\mathrm{E}(\frac{\mathrm{X}}{n_{\mathrm{X}}}) + (1-w)\mathrm{E}(\frac{\mathrm{Y}}{n_{\mathrm{Y}}}) \\
&= wp + (1-w)p \\
&= (w + 1 - w)p \\
&= p
\end{aligned}$$

To show the variance, we must additionally assume that X and Y are independant:

$$\begin{aligned}
\mathrm{Var}(\hat{P}^w) &= \mathrm{Var}(w\frac{\mathrm{Y}}{n_\mathrm{X}} + (1-w)\frac{\mathrm{Y}}{n_\mathrm{Y}}) \\
&= w^2\mathrm{Var}(\frac{\mathrm{X}}{n_\mathrm{X}}) + (1-w)^2\mathrm{Var}(\frac{\mathrm{Y}}{n_\mathrm{Y}}) \\
&= w^2\frac{p(1-p)}{n_\mathrm{X}} + (1-w)^2\frac{p(1-p)}{n_\mathrm{Y}} \\
&= p(1-p)\left(\frac{w^2}{n_\mathrm{X}} + \frac{(1-w)^2}{n_\mathrm{Y}}\right)
\end{aligned}$$

1.6 Part f)

To find if a certain w value minimuzes $\hat{P}^w$, I will take the derivative to find if it is a stationary point and the double deriviative to confirm it is a minimum.

$$\frac{\mathrm{d}}{\mathrm{d}w}\hat{P}^w = \frac{\mathrm{d}}{\mathrm{d}w}p(1-p)\left(\frac{w^2}{n_\mathrm{X}} + \frac{(1-w)^2}{n_\mathrm{Y}}\right)$$

To solve this derivative sensibly, I will assume $0 = \frac{\mathrm{d}p}{\mathrm{d}w} = \frac{\mathrm{d}n_\mathrm{X}}{\mathrm{d}w} = \frac{\mathrm{d}n_\mathrm{Y}}{\mathrm{d}w}$, as in, our choice of w is independant of p, n_X and n_Y.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}w}\hat{P}^w &= p(1-p)\left[\frac{1}{n_\mathrm{X}}\frac{\mathrm{d}}{\mathrm{d}w}(w^2) + \frac{1}{n_\mathrm{Y}}\frac{\mathrm{d}}{\mathrm{d}w}((1-w)^2)\right] \\
&= p(1-p)\left(\frac{2w}{n_\mathrm{X}} + \frac{-2(1-w)}{n_\mathrm{Y}}\right) \\
&= 2p(1-p)\left(\frac{w}{n_\mathrm{X}} + \frac{w}{n_\mathrm{Y}} - \frac{1}{n_\mathrm{Y}}\right) \\
\frac{\mathrm{d}^2}{\mathrm{d}w^2}\hat{P}^w &= 2p(1-p)\left(\frac{1}{n_\mathrm{X}}\frac{\mathrm{d}}{\mathrm{d}w}(w) + \frac{1}{n_\mathrm{Y}}\frac{\mathrm{d}}{\mathrm{d}w}(w) - \frac{1}{n_\mathrm{Y}}\right) \\
&= \frac{2p(1-p)}{n_\mathrm{X}}
\end{aligned}$$

Substituting $w = \frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}$:

$$\begin{aligned}
\left.\frac{\mathrm{d}}{\mathrm{d}w}\hat{P}^w\right|_{w=\frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}} &= 2p(1-p)\left(\frac{\frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}}{n_\mathrm{X}} + \frac{\frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}}{n_\mathrm{Y}} - \frac{1}{n_\mathrm{Y}}\right) \\
&= 2p(1-p)\left(\frac{1}{n_\mathrm{Y}+n_\mathrm{X}} + \frac{n_\mathrm{X}}{n_\mathrm{Y}(n_\mathrm{Y}+n_\mathrm{X})} - \frac{1}{n_\mathrm{Y}}\right) \\
&= 2p(1-p)\left(\frac{n_\mathrm{Y}}{n_\mathrm{Y}(n_\mathrm{Y}+n_\mathrm{X})} + \frac{n_\mathrm{X}}{n_\mathrm{Y}(n_\mathrm{Y}+n_\mathrm{X})} - \frac{n_\mathrm{X}+n_\mathrm{Y}}{n_\mathrm{Y}(n_\mathrm{Y}+n_\mathrm{X})}\right) \\
&= 2p(1-p)(0) \\
&= 0
\end{aligned}$$

Therefore $w = \frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}$ is a stationary point of $\hat{P}^w$. To prove it is a minimum:

$$\frac{\mathrm{d}^2}{\mathrm{d}w^2}\bigg|_{w=\frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}} = \frac{2p(1-p)}{n_\mathrm{X}}$$

Since $p \geq 0$ and $n_\mathrm{X} > 0$, $\frac{\mathrm{d}^2}{\mathrm{d}w^2}\hat{P}^w > 0$ at $w = \frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}$, therefore $w = \frac{n_\mathrm{X}}{n_\mathrm{Y}+n_\mathrm{Y}}$ is a minimum for $\hat{P}^w$.