Two random variables

In the context of an experiment, the success rate in group A and B are Bernoulli random variables with expected value μ_A, μ_B and variance ${\sigma^2}_A, {\sigma^2}_B$ respectively :

$A \sim Bernoulli(P_A)$ with $\mu_A = P_A$ and ${\sigma^2}_A = P_A(1 - P_A)$

$B \sim Bernoulli(P_B)$ with $\mu_B = P_B$ and ${\sigma^2}_B = P_B(1 - P_B)$

Combined random variables

If we want to compare the success rate between two Bernoulli random variables A and B, we can create a random variable $C = A - B$ (which mean is expected to be zero). Assuming A and B are IID :

$$\mu_C = \mu_A - \mu_B$$

$${\sigma^2}_C = {\sigma^2}_A + {\sigma^2}_B$$

According to the central limit theorem, if $C_1, C_2, \ldots$ are random samples each of size n taken from C, then the sampling distribution of means $\bar{C}$ will be approximately normal for large sample sizes (over 30) with the following statistical properties.

$$\mu_{\bar{C}} = \mu_C$$

$${\sigma^2}_{\bar{C}} = \frac{{\sigma^2}_C}{n}$$

Therefore :

$$\bar{C} \sim N(\mu_{\bar{C}}, {\sigma^2}_{\bar{C}})$$

$$\bar{C} \sim N(\mu_C, \frac{{\sigma^2}_C}{n})$$

$$\bar{C} \sim N(\mu_A - \mu_B, \frac{{\sigma^2}_A}{n_A} + \frac{{\sigma^2}_B}{n_B})$$

Hypothesis test

We can now set the hypothesis for our test :

$$H_0 : \mu_{\bar{C}} = 0$$

$$H_1 : \mu_{\bar{C}} \neq 0$$

$$Z = \frac{\hat{P}_A - \hat{P}_B - 0}{\sqrt{\frac{{\sigma^2}_A}{n_A} + \frac{{\sigma^2}_B}{n_B}}}$$

Since the variance of both random variables is the same under the null hypothesis, we can rewrite the test statistic using a pooled variance based on the Satterthwaite Approximation :

$$Z = \frac{\hat{P}_A - \hat{P}_B}{\sqrt{\frac{{\sigma^2}_p}{n_A} + \frac{{\sigma^2}_p}{n_B}}} = \frac{\hat{P}_A - \hat{P}_B}{\sqrt{\frac{P_p(1 - P_p)}{n_A} + \frac{P_p(1 - P_p)}{n_B}}} = \frac{\hat{P}_A - \hat{P}_B}{\sqrt{P_p(1 - P_p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Where P_p is the weighted average of P_A and P_B

$$P_p = \frac{P_A n_A + P_B n_B}{n_A + n_B}$$

We approximate the probabilities P_A and P_B with their empirical equivalent $\hat{P}_A \approx P_A$ and $\hat{P}_B \approx P_B$ to compute the pooled probability $\hat{P}_p \approx P_p$ and the Z score.