Setup

We consider an evaluation setup with

k questions, n_0 trials for the old system, and n_1 trials for the new system.

Old system: binary outcomes per trial and question

For each trial $i = 1, \ldots, n_0$ and each question $j = 1, \ldots, k$, let

$$x_{ij} \in \{0, 1\}$$

be a binary indicator that the answer to question j in trial i is correct for the *old system*.

For each question j, the empirical probability of correctness is

$$\hat{p}_{0,j} = \frac{1}{n_0} \sum_{i=1}^{n_0} x_{ij}.$$

For each trial i, define the total number of correct answers as

$$S_i^{(0)} = \sum_{j=1}^{k} x_{ij}.$$

The average number of correct answers per trial for the old system is

$$\overline{S}^{(0)} = \frac{1}{n_0} \sum_{i=1}^{n_0} S_i^{(0)}.$$

New system: binary outcomes per trial and question

For the *new system*, we have n_1 independent trials. For each trial $i = 1, \ldots, n_1$ and each question $j = 1, \ldots, k$, let

$$y_{ij} \in \{0, 1\}$$

be a binary indicator that the answer to question j in trial i is correct.

For each question j, the empirical probability of correctness for the new system is

$$\hat{p}_{1,j} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{ij}.$$

For each trial i, define the total number of correct answers as

$$S_i^{(1)} = \sum_{j=1}^{k} y_{ij}.$$

The average number of correct answers per trial for the new system is

$$\overline{S}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} S_i^{(1)}.$$

Main question

Our goal is to test whether the average number of correct answers per trial differs between the new and the old systems. Formally, we are interested in whether

$$\overline{S}^{(1)} \text{ is significantly different from } \overline{S}^{(0)}.$$

Equivalently, we are testing the null hypothesis

$$H_0 : \mathbb{E}[S^{(1)}] = \mathbb{E}[S^{(0)}] \qquad \text{vs.} \qquad H_1 : \mathbb{E}[S^{(1)}] \neq \mathbb{E}[S^{(0)}].$$

Test statistic

We want to compare the mean total number of correct answers per trial between the old and new systems. The natural test statistic is

$$T = \frac{\overline{S}^{(1)} - \overline{S}^{(0)}}{SE\big(\overline{S}^{(1)} - \overline{S}^{(0)}\big)}.$$

(One may also use the opposite sign $\overline{S}^{(0)} - \overline{S}^{(1)}$; the resulting T will just have the opposite sign but the same absolute value.)

Step 1: Variance of the sample means

For the old system, recall

$$\overline{S}^{(0)} = \frac{1}{n_0} \sum_{i=1}^{n_0} S_i^{(0)}, \qquad S_i^{(0)} = \sum_{j=1}^{k} x_{ij}.$$

Assuming trials are independent,

$$\mathrm{Var}\big(\overline{S}^{(0)}\big) = \frac{1}{n_0}\,\mathrm{Var}\big(S_i^{(0)}\big).$$

Similarly, for the new system,

$$\mathrm{Var}\big(\overline{S}^{(1)}\big) = \frac{1}{n_1}\,\mathrm{Var}\big(S_i^{(1)}\big).$$

Step 2: Variance of the difference of means

Because the samples for the old and new systems are independent, the variance of the difference is

$$\mathrm{Var}\big(\overline{S}^{(1)} - \overline{S}^{(0)}\big) = \mathrm{Var}\big(\overline{S}^{(1)}\big) + \mathrm{Var}\big(\overline{S}^{(0)}\big).$$

Step 3: Variance of a single trial total score

Under the assumption that within a single trial, question outcomes x_{ij} are independent across j, we have

$$\mathrm{Var}\big(S_i^{(0)}\big) = \sum_{j=1}^{k} \mathrm{Var}(x_{ij}) = \sum_{j=1}^{k} p_{0,j}\big(1 - p_{0,j}\big).$$

Similarly,

$$\mathrm{Var}\big(S_i^{(1)}\big) = \sum_{j=1}^{k} \mathrm{Var}(y_{ij}) = \sum_{j=1}^{k} p_{1,j}\big(1 - p_{1,j}\big).$$

Step 4: Standard error of the difference

Therefore,

$$SE\big(\overline{S}^{(1)} - \overline{S}^{(0)}\big) = \sqrt{\frac{1}{n_1}\sum_{j=1}^{k} p_{1,j}\big(1 - p_{1,j}\big) + \frac{1}{n_0}\sum_{j=1}^{k} p_{0,j}\big(1 - p_{0,j}\big)}.$$

Step 5: Final test statistic

Putting everything together, the test statistic is

$$T = \frac{\overline{S}^{(1)} - \overline{S}^{(0)}}{\sqrt{\frac{1}{n_1}\sum_{j=1}^{k} p_{1,j}(1-p_{1,j}) + \frac{1}{n_0}\sum_{j=1}^{k} p_{0,j}(1-p_{0,j})}}$$

$$= \frac{\sum_{j=1}^{k} \hat{p}_{1,j} - \sum_{j=1}^{k} \hat{p}_{0,j}}{\sqrt{\frac{1}{n_1}\sum_{j=1}^{k} p_{1,j}(1-p_{1,j}) + \frac{1}{n_0}\sum_{j=1}^{k} p_{0,j}(1-p_{0,j})}}.$$

Practical note: small n_1 case

When the number of trials for the new system n_1 is sufficiently large, the variance term $\sum_{j=1}^{k} p_{1,j}(1-p_{1,j})$ can be estimated reliably using the empirical probabilities $\hat{p}_{1,j}$.

However, when n_1 is small (say $n_1 \leq 5$), these estimates may be very noisy, leading to unstable standard error calculations and unreliable test statistics. In this case, a common and reasonable approximation is to *substitute the variance of the new system with that of the old system*, i.e.

$$p_{1,j}(1-p_{1,j}) \;\approx\; p_{0,j}(1-p_{0,j}).$$

This yields the following modified standard error:

$$SE_{\text{small-}n_1}\big(\overline{S}^{(1)} - \overline{S}^{(0)}\big) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_0}\right)\sum_{j=1}^{k} p_{0,j}\big(1-p_{0,j}\big)}.$$

Correspondingly, the test statistic becomes

$$T_{\text{small-}n_1} = \frac{\overline{S}^{(1)} - \overline{S}^{(0)}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_0}\right)\sum_{j=1}^{k} p_{0,j}(1-p_{0,j})}}.$$

This approximation is justified when:

- The new system is not expected to have drastically different per-question correctness probabilities compared to the old system;
- n_1 is too small to estimate $\hat{p}_{1,j}$ reliably (e.g. only a few trials);
- A conservative estimate of variance is preferred.