Comparing two population proportions.

- 問題: 同一產品有二生產線A和B可生產, 不良率分別爲p_1, p_2. 想知道是否$p_1 = p_2$. 假設觀察到A生產線生產n_1個產品, 其中X個是不良品, 而B生產線生產n_2個產品, 其中Y個是不良品. 希望根據 X 和 Y來檢定是否$p_1 = p_2$.

- 上述問題可改寫如下: Suppose that X and Y are independent, $X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$. We want to test

$$H_0 : p_1 = p_2 \text{ versus } H_1 : p_1 \neq p_2 \tag{1}$$

based on X and Y.

- Let

$$Z = \frac{X/n_1 - Y/n_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{X+Y}{n_1+n_2}\right)\left(1 - \frac{X+Y}{n_1+n_2}\right)}} \tag{2}$$

and

$$Z_0 = \frac{X/n_1 - Y/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

When n_1 and n_2 are both large, the distribution of Z_0 is approximated $N(0,1)$, so under $H_0 : p_1 = p_2$, the distribution of Z is approximated $N(0,1)$. The (approximate) z test rejects $H_0 : p_1 = p_2$ at level a if and only if

$$|Z| > z_{a/2}.$$

The size of the approximate z test is approximately a for large n_1, n_2. The p-value for the test is $2P(N(0,1) > \text{ observed } |Z|\ )$.

  - The exact distribution of Z under H_0 depends on $p = p_1 = p_2$.
  - $Z_0 \approx N(0,1)$ since for $X_0 \sim Bin(n,p)$ and for large n,

$$\frac{X_0/n - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

  and $X_0/n \approx p$.

- Example 1. Suppose that a squirrel usually hides nuts in two different regions: A and B. Suppose that we observed that the squirrel hid nuts 300 times, 100 times in Region A and 200 times in Region B. Moreover, in Region A, 40 out of the 100 times the nuts were lost. In Region B, 100 out of the 200 times the nuts were lost. Do we have strong evidence that the probabilities of losing nuts are different in the two regions?

  Sol. The observed $|Z|$ value is

$$\frac{|40/100 - 100/200|}{\sqrt{(1/100 + 1/200)(140/300)(1 - 140/300)}} = 1.636634 \approx 1.64.$$

  The p-value is

$$2P(N(0,1) > 1.64) = 2 \times (0.5 - 0.4495) = 0.101 > 0.05,$$

  so we do not have strong evidence to conclude that the two probabilities are different.

- One-side testing problems. Suppose that X and Y are independent, $X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$. In addition to the testing problem in (1), we can also consider the testing problem

$$H_0 : p_1 \le p_2 \text{ versus } H_1 : p_1 > p_2 \tag{3}$$

and the testing problem

$$H_0 : p_1 \ge p_2 \text{ versus } H_1 : p_1 < p_2 \tag{4}$$

based on X and Y. Let Z be the statistic defined in (2).

For the testing problem in (3), the approximate z test rejects $H_0 : p_1 \le p_2$ at level a if and only if

$$Z > z_a,$$

and the p-value for the test is $P(N(0,1) > \text{ observed } Z\ )$.

For the testing problem in (4), the approximate z test rejects $H_0 : p_1 \le p_2$ at level a if and only if

$$Z < -z_a,$$

and the p-value for the test is

$$P(N(0,1) > -\text{ observed } Z\ ) = P(N(0,1) < \text{ observed } Z\ ).$$

Example 2. In Example 1, do we have some evidence that the probability of losing nuts in Region B is higher (than the probability of losing nuts in Region A)?

The observed Z value is

$$\frac{100/200 - 40/100}{\sqrt{(1/100 + 1/200)(140/300)(1 - 140/300)}} = 1.636634 \approx 1.64.$$

The p-value is

$$P(N(0,1) > 1.64) = (0.5 - 0.4495) = 0.0505 < 0.1,$$

so there is some evidence that the probability of losing nuts in Region B is higher.