Two-sample z-test for proportions

Use case

We have two binary categorical variables. One identifies membership in one of two independent (i.e., disjoint) populations. The other represents our variable of interest. We want to perform inference on the true proportion of each population that falls into one value of the variable of interest (e.g., “yes” on a survey). Our true proportions for popuplation 1 and 2 are denoted by p_1 and p_2, respectively.

We draw a sample of size n_1 from population 1, yielding X_1 successes. Let $\hat{p}_1 = \frac{X_1}{n_1}$. $\hat{p}_1$ is an *unbiased estimator* of p_1.

We draw a sample of size n_2 from population 2, yielding X_2 successes. Let $\hat{p}_2 = \frac{X_2}{n_2}$. $\hat{p}_2$ is an *unbiased estimator* of p_2.

Preconditions

Preconditions for hypothesis testing

For the normal approximation used in this test to be valid, n_1 and n_2 should be large enough such that:

1. $n_1 p_1 \geq 10$
2. $n_1(1 - p_1) \geq 10$
3. $n_2 p_2 \geq 10$
4. $n_2(1 - p_2) \geq 10$.

Preconditions for confidence interval

For a confidence interval to be valid, the number of successes and failures in each sample should be at least 10.

Hypotheses

Null hypothesis, H_0**:** $p_1 - p_2 = 0$

Alternative hypothesis, H_a**:** $p_1 - p_2 \neq / > / < 0$

Test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}, \text{ where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

p-value

Where Z is a standard normal random variable,

1. For $H_a: p_1 - p_2 < 0$:
   $p = P(Z \leq z)$
2. For $H_a: p_1 - p_2 > 0$:
   $p = P(Z \geq z)$
3. For $H_a: p_1 - p_2 \neq 0$:
   $p = 2P(Z \leq -|z|)$

p-value (Python)

1. For $H_a: p_1 - p_2 < 0$: `p = stats.norm.cdf(z)`
2. For $H_a: p_1 - p_2 > 0$: `p = 1 - stats.norm.cdf(z)`
3. For $H_a: p_1 - p_2 \neq 0$: `p = 2 * stats.norm.cdf(-abs(z))`

Associated confidence interval

$$\text{C\% confidence interval} = (\hat{p}_1 - \hat{p}_2) \pm z^\star \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\text{choose } z^\star \text{ s.t. area on standard normal distribution from } (-z^\star, z^\star) = \text{C}$$

See the preconditions for the confidence interval.

Minimum sample size to achieve margin of error:

$$n = (\frac{z^\star}{m^\star})^2(p_1^\star(1-p_1^\star) + p_2^\star(1-p_2^\star))$$

$p_1^\star$ is an “educated guess” about the value of p_1 ($p_2^\star$ and p_2, respectively). If you have access to a sample proportion, $\hat{p}_1$, set $p_1^\star = \hat{p}_1$. Otherwise, a *conservative approach* is to set $p_1^\star = p_2^\star = 0.5$.

Related tests

If you want to perform inference on the proportion of just one population, a one-sample z-test for proportions is appropriate.

If you want to perform inference about the proportions of a non-binary categorical variable, a chi-square test of association is appropriate.

Considerations

A two-sample z-test for proportions conducted at confidence level α will reject the null hypothesis if and only if the value corresponding to the null hypothesis,

0, is completely outside of the
$C = 1 - \alpha$ confidence interval for the true difference between
p_1 and
p_2.

Neither the z-test nor the confidence interval use the standard deviation of the difference sample proportion,
$\sigma_{\hat{p}_1 - \hat{p}_2}$.

1. The hypothesis test uses the standard error under the null hypothesis, $SE^{H_0}_{\hat{p}_1 - \hat{p}_2}$.
2. The confidence interval uses the standard error of the difference in sample proportions, $SE_{\hat{p}_1 - \hat{p}_2}$.