Two samples z-test for difference of two proportions

Function of the test: This test allows us to compare two population proportions of success i.e., to know if they are really different or not.

Test assumptions:

1. Measurements are **categorical** (e.g., Yes or No to the question "Do you smoke?") and taken from two distinct groups (e.g., Female, Male)
2. Samples are drawn randomly from each population using SRS
3. Samples are independent
4. Both samples are large i.e., at least 30 observations are drawn in each sample, which guarantee the normal distribution of sampling distribution of difference of sample proportion of successes.

Hypothesis to test:

Let π_1 and π_2 be the proportion of success in population I and II respectively. Then we have following hypothesis testing problems:

Two-tailed test	Left tailed test	Right tailed test
H_0: $\pi_1 = \pi_2$	H_0: $\pi_1 \geq \pi_2$	H_0: $\pi_1 \leq \pi_2$
H_1: $\pi_1 \neq \pi_2$	H_1: $\pi_1 < \pi_2$	H1: $\pi_1 > \pi_2$

Test Statistic

The appropriate formula for the test statistic comparing two proportions is given by,

$$Z = \frac{\text{(observed difference in sample proportions) - (hypothesized difference of proportions)}}{\text{Statndard error of difference of proportions}}$$

$$= \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where,

p_1 = X_1/n_1 = proportion of success for sample 1

p_2 = X_2/n_2 = proportion of success for sample 2

n_1 = Number of observations in sample 1

n_2 = Number of observations in sample 2

$\hat{\pi}$ = Pooled estimate of common proportion of success under H_0

$$= \frac{X_1+X_2}{n_1+n_2} = \frac{n_1p_1+n_2p_2}{n_1+n_2}$$

$1-\hat{\pi}$ = Pooled estimate of common proportion of failure

Problem: Z test for difference of population proportion

In a study of obesity, the following results were obtained from samples of males and females between the ages of 20 and 35.

	Sample size	Number overweight
Males	150	21
Females	200	48

Can we conclude from these data that in the sampled population there is a difference in the proportions that are overweight? Let $\alpha = 0.05$.

Solution

Each population is dichotomized as:

- overweight people (success)
- normal weight people (failure)

Let X_1 and X_2 be the no. of overweight in sample 1 and sample 2

Given,

$n_1 = 150$ $X_1 = 21$

$n_2 = 200$ $X_2 = 48$

Step 1: Setting up Null and Alternative hypothesis

Let π_1 and π_2 be the proportion of overweight (success) in population I (male) and II (female) respectively

Null Hypothesis

H_0: $\pi_1 = \pi_2$ (There is no significant difference in the proportion of overweight among population of male and female)

Alternative Hypothesis

H_1: $\pi_1 \neq \pi_2$ (There is significant difference in the proportion of overweight among population of male and female)

Step 2: Choice of α for the test

The level of significance = Probability of making type I error employed for the test = 0.05

Step 3: Test Statistic

The appropriate formula for the test statistic comparing two proportions is given by,

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where,

p_1 = proportion of success for sample 1

p_2 = proportion of success for sample 2

n_1 = Number of observations in sample 1

n_2 = Number of observations in sample 2

$\hat{\pi}$ = Pooled estimate of common proportion of success under H_0

$$= \frac{X_1+X_2}{n_1+n_2} = \frac{n_1\, p_1 + n_2\, p_2}{n_1+n_2}$$

$1 - \hat{\pi}$ = Pooled estimate of common proportion of failure

The test statistic has standard normal distribution with mean = 0 and standard deviation = 1

Step 4: Critical or Tabulated Z

The test is two sided and the given level of α for the test is 5 %.

The test has two rejection regions

The critical value of Z from the Z table is given by,

$Z_c = 1.96$

Acceptance Region: $-1.96 < Z < +1.96$

Rejection Region: $Z \geq = +1.96$ OR $Z \leq -1.96$

Step 5: Calculated or Observed Z

p_1 = proportion of overweight among males in the sample = X_1/n_1 = 21 / 150 = 0.14

p_2 = Proportion of overweight among females in the sample = X_2/n_2 = 48 / 200 = 0.24

The pooled estimate of common proportion of success is given by,

$$\hat{\pi} = \frac{X_1+X_2}{n_1+n_2} = \frac{21+48}{150+200} = 0.1971$$

$1 - \hat{\pi} = 1 - 0.1971 = 0.8029$

Now, calculated z is given by,

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.14-0.24}{\sqrt{0.1971 \; x \; 0.8029\left(\frac{1}{150}+\frac{1}{200}\right)}} = -\,2.3273$$

Hence, cal Z = - 2.3273

Step 6: Statistical Decision

Since cal z = - 2.3273 < critical Z = -1.96 (observed z falls in the lower critical region) we reject our null hypothesis in favor of alternative hypothesis at 5 % level of significance.

Step 7: Conclusion

There is significant difference in the proportion of overweight among population of male and female. Since p_1 = 0.14 and p_2 = 0.24, we can say proportion of overweight among females is significantly higher than that among males.