Two sample test for proportions

In the videos, you learnt how to test if the means from two populations were different from one another. In this reading item, you will see a similar test but for comparing proportions from two different populations. A common application of two sample test for proportions is in A/B testing.

An example

Imagine you want to compare the proportion of households that own a car in Chicago (p_1) with the proportion of households that do in New York (p_2). Note that p_1 and p_2 are **population** proportions.

A possible set of hypothesis for this problem is

$$H_0 : p_1 = p_2 \text{ vs. } H_1 = p_1 \neq p_2$$

Consider for this test a significance level of 0.05

Suppose you randomly sample $n_1 = 100$ households from Chicago, 62 of which own a car, and $n_2 = 120$ households from New York, 58 of which own a car.

Defining X ="number of households that own a car in Chicago" and Y ="number of households that own a car in New York: , a good approximation for p_1 and p_2 are

$$\hat{p}_1 = \frac{X}{100} \quad \hat{p}_2 = \frac{Y}{120}$$

Naturally, a good approximation for $\Delta = p_1 - p_2$ is

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2$$

In order to get a good test statistic, you need to find the distribution of $\hat{\Delta}$. First, note that in n_1 and n_2 are big enough, which in this case they are.

$$\hat{p}_1 = \frac{X}{100} \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{100}\right) \quad \hat{p}_2 = \frac{Y}{120} \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{120}\right)$$

so that $\hat{\Delta} = \hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{100} + \frac{p_2(1-p_2)}{120}\right)$

If H_0 is true, then $p_1 = p_2 = p$. This simplifies the expression quite a bit! In this case,

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, \frac{p(1-p)}{100} + \frac{p(1-p)}{120}\right) = N\left(0, p(1-p)\left(\frac{1}{100} + \frac{1}{120}\right)\right)$$

Standardizing, you get that

$$\frac{\frac{X}{100} - \frac{Y}{120} - 0}{\sqrt{p(1-p)\left(\frac{1}{100} + \frac{1}{120}\right)}} \sim \mathcal{N}(0,1)$$

Unfortunately, this statistic is not good enough, because even if H_0 is true, you do not know the value of p. However, you can replace it by the aggregated sample proportion. If H_0 is true, and $p_1 = p_2 = p$, then $X \sim Binomial(100, p)$ and $Y \sim Binomial(120, p)$, so you can use both samples to get a better estimate of p:

$$\hat{p} = \frac{X + Y}{100 + 120} = \frac{X + Y}{220}$$

Replacing p with $\hat{p}$ you finally get the test statistic

$$Z = \frac{\frac{X}{100} - \frac{Y}{120} - 0}{\sqrt{\frac{(X+Y)}{220}(1 - \frac{X+Y}{220})\left(\frac{1}{100} + \frac{1}{120}\right)}} \sim \mathcal{N}(0,1)$$

To simplify the expression even further, observe that $\left(\frac{1}{100} + \frac{1}{120}\right) = \left(\frac{100+120}{100 \cdot 120}\right)$, so that you can rewrite the test statistic as

$$Z = \frac{\frac{X}{100} - \frac{Y}{120} - 0}{\sqrt{(X+Y)(1 - \frac{X+Y}{220})}} \sqrt{100 \cdot 120} \sim \mathcal{N}(0,1)$$

With the observations you have ($x = 62,\ y = 58$), the observed statistic results $z = 2.0271$. Since you are proposing a two-sided test, then the p-value is the probability that $Z > 2.0271$ or $Z < -2.0271$:

$$p\text{-value} = \mathbf{P}(|Z| > 2.0271) = 0.04265$$

Conclusion: Since the p-value is smaller than the significance level (0.05), then you reach the conclusion that you have enough evidence to reject the null hypothesis, and accept that the two population proportions are different

General Case

You have two populations or groups you want to compare.

- $p_1 - p_2$, is the difference in the population proportion between two groups. (i.e. the difference in the proportion of households that own a car in Chicago and New York)
- x is the observed number of individuals in the sample from the specified category from one of the groups(i.e. number households that own a car in Chicago)
- y is the observed number of individuals in the sample from the specified category from one of the groups(i.e. number households that own a car in New York)
- n_1 is the sample size for group 1 (sample size from Chicago)
- n_2 is the sample size for group 2 (sample size from New York)

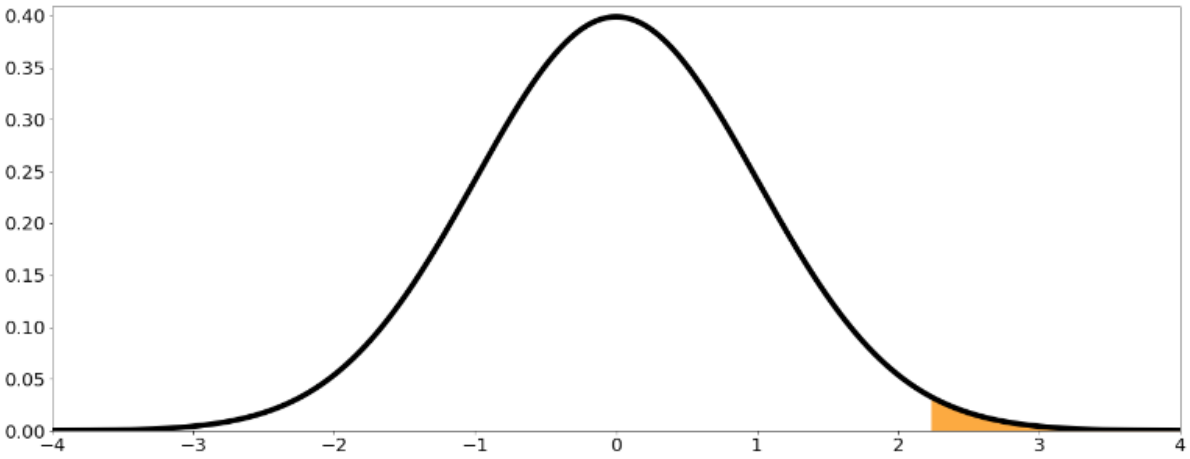
Then the test statistic is

$$Z = \frac{\frac{X}{n_1} - \frac{Y}{n_2} - 0}{\sqrt{(X+Y)(1 - \frac{X+Y}{n_1+n_2})}} \sqrt{n_1 \cdot n_2} \sim \mathcal{N}(0,1)$$

and the observed statistic is $z \frac{\frac{x}{n_1} - \frac{y}{n_2} - 0}{\sqrt{(x+y)(1 - \frac{x+y}{n_1+n_2})}} \sqrt{n_1 \cdot n_2} \sim \mathcal{N}(0,1)$.

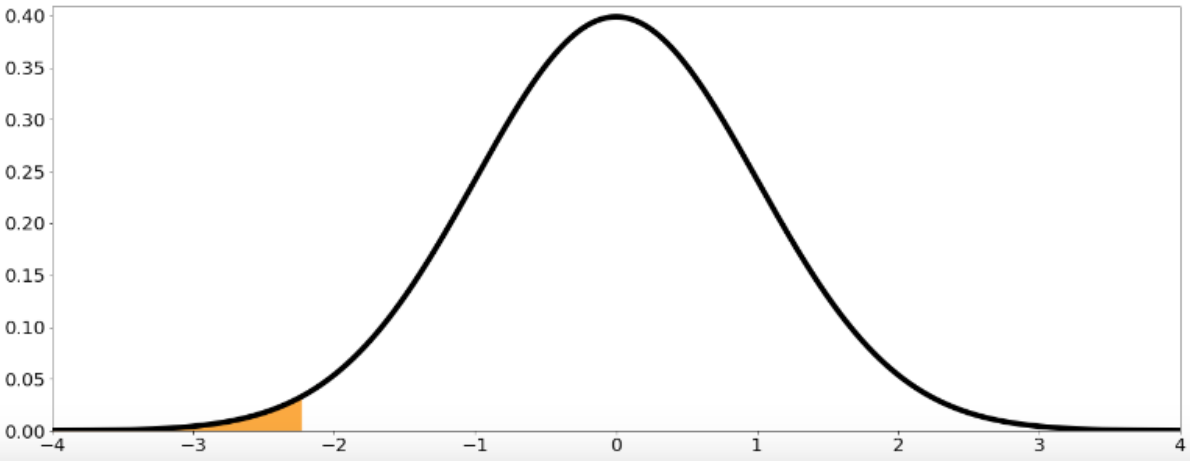
Depending on the type of hypothesis, you have different expressions for the p-value:

- Right-tailed test: $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$:
$$p\text{-value} = \mathbf{P}(Z > z)$$



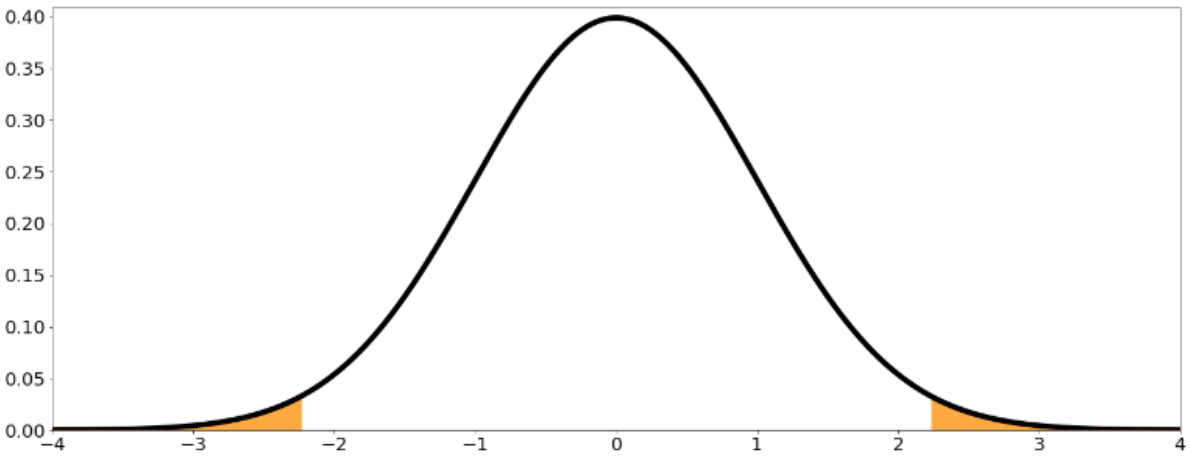
- Left-tailed test: $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 < 0$

$$p\text{-value} = \mathbf{P}(Z < z)$$



- Two-tailed test: $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$

$$p\text{-value} = \mathbf{P}(|Z| > |z|)$$



For this results to be valid, the following conditions need to be satisfied:

- There are two simple random samples that are independent from one another. This means that you have one sample from population 1, and another from population 2, and that the samples are independent between both groups.
- Each population size needs to be at least 20 times bigger than the sample size. This is necessary to ensure that all samples are independent.
- The individuals in each sample can be divided into two categories: wither they belong to the specified category or they don't
- Both sample sizes need to be at least 10. This condition needs to be verified so that the Gaussian approximation holds when the assumption that H_0 is true.