10.5: Comparing Two Independent Population Proportions

Proportions have their base in the binomial probability distribution. As a probability distribution, we know the mean and standard deviation of the distribution. As a binary or categorical sample data set, we lose this knowledge from when we knew the population parameters. We do not know the population mean, $\mu = np$, or variance, $\sigma^{\angle} = npq$. We can gather data that we know comes from a binary distribution but not know the specific parameter. It is then that we have moved from probability to inferential statistics.

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or even perhaps 20 times the size of the sample. This keeps each population from being over-sampled and causing biased results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance in the sampling. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the two population proportions.

Like the case of differences in sample means, we construct a sampling distribution for differences in sample proportions: $(p'_A - p'_B)$ where $p'_A = \frac{X_A}{n_A}$ and $p'_B = \frac{X'_B}{n_B}$ are the sample proportions for the two sets of data in question. X_A and X_B are the number of successes in each sample group respectively, and n_A and n_B are the respective sample sizes from the two groups. Again we go the Central Limit theorem to find the distribution of this sampling distribution for the differences in sample proportions. And again we find that this sampling distribution, like the ones past, are normally distributed as proved by the Central Limit Theorem, as seen in Figure $10.5.1$.

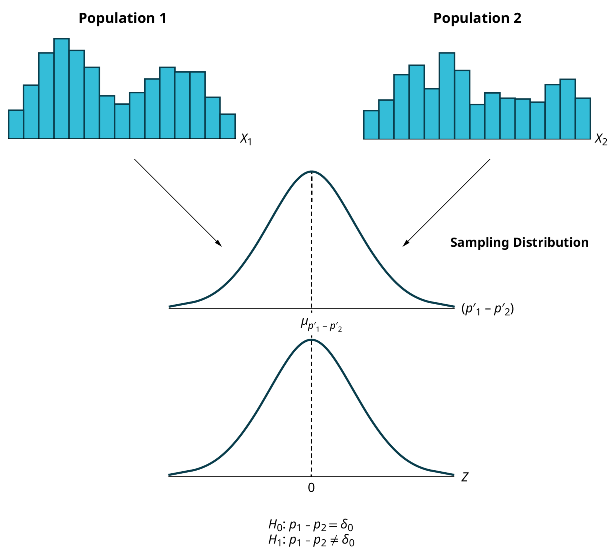

Figure 10.5.1: Copy and Paste Caption here. (Copyright; author via source)

Generally, the null hypothesis allows for the test of a difference of a particular value, δ_0, just as we did for the case of differences in means.

$$\begin{aligned} H_0 &: p_1 - p_2 = \delta_0 \\ H_1 &: p_1 - p_2 \neq \delta_0 \end{aligned} \tag{10.5.1}$$

Most common, however, is the test that the two proportions are the same. That is,

$$\begin{aligned} H_0 &: p_{\mathrm{A}} = p_B \\ H_a &: p_{\mathrm{A}} \neq p_B \end{aligned} \tag{10.5.2}$$

To conduct the test, we use a pooled proportion, p_c.
The pooled proportion is calculated as follows:

$$p_c=\frac{x_A+x_B}{n_A+n_B} \tag{10.5.3}$$

The test statistic (z-score) is:

\[Z_c=\dfrac{\left(p_A^{\prime}-p_B^{\prime}\right)-\delta_0}{\sqrt{p_c\left(1-p_c\right)\left(\dfrac{1}{n_A}+\dfrac{1}{n_B}\right)}}\]

where δ_0 is the hypothesized differences between the two proportions and p_c is the pooled variance from the formula above.

? Exercise 10.5.1

A bank has recently acquired a new branch and thus has customers in this new territory. They are interested in the default rate in their new territory. They wish to test the hypothesis that the default rate is different from their current customer base. They sample 200 files in area A, their current customers, and find that 20 have defaulted. In area B, the new customers, another sample of 200 files shows 12 have defaulted on their loans. At a 10% level of significance can we say that the default rates are the same or different?

Answer

This is a test of proportions. We know this because the underlying random variable is binary, default or not default. Further, we know it is a test of differences in proportions because we have two sample groups, the current customer base and the newly acquired customer base. Let A and B be the subscripts for the two customer groups. Then p_A and p_B are the two population proportions we wish to test.

Random Variable: $P_A^{\prime}-P_B^{\prime}=$ difference in the proportions of customers who defaulted in the two groups.

$$\begin{aligned} &H_0: p_A=p_B \\ &H_a: p_A \neq p_B \end{aligned} \tag{10.5.4}$$

The words "**is a difference**" tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c=\frac{x_A+x_B}{n_A+n_B}=\frac{20+12}{200+200}=0.08 \quad 1-p_c=0.92 \tag{10.5.5}$$

$\left(p_A^{\prime}-p_B^{\prime}\right)=0.04$ follows an approximate normal distribution.

Estimated proportion for group A: $p^{\prime}{ }_A=\frac{x_A}{n_A}=\frac{20}{200}=0.1$

Estimated proportion for group B: $p^{\prime}{ }_B=\frac{x_B}{n_B}=\frac{12}{200}=0.06$

The estimated difference between the two groups is : $p_A^{\prime}-p_B^{\prime}=0.1-0.06=0.04$.

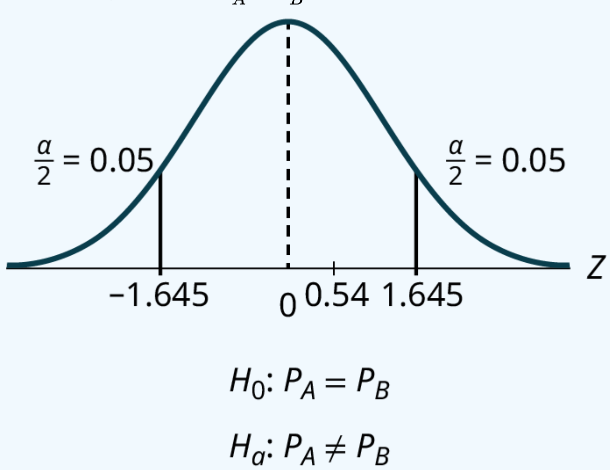

Figure 10.5.1: Copy and Paste Caption here. (Copyright; author via source)

$$Z_c = \frac{(\mathrm{P}'_A - \mathrm{P}'_B) - \delta_0}{\sqrt{P_c\,(1 - P_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = 1.47 \tag{10.5.6}$$

The calculated test statistic is 1.47 and is not in the tail of the distribution.

Make a decision: Since the calculate test statistic is not in the tail of the distribution we cannot reject H_0.

Conclusion: At a 1\% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference between the proportions of customers who defaulted in the two groups.

Try It 10.5.1

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.