Session 5

Difference Between Proportions

Many statistical applications involve comparisons between two independent sample proportions.

Difference Between Proportions: Theory

Suppose we have two populations with proportions equal to P_1 and P_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least 10 times bigger than their sample.)
- The samples from each population are big enough to justify using a normal distribution to model differences between proportions. The sample sizes will be big enough when the following conditions are met: $n_1P_1 \geq 10$, $n_1(1 - P_1) \geq 10$, $n_2P_2 \geq 10$, and $n_2(1 - P_2) \geq 10$.
- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.

Given these assumptions, we know the following.

- The set of differences between sample proportions will be normally distributed. We know this from the central limit theorem.
- The expected value of the difference between all possible sample proportions is equal to the difference between population proportions. Thus, $E(p_1 - p_2) = P_1 - P_2$.
- The standard deviation of the difference between sample proportions (σ_d) is approximately equal to:

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$$\sigma^2_d = \sigma^2_{P1 - P2} = \sigma^2_1 + \sigma^2_2$$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then

$$\sigma^2_1 = P_1(1 - P_1) / n_1 \qquad \text{And} \qquad \sigma^2_2 = P_2(1 - P_2) / n_2$$

Therefore,

$$\sigma^2_d = [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2]$$

And

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$

Difference Between Proportions: Sample Problem

In this section, we work through a sample problem to show how to apply the theory presented above. The approach presented is valid whenever we need to analyze differences between independent sample proportions. In this example, differences between proportions are modeled with a normal distribution

Problem 1

In one state, 52% of the voters are Republicans, and 48% are Democrats. In a second state, 47% of the voters are Republicans, and 53% are Democrats. Suppose 100 voters are surveyed from each state. Assume the survey uses simple random sampling.

What is the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state?

(A) 0.04
(B) 0.05
(C) 0.24
(D) 0.71
(E) 0.76

Solution

The correct answer is C. For this analysis, let P_1 = the proportion of Republican voters in the first state, P_2 = the proportion of Republican voters in the second state, p_1 = the proportion of Republican voters in the sample from the first state, and p_2 = the proportion of Republican voters in the sample from the second state. The number of voters sampled from the first state (n_1) = 100, and the number of voters sampled from the second state (n_2) = 100.

The solution involves four steps.

- Make sure the samples from each population are big enough to model differences with a normal distribution. Because $n_1P_1 = 100 * 0.52 = 52$, $n_1(1 - P_1) = 100 * 0.48 = 48$, $n_2P_2 = 100 * 0.47 = 47$, and $n_2(1 - P_2) = 100 * 0.53 = 53$ are each greater than 10, the sample size is large enough.

- Find the mean of the difference in sample proportions: $E(p_1 - p_2) = P_1 - P_2 = 0.52 - 0.47 = 0.05$.

- Find the standard deviation of the difference.

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$
$$\sigma_d = \text{sqrt}\{ [(0.52)(0.48) / 100] + [(0.47)(0.53) / 100] \}$$
$$\sigma_d = \text{sqrt} (0.002496 + 0.002491) = \text{sqrt}(0.004987) = 0.0706$$

- Find the probability. This problem requires us to find the probability that p_1 is less than p_2. This is equivalent to finding the probability that $p_1 - p_2$ is less than zero. To find this probability, we need to transform the random variable ($p_1 - p_2$) into a z-score. That transformation appears below.

$$z_{p1 - p2} = (x - \mu_{p1 - p2}) / \sigma_d = = (0 - 0.05)/0.0706 = -0.7082$$

We find that the probability of a z-score being -0.7082 or less is 0.24.

Therefore, the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state is 0.24.