Session 6

Difference Between Means

Many statistical applications involve comparisons between two independent sample means.

Difference Between Means: Theory

Suppose we have two populations with means equal to μ_1 and μ_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 are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.
- The set of differences between sample means are normally distributed. This will be true if each population is normal or if the sample sizes are large. (Based on the central limit theorem, sample sizes of 40 are large enough).

Given these assumptions, we know the following.

- The expected value of the difference between all possible sample means is equal to the difference between population means. Thus, $E(x_1 - x_2) = \mu_d = \mu_1 - \mu_2$.
- The standard deviation of the difference between sample means (σ_d) is approximately equal to:

$$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

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,

$$\sigma^2_d = \sigma^2_{(x1 - x2)} = \sigma^2_{x1} + \sigma^2_{x2}$$

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

$$\sigma^2_{x1} = \sigma^2_1 / n_1 \qquad \text{And} \qquad \sigma^2_{x2} = \sigma^2_2 / n_2$$

Therefore,

$$\sigma_d^2 = \sigma_1^2 / n_1 + \sigma_2^2 / n_2 \qquad \text{And} \qquad \sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

Problem 1

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.

In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have *at most* three more days of absences than the female sample?

(A) 0.025
(B) 0.035
(C) 0.045
(D) 0.055
(E) None of the above

Solution

The correct answer is B. The solution involves three or four steps, depending on whether you work directly with raw scores or z-scores. The "raw score" solution appears below:

- Find the mean difference (male absences minus female absences) in the population.

$$\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$$

- Find the standard deviation of the difference.

$$\sigma_d = \text{sqrt}(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$
$$\sigma_d = \text{sqrt}(7^2/100 + 6^2/50) = \text{sqrt}(49/100 + 36/50) = \text{sqrt}(0.49 + .72) = \text{sqrt}(1.21) = 1.1$$

- Find the probability. This problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3. To find this probability: 3, is the normal random variable; 5, is the mean; and 1.1, is the standard deviation. We find that the probability of the mean difference (male absences minus female absences) being 3 or less is about 0.035.

Thus, the probability that the difference between samples will be no more than 3 days is 0.035.