Probability and Statistics

Hesty Dewi Maria Siagian

A lot of statistical work relies on picking random samples from a population. One basic thing we often do is find the average of numbers in a sample. When we pick a sample randomly, the average we get forms a certain pattern. What can we say about this pattern?

Example 1

Imagine you're collecting data on the heights of students at a school. You randomly choose groups of two students each and find the average height in each group. Let's call this average the "group mean." The heights of the students in each group might vary, so the group means will also vary. Let's explore some questions about this situation:

1. First, think about a single student's height. If 40% of students are 5 feet tall, 40% are 6 feet tall, and 20% are 7 feet tall, what's the average ($\mathrm{E}(X)$) height among all students? And how much do individual heights differ from this average? This is like finding out how much the heights typically vary ($\mathrm{SD}(X)$) from the average among all students.
   Answer: The average height ($\mathrm{E}(X)$) among all students is calculated as (5 * 0.4) + (6 * 0.4) + (7 * 0.2) = 5.8 feet. The variability from this average is measured by the standard deviation, which is approximately 0.748 feet ($\mathrm{SD}(X) = \sqrt{0.56}$). This means that individual heights deviate from the average height by about 0.748 feet on average.

2. Now, you select two students at random and calculate the average height for this small group. What does the distribution of these group averages represent? In other words, how do the average heights of these pairs of students tend to behave?
   Answer: The distribution of these group averages represents how the average height of pairs of students tends to vary across many random selections of two students. It shows the likelihood of getting different average heights from different pairs of students.

3. Lastly, you're interested in how the variability of these group averages changes. What does the standard deviation ($\mathrm{SD}(\bar{X}_2)$) of these group averages tell you about the way the average height of pairs of students can differ from the overall average height?
   Answer: The standard deviation of these group averages ($\mathrm{SD}(\bar{X}_2)$) measures the variability of the average height of pairs of students. Over many samples of two students, the average height of pairs deviates from the overall average height (5.8 feet) by about 0.53 feet on average. This indicates how much the group mean tends to differ from the overall mean across different pairs of students.

Numerous statistical challenges involve drawing samples from a larger population. The **population distribution** outlines how a variable's values are spread across all members of the population. The **population mean** (μ) stands for the average value of the variable across the entire population. The **population standard deviation** (σ) signifies the spread of individual values in the entire population.

A **(simple) random** sample of size n comprises a set of random variables $X_1, ..., X_n$ that are independent and identically distributed (i.i.d.).

- The concept of independence assumes that the selection of individuals for the sample is unrelated to each other (akin to drawing with or without replacement from a very large population).
- Identical distribution assumes that all individuals are drawn from the same population, ensuring all individual values come from the same population distribution.

The **sample mean** signifies the sum of values in the sample. Since the sample is selected at random, the sample mean ($\bar{X}_n$) becomes a random variable with its own distribution. This distribution elucidates how sample means fluctuate from one sample to another across multiple random samples of size n.

Across numerous random samples, sample means do not systematically over or under-estimate the population mean. (In other words, the sample mean ($\bar{X}_n$) serves as an unbiased estimator of the population mean μ.)

The variability in sample means is influenced by the variability of individual variable values; greater variability among individual values in the population leads to more variance among sample means. However, sample means exhibit less variation than individual variable values. Furthermore, the variability of sample means diminishes as sample size grows.

Over numerous random samples, larger sample sizes result in less fluctuation in sample means from one sample to another, compared to smaller sample sizes. The standard deviation of sample means decreases as sample size increases, but this reduction follows a "square root rule." For instance, if sample size quadruples, the standard deviation of the sample mean decreases by a factor of $\sqrt{4} = 2$.

Considering the population mean μ and standard deviation σ, the above equations determine the mean and standard deviation of the distribution of sample means across various sample sizes. However, what about the shape of this distribution of sample means from sample to sample?