The Expected Value of the Variance of the Sample Mean in Large Sample Theory: The Importance of the Independent and Identically Distributed Assumption

WAKABAYASHI Ren* and OH Kyogai**

* First author, Faculty of Engineering, Kagoshima University, Kagoshima 8900065, Japan; k7971049@kadai.jp.

** Co-first author and corresponding author, Graduate School of Humanities and Social Sciences, Kagoshima University, Kagoshima 8900065, Japan; kyogaiw@leh.kagoshima-u.ac.jp.

1. Introduction

In statistics, the Independent and Identically Distributed (IID, i.i.d., iid) assumption is an important assumption that a sample of size n consisting of random variables drawn from a population distribution are independent of each other and follow the same probability distribution. Based on large sample theory, this study focuses on the iid assumption, a fundamental assumption in statistics, and examines the role of the iid assumption in calculating the expected value of the variance of the sample mean.

Accurate estimation of the sample variance is essential in statistics. Therefore, estimating the variance of the sample mean is a necessary step. The iid assumption plays a crucial role in determining the expected value of the variance of the sample mean, which is fundamental to understanding the properties of the sample variance.

Here, we explain the conditions for the iid assumption, which consists of two conditions: (1) sample independence and (2) identically distributed population. (1) The sample independence is guaranteed by the random sampling procedure. (2) The identically distributed population implies that each sample follows the same distribution as the population. If the population is identically distributed, then, the samples extracted will also be identically distributed regardless of the sampling method, and the expected value of the variance for each sample will be equal. Conversely, if the population is not identically distributed, then, the samples extracted regardless of the sampling method will not necessarily be identically distributed, and the expected variances may differ. Therefore, the assumption of identically distributed population is a crucial prerequisite for many statistical analyses.

Various sampling methods exist, but here we focus on Simple Random Sampling (SRS). SRS can be performed with or without replacement. Sampling with replacement yields independent samples, while sampling without replacement does not. Large sample theory assumes that the sample size is very large. Therefore, in large sample theory, both sampling with replacement and sampling without replacement satisfy the condition

of independence for SRS. Based on the above, this study aims to investigate how the iid assumption influences the expected value of the variance of the sample mean, particularly within the context of large sample theory and identically distributed population.

In large sample theory, the sample mean and its variance, derived from an iid sample, also satisfy the iid condition. The variance of the sample mean is used in the calculation of the sample variance. When the iid condition is satisfied, the derivation of the sample variance becomes significantly simpler. By examining the process of deriving the expected value of the variance of the sample mean in detail, it is possible to clearly see the impact of the iid assumption on the calculation of the expected value. This simplifies theoretical analysis and enhances practical utility.

The structure of this study is as follows. Section 2 outlines the population mean and the population variance. Section 3 discusses the properties of the sample mean and its variance. The purpose of this study is to examine the importance of the iid assumption under large sample theory. To this end, we employed the following approach to clarify the specific role of the iid assumption in the process of deriving the expected value of the variance of the sample mean. The two conditions of the iid assumption, (1) sample independence and (2) identically distributed population, were examined by developing them in the respective derivation processes in the order of (1) and (2), and then in the order of (2) and (1). Finally, we summarize the findings and present our conclusions.

2. Population

In statistics, the data set that constitutes the original distribution is called the population, and the data set that is actually observed is called the sample. The sample is a subset of the population. As mentioned earlier, the original distribution refers to (2) the identically distributed population. Because the population is identically distributed, each sample follows the same distribution as the population. The size of the population is generally denoted by N, and the size of the sample is denoted by n. The mean of the population distribution is called the population mean, generally denoted by μ, and is defined as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(A).}$$

The variance of a population is called the population variance and is generally expressed as σ^2. The population variance is defined by the following equation:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \text{(B).}$$

3. Sample mean $\bar{X}$ and its variance $V(\bar{X})$

Given a set of n independent samples from the population, the sample mean is defined as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{(C).}$$

From the definition, $\bar{X}$ is the arithmetic mean of the theoretical observations, and since each theoretical observation is a random variable, respectively, $\bar{X}$ is also a random variable. The expected value of the sample mean is derived as follows:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu \qquad \text{(D).}$$

For the third identity $[E(X_i) = \mu]$ in equation (D), we would like to emphasize that it is due to the definition of the population mean (A) and relates to the population, not the sample. The sample mean $\bar{X}$ is computed based on the sample data drawn from the population. However, the expected value of the sample mean, $E(\bar{X})$, represents what the average of those sample means would be if an infinite number of samples were repeatedly drawn from the population.

Now let us find the expected value of the variance of the sample mean, $V(\bar{X})$. First, let's start with the definition of variance,

$$V(\bar{X}) = E[\{\bar{X} - E(\bar{X})\}^2] = E[(\bar{X} - \mu)^2]$$

is obtained. Then, as defined in Equation (C),

$$V(\bar{X}) = E\left[\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right\}^2\right] = \frac{1}{n^2} E\left[\left\{\sum_{i=1}^{n} (X_i - \mu)\right\}^2\right]$$

is obtained. Expanding further, we obtain the following equation:

$$V(\bar{X}) = \frac{1}{n^2} E\left\{\sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i \neq j}^{n} (X_i - \mu)(X_j - \mu)\right\}$$

$$= \frac{1}{n^2}\left\{\sum_{i=1}^{n} E[(X_i-\mu)^2] + \sum_{i\neq j}^{n} E\big[(X_i-\mu)(X_j-\mu)\big]\right\}$$

$$= \tfrac{1}{n^2}\left\{\textstyle\sum_{i=1}^{n} V(X_i) + \sum_{i\neq j}^{n} Cov\big(X_i, X_j\big)\right\} \qquad \text{(E).}$$

It is crucial to emphasize that the iid condition is not yet used in the derivation of Equation (E). Equation (E) merely expresses the definitions of variance and covariance in different forms.

Here, we develop equation (E) for the iid conditions: (1) sample independence and (2) identically distributed population, first in the order of (1) and (2), and then in the order of (2) and (1). The iid condition consists of two conditions: (1) sample independence and (2) identically distributed population. These two conditions can be satisfied independently. Therefore, both cases can occur: (1) where the condition of identically distributed population is satisfied but the condition of sample independence is not, and (2) where the condition of sample independence is satisfied but the condition of identically distributed population is not. First, applying the condition (1) sample independence, to equation (E),

$$V(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

is obtained. This implies $\left[\sum_{i\neq j}^{n} Cov\big(X_i, X_j\big) = 0\right]$. Expanding $\sum_{i\neq j}^{n} Cov\big(X_i, X_j\big)$ yields

$$\textstyle\sum_{i\neq j}^{n} Cov\big(X_i, X_j\big) = \sum_{i\neq j}^{n} \rho_{i,j}\sigma_i\sigma_j = 0.$$

Here, it is important to note that $\big(\sigma_i \neq \sigma_j \neq 0\big)$ because condition (2), identically distributed population, is not yet applied. Therefore, the result $\left[\sum_{i\neq j}^{n} Cov\big(X_i, X_j\big) = 0\right]$ is solely due to the sample independence condition (1), which means $\big(\rho_{i,j} = 0\big)$. In summary, condition (1) sample independence implies $\big(\rho_{i,j} = 0\big)$. Not satisfying condition (2), identically distributed population, implies $\big(\sigma_i \neq \sigma_j \neq 0\big)$. Applying condition (2), identically distributed population, we have $\big(\sigma_i = \sigma_j\big)$. Then,

$$V(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}$$

which leads to the well-known result.

Conversely, if we apply the condition (2), identically distributed population, to equation (E),

$$V(\bar{X}) = \frac{1}{n^2}\{n\sigma^2 + n(n-1)\rho\sigma^2\} = \frac{\sigma^2}{n} + \frac{(n-1)}{n}\rho\sigma^2$$

is obtained. The assumption of identically distributed population means that all elements follow the same probability distribution. A consequence of identical distribution is that the expected values of the variance and correlation coefficients of the samples coincide with the variance and correlation coefficients of the population. Applying the condition of (1) sample independence, we have $(\rho_{i,j} = 0)$, which leads to the familiar result:

$$V(\bar{X}) = \frac{\sigma^2}{n}.$$

4. Conclusion

This study, grounded in large sample theory, underscores the importance of the fundamental statistical assumptions of (1) sample independence and (2) identically distributed population in deriving the expected value of the variance of the sample mean. The research systematically and critically investigated how the order in which these assumptions hold influences this expectation, clarifying their specific effects. Specifically, we calculated the expected value of the variance of the sample mean for both scenarios – (1) sample independence followed by identical distribution, and (2) identically distributed population followed by sample independence – and thoroughly analyzed the resulting differences. Our findings indicate that the order of these assumptions must be considered for accurately evaluating the expected variance of the sample mean. These results are expected to contribute to more accurate statistical inference and provide valuable guidance for data analysis across various fields.

References

[1] Hotta, Keisuke (2013), *Population, Sample and Sample Distribution,* PDF document. [https://www.bunkyo.ac.jp/~hotta/lab/courses/2013/2013dist/13dist_3.pdf]

[2] Taku Yamamoto (2004), *Econometrics* (New Economics Library 12), Shinsei-sha.

[3] Tokyo University, Statistics Section (1991), *Introduction to Statistics I: Fundamental Statistics,* University of Tokyo Press.

[4] Tokyo University, Statistics Section (1992), *Introduction to Statistics III: Statistics for the Natural Sciences,* University of Tokyo Press.

[5] Tokyo University, Statistics Section (1994), *Introduction to Statistics II: Statistics for the Humanities and Social Sciences,* University of Tokyo Press.