7.2: Sample Variance

In Section 6.2, we introduced the sample mean $\bar{X}$ as a tool for understanding the mean of a population. In this section, we formalize this idea and extend it to define the *sample variance*, a tool for understanding the *variance* of a population.

Estimating μ and σ^2

Up to now, μ denoted the mean or expected value of a random variable. In other words, it represented a parameter of a probability distribution. In the context of statistics, the main focus is more generally a population of objects, where the objects could be actual individuals and we are interested in a certain characteristic of the individuals, e.g., height or IQ. Oftentimes, we can model the distribution of values in a population using a certain probability distribution, and so it makes sense that the mean of a population is also denoted with μ. However, we may not always have a specific probability distribution in mind when considering the values of a population. Thus, we can generalize the interpretation of μ as the mean of a population as provided in the following definition. Note that the definition also provides a more general interpretation of σ^2 the variance of a population.

Definition 7.2.1

Suppose that a population has N elements, denoted $x_1, x_2, \ldots, x_N$. Then the ***population mean*** μ is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{7.2.1}$$

and the ***population variance*** σ^2 is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{7.2.2}$$

As we saw in Section 6.2, we can collect a random sample from a population and use the sample mean to estimate the population mean. More formally, let $X_1, \ldots, X_n$ be a collection of independent random variables representing a random sample of observations drawn from a population of interest. Then the sample mean, given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{7.2.3}$$

can be used to estimate the value of the population mean μ.

Note the use of lower case letters "x_i" in Definition 7.2.1 for the elements in the population. This is in contrast to the upper case letters "X_i" used to denote the elements of the random sample. Because the values in a population are fixed, though unknown in practice, it would not be appropriate to represent them with capital letters which are reserved for random variables per convention.

We argue that the sample mean $\bar{X}$ is the "obvious" estimate of the population mean μ because the population elements in Equation 7.2.1 are simply replaced by the corresponding sample elements in Equation 7.2.3. In addition to being the natural choice for estimating μ, $\bar{X}$ has another desirable property, which has to do with the following result, stated in Corollary 6.2.1 for normally distributed populations.

Theorem 7.2.1

For a random sample of size n from a population with mean μ and variance σ^2, it follows that

$$\mathrm{E}[\bar{X}] = \mu,$$
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof

Let $X_1, \ldots, X_n$ denote the elements of the random sample. Then $X_1, \ldots, X_n$ are independent random variables each having the same distribution as the population. In other words, we know that $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, for

$i = 1, \ldots, n$. Given this, and using the linearity of expected value and the independence of the sample elements, we have the following:

$$\begin{aligned} \mathrm{E}[\bar{X}] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu \\ \mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \end{aligned}$$

Theorem 7.2.1 provides formulas for the expected value and variance of the sample mean, and we see that they both depend on the mean and variance of the population. The fact that the expected value of the sample mean is exactly equal to the population mean indicates that the sample mean is an ***unbiased*** estimator of the population mean. This is because on average, we expect the value of $\bar{X}$ to equal the value of μ, which is precisely the value it is being used to estimate. This is a very desirable property for estimators to have as it lends more confidence to using their values in understanding the unknown population characteristic. We will keep the goal of using unbiased estimators as we now consider estimating the population variance.

Before we tackle the problem of estimating population variance, we again point out that the variance of the sample mean depends on the population variance. Thus, if we are interested in using the variance of $\bar{X}$ to quantify its accuracy in estimating the population mean, we need to know the value of σ^2, which is unlikely. (We talked about this at the end of Section 6.2 in the context of computing error probabilities. See Example 6.2.5.) So we have a specific need for an estimate of σ^2, not just for understanding the distribution of the population better.

Given Equation 7.2.2 in Definition 7.2.1, an "obvious" estimate of σ^2 is given by simply replacing the population elements by the corresponding sample elements, as we did for estimating μ. This gives the following formula for $\hat{\sigma}^2$ (note the "hat" ^), which is our first attempt at estimating σ^2:

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

The problem with this "obvious" estimate is that it is not unbiased. The following theorem (stated without proof) gives the expected value of $\hat{\sigma}^2$.

Theorem **7.2.2**

For a random sample of size n from a population with mean μ and variance σ^2, it follows that

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \sigma^2\left(\frac{n-1}{n}\right).$$

As we can see in Theorem 7.2.2, the expected value of $\hat{\sigma}^2$ does not equal σ^2, so it is not an unbiased estimator. Furthermore, note that $\hat{\sigma}^2$ actually ***underestimates*** the value of σ^2, on average, since its expected value is multiplied by a factor less than 1: $(n-1)/n < 1$. This is not good. Putting this again in the context of using the variance of the sample mean to quantify its accuracy in estimating the population mean, if we use $\hat{\sigma}^2$ to estimate σ^2, we would consistently report *higher* accuracy than what is actually being obtained, since smaller variance means less spread or greater confidence in our estimate of μ. This would make our analysis unreliable and misleading.

We can find an *unbiased* estimate of σ^2 by modifying our first attempt in $\hat{\sigma}^2$. The modification is to simply multiply by the reciprocal of the factor on σ^2 in the expected value of $\hat{\sigma}^2$. In doing this, we note that expected value of the modification will equal σ^2, following from the linearity of expected value:

$$\mathrm{E}\left[\left(\frac{n}{n-1}\right)\hat{\sigma}^2\right] = \left(\frac{n}{n-1}\right)\mathrm{E}\left[\hat{\sigma}^2\right] = \left(\frac{n}{n-1}\right)\sigma^2\left(\frac{n-1}{n}\right) = \sigma^2$$

We can simplify the modification of $\hat{\sigma}^2$ algebraically as follows:

$$\left(\frac{n}{n-1}\right)\hat{\sigma}^2 = \left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

This leads to the following definition of the ***sample variance***, denoted S^2, our unbiased estimator of the population variance:

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

The next theorem provides a sampling distribution for the sample variance *in the case that the population is normally distributed*.

Theorem 7.2.3

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then, it follows that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Theorem 7.2.3 states that the distribution of the sample variance, when sampling from a normally distributed population, is chi-squared with $(n-1)$ degrees of freedom. Note that without knowing that the population is normally distributed, we are not able to say anything about the distribution of the sample variance, not even approximately. There is no "CLT-like" result for the sample variance. Further note that it is not the distribution of S^2 alone, but rather, we multiply by one less than the sample size and divide by the population variance to get the result. This may not seem like a very useful result, given that it is the distribution of a quantity involving both the estimator S^2 and the parameter it is estimating σ^2. But you will see in the study of statistics how this can be utilized to quantify the error in using S^2 to estimate σ^2. But we can use Theorem 2.7.3 to help us in the context of understanding the accuracy of our estimate given by the sample mean. Before we state the result, we need two additional properties regarding the probabilistic qualities of $\bar{X}$ and S^2 as random variables. We state these properties without proof.

Theorem 7.2.4

1. $\bar{X}$ is independent of the collection of random variables given by $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$.
2. $\bar{X}$ and S^2 are independent.

Note that the second property in Theorem 7.2.4 follows immediately from the first one given our definition of S^2. Using these properties, we can prove the following result.

Theorem 7.2.5

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then, it follows that

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1} \tag{7.2.4}$$

Proof

We rewrite the quotient by dividing top and bottom by the quantity $\sqrt{\sigma^2/n}$:

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} = \frac{\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right)}{\sqrt{\frac{S^2/n}{\sigma^2/n}}} = \frac{\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right)}{\sqrt{\frac{S^2}{\sigma^2}}} \tag{7.2.5}$$

Note that the quantity in the numerator is the standardization of a normally distributed random variable, thus it has the standard normal distribution. For the denominator, we can further modify the expression under the square root by multiplying top and bottom by the quantity $(n-1)$:

$$\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}} = \sqrt{\left(\frac{(n-1)S^2}{\sigma^2}\right)\frac{1}{n-1}}$$

We know from Theorem 7.2.3 that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and so the denominator in Equation 7.2.5 is the square root of a chi-squared distributed random variable divided by its degrees of freedom. Also note from Theorem 7.2.4 that the numerator and denominator in Equation 7.2.4 are independent random variables, since they are functions of $\bar{X}$ and S^2, respectively. Thus, we have shown that the quantity we started with in Equation 7.2.4 is equal to a random variable with a standard normal distribution divided by the square root of an independent random variable with a chi-squared distribution divided by its degrees of freedom. This is precisely the definition of the t distribution given in Definition 7.1.3.

Notice what the result of Theorem 7.2.5 says: when sampling from a normally distributed population, if we take the sample mean and subtract its expected value μ and divide by its standard deviation *where the population variance σ^2 is estimated by the sample variance S^2*, then the resulting random variable has a t distribution with $(n-1)$ degrees of freedom. The distribution is no longer the standard normal distribution because we have now estimated the population variance, which has the effect of increasing the overall variability in the quantity given in Equation 7.2.4. To account for that increased variability, we need a distribution with *thicker tails*, which is precisely what the t distribution provides. Notice also that the degrees of freedom of the t distribution that models the quantity in Equation 7.2.4 is one less than the sample size because we lose a degree of freedom by using the sample variance to estimate the population variance. This result provides the foundation for many statistical inference techniques.