Sample Variance

Wenhao Jiang

2022-09-27

Sample Mean and Variance

As we mentioned in the class, sample mean or variance, as other sample *estimators*, is an estimate of the population mean or variance. In other words, what we care about is not the sample itself, but what we can infer from the sample about the population characteristics.

In Statistics, sample *estimators* generally need to have certain properties, one of which is *unbiasedness*. A sample estimator (*e.g.*, sample mean) is defined to be *unbiased* if and only if its *expectation* equals to the true population parameter (*e.g.*, population mean) to be estimated. In an informal language, when we have a large-to-infinite number of samples each randomly drawn from the population, the mean of the estimators from these samples should be equal to the true parameter in the population. Note that we will briefly come across *unbiasedness* in Chapter 5.

For example, we learned in class that the *unbiased* sample estimator of the population mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We define sample mean in this way because when we draw large-to-infinite number of random samples each containing n observations from the population, the mean value of these sample means (*i.e.*, the *expectation* of the sample mean) would be equal to the true population mean.

By definition, the true population mean is the *expected* value of x_i, *i.e.*, $\mu = \mathbb{E}[x_i]$. We first want to prove the *unbiasedness* of sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ by showing that $\mathbb{E}[\bar{x}] = \mu$

Proof 1

$$
\begin{aligned}
\mathbb{E}[\bar{x}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\mathbb{E}[x_1 + x_2 + ... + x_n] \\
&= \frac{1}{n} \times n \times \mathbb{E}[x_i] \\
&= \mathbb{E}[x_i] \\
&= \mu
\end{aligned}
$$

Now denote the sample variance as s^2 and the population variance *to be estimated* as σ^2. We would expect the desirable sample variance to be *unbiased* relative to the population

variance, that is,

$$\mathbb{E}[s^2] = \sigma^2$$

Suppose we define $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, with n instead of $n-1$ as the denominator.

Proof 2

$$\mathbb{E}[s^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i^2] - \mathbb{E}[\bar{x}^2]$$

We will cover probably later and you may try proving it by yourself by the definition of variance we learned in class that

$$var(x_i) = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2$$
$$and\ var(\bar{x}) = \mathbb{E}[\bar{x}^2] - \mathbb{E}[\bar{x}]^2$$

Therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i^2] - \mathbb{E}[\bar{x}^2] = \frac{1}{n}\sum_{i=1}^{n}\left(var(x_i) + \mathbb{E}[x_i]^2\right) - var(\bar{x}) - \mathbb{E}[\bar{x}]^2$$

$$= \frac{1}{n} \times n \times \sigma^2 + \frac{1}{n} \times n \times \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \frac{n-1}{n}\sigma^2$$

$$\neq \sigma^2$$

And it will be straightforward to prove that only when sample variance is defined by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ will the desired property $\mathbb{E}[s^2] = \sigma^2$ be satisfied.

Note: You may be curious why $var(\bar{x}) = \frac{\sigma^2}{n}$ in the proof above. We will explain it later when we cover Central Limit Theorem. The way to prove it is:

Proof 3

$$var(\bar{x}) = var\left(\frac{1}{n}(x_1 + x_2 + ... + x_n)\right)$$
$$= \left(\frac{1}{n}\right)^2 (var(x_1) + var(x_2) + ... + var(x_n)) \because x_i \sim i.i.d.$$
$$= \left(\frac{1}{n}\right)^2 \times n \times \sigma^2$$
$$= \frac{\sigma^2}{n}$$