8.2.2 Point Estimators for Mean and Variance

The above discussion suggests the sample mean, $\overline{X}$, is often a reasonable point estimator for the mean. Now, suppose that we would like to estimate the variance of a distribution σ^2. Assuming $0 < \sigma^2 < \infty$, by definition

$$\sigma^2 = E[(X - \mu)^2].$$

Thus, the variance itself is the mean of the random variable $Y = (X - \mu)^2$. This suggests the following estimator for the variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \mu)^2.$$

By linearity of expectation, $\hat{\sigma}^2$ is an unbiased estimator of σ^2. Also, by the weak law of large numbers, $\hat{\sigma}^2$ is also a consistent estimator of σ^2. However, in practice we often do not know the value of μ. Thus, we may replace μ by our estimate of the μ, the sample mean, to obtain the following estimator for σ^2:

$$\overline{S}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X})^2.$$

Using a little algebra, you can show that

$$\overline{S}^2 = \frac{1}{n} \left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2 \right).$$

Example 8.5

Let X_1, X_2, X_3, ..., X_n be a random sample with mean $EX_i = \mu$, and variance $\mathrm{Var}(X_i) = \sigma^2$. Suppose that we use

$$\overline{S}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X})^2 = \frac{1}{n} \left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2 \right)$$

to estimate σ^2. Find the bias of this estimator

$$B(\overline{S}^2) = E[\overline{S}^2] - \sigma^2.$$

Solution

First note that

$$\begin{aligned} E\overline{X}^2 &= (E\overline{X})^2 + \text{Var}(\overline{X}) \\ &= \mu^2 + \frac{\sigma^2}{n}. \end{aligned}$$

Thus,

$$\begin{aligned} E[\overline{S}^2] &= \frac{1}{n}\left(\sum_{k=1}^{n} EX_k^2 - nE\overline{X}^2\right) \\ &= \frac{1}{n}\left(n(\mu^2+\sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

Therefore,

$$\begin{aligned} B(\overline{S}^2) &= E[\overline{S}^2] - \sigma^2 \\ &= -\frac{\sigma^2}{n}. \end{aligned}$$

We conclude that $\overline{S}^2$ is a biased estimator of the variance. Nevertheless, note that if n is relatively large, the bias is very small. Since $E[\overline{S}^2] = \frac{n-1}{n}\sigma^2$, we can obtain an unbiased estimator of σ^2 by multiplying $\overline{S}^2$ by $\frac{n}{n-1}$. Thus, we define

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right).$$

By the above discussion, S^2 is an unbiased estimator of the variance. We call it the **sample variance**. We should note that if n is large, the difference between S^2 and $\overline{S}^2$ is very small. We also define the **sample standard deviation** as

$$S = \sqrt{S^2}.$$

Although the sample standard deviation is usually used as an estimator for the standard deviation, it is a biased estimator. To see this, note that S is random, so $\text{Var}(S) > 0$. Thus,

$$\begin{aligned} 0 < \mathrm{Var}(S) &= ES^2 - (ES)^2 \\ &= \sigma^2 - (ES)^2. \end{aligned}$$

Therefore, $ES < \sigma$, which means that S is a biased estimator of σ.

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample with mean $EX_i = \mu < \infty$, and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. The **sample variance** of this random sample is defined as

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2 = \frac{1}{n-1} \left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2 \right).$$

The sample variance is an unbiased estimator of σ^2. The **sample standard deviation** is defined as

$$S = \sqrt{S^2},$$

and is commonly used as an estimator for σ. Nevertheless, S is a biased estimator of σ.

You can use the `mean` command in MATLAB to compute the sample mean for a given sample. More specifically, for a given vector $x = [x_1, x_2, \cdots, x_n]$, `mean(x)` returns the sample average

$$\frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Also, the functions `var` and `std` can be used to compute the sample variance and the sample standard deviation respectively.

Example 8.6

Let T be the time that is needed for a specific task in a factory to be completed. In order to estimate the mean and variance of T, we observe a random sample T_1,T_2,$\cdots$, T_6. Thus, T_i's are i.i.d. and have the same distribution as T. We obtain the following values (in minutes):

$$18, 21, 17, 16, 24, 20.$$

Find the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

Solution

The sample mean is

$$\begin{aligned}\overline{T} &= \frac{T_1 + T_2 + T_3 + T_4 + T_5 + T_6}{6} \\ &= \frac{18 + 21 + 17 + 16 + 24 + 20}{6} \\ &= 19.33\end{aligned}$$

The sample variance is given by

$$S^2 = \frac{1}{6-1}\sum_{k=1}^{6}(T_k - 19.333)^2 = 8.67$$

Finally, the sample standard deviation is given by

$$S = \sqrt{S^2} = 2.94$$

You can use the following MATLAB code to compute the above values:

```
t = [18, 21, 17, 16, 24, 20];
m = mean(t);
v = var(t);
s = std(t);
```
