

### Mean and variance estimation

Consider a sample  $x_1, \dots, x_N$  from a random variable  $X$ . The sample may have been obtained through  $N$  independent but statistically identical experiments. From this sample, we want to estimate the mean  $\mu$  and variance  $\sigma^2$  of the random variable  $X$  (i.e., we want to estimate “population” quantities).

The *sample mean*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

is an estimate of the mean  $\mu$ . The expectation value of the sample mean is the population mean,  $E(\bar{x}) = \mu$ , and the variance of the sample mean is  $\text{var}(\bar{x}) = \sigma^2/N$ . Since the expectation value of the sample mean is the population mean, the sample mean is said to be an *unbiased estimator* of the population mean. And since the variance of the sample mean approaches zero as the sample size increases (i.e., fluctuations of the sample mean about the population mean decay to zero with increasing sample size), the sample mean is said to be a *consistent estimator* of the population mean.

These properties of the sample mean are a consequence of the fact that if  $x_1, \dots, x_N$  are mutually uncorrelated random variables with variances  $\sigma_1^2, \dots, \sigma_N^2$ , the variance of their sum  $z = x_1 + \dots + x_N$  is

$$\sigma_z^2 = \sigma_1^2 + \dots + \sigma_N^2. \quad (1)$$

If we view the members of the sample  $x_1, \dots, x_N$  as realizations of identically distributed random variables with mean  $E(x_i) = \mu$  and variance  $\text{var}(x_i) = \sigma^2$ , it follows by the linearity of the expectation value operation that the expectation value of the sample mean is the population mean:  $E(\bar{x}) = N^{-1} \sum_i E(x_i) = \mu$ . The variance of the sample mean follows from (1):  $\text{var}(\bar{x}) = N^{-2} \sum_i \sigma_i^2 = \sigma^2/N$ .

Moreover, the Central Limit Theorem states that, under fairly general conditions, the distribution of the sample mean  $\bar{x}$  approaches a normal distribution  $\mathcal{N}(\mu, \sigma^2/N)$  with mean  $\mu$  and variance  $\sigma^2/N$  as the sample size  $N$  increases [see, e.g., Johnson and Wichern (2002, chapter 4.5) or Papoulis (1991, chapter 8)].

The *sample variance*

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

is an unbiased estimator of the population variance  $\sigma^2$ , that is,  $E(s^2) = \sigma^2$ . To see

this, we calculate

$$\begin{aligned} E[(x_i - \mu)(\bar{x} - \mu)] &= \frac{1}{N} \sum_{j=1}^N E[(x_i - \mu)(x_j - \mu)] \\ &= \frac{1}{N} E(x_i - \mu)^2 \\ &= \frac{\sigma^2}{N}, \end{aligned}$$

where we have used the assumption that the  $x_i$  are mutually uncorrelated. With  $\text{var}(\bar{x}) = E[(\bar{x} - \mu)^2] = \sigma^2/N$ , it then follows that

$$\begin{aligned} E[(x_i - \bar{x})^2] &= E\left\{[(x_i - \mu) - (\bar{x} - \mu)]^2\right\} \\ &= \sigma^2 + \frac{\sigma^2}{N} - 2\frac{\sigma^2}{N} \\ &= \frac{N-1}{N}\sigma^2. \end{aligned}$$

Thus,

$$E(s^2) = \frac{1}{N-1} \sum_{i=1}^N E[(x_i - \bar{x})^2] = \sigma^2.$$

The denominator  $N - 1$  in the sample variance is necessary to ensure unbiasedness of the variance estimator. The denominator  $N$  would only be correct if fluctuations about the population mean  $\mu$  and not about the sample mean  $\bar{x}$  would appear in the expression for the sample variance. With the denominator  $N - 1$ , one obtains an indefinite sample variance for a sample of size  $N = 1$ , as expected. With the denominator  $N$ , the sample variance would vanish, yielding an obviously incorrect estimate of the population variance. The denominator  $N - 1$  appears because, after estimation of the sample mean, only  $N - 1$  degrees of freedom are available for the estimation of the variance, since the variables  $x_1, \dots, x_N$  and the sample mean satisfy the constraint

$$\sum_{i=1}^N (x_i - \bar{x}) = 0.$$

## References

- Johnson, R. A., and D. W. Wichern, 2002: *Applied Multivariate Statistical Analysis*. 5th ed., Prentice-Hall, 767 pp.
- Papoulis, A., 1991: *Probability, Random Variables, and Stochastic Processes*. 3rd ed., McGraw Hill, 666 pp.