The method of moments

THEOREM [Khinchin, 1929] Let $X_1, X_2, \dots$ be a sequence of mutually independent random variables with a common distribution. Let the expectation $\mathbb{E}[X_i] = \mu$ be finite and let $\overline{X}_n = (X_1 + \ldots + X_n)/n$ be the partial means. Then for every $\varepsilon > 0$,

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right)$$

approaches 0, as $n \to \infty$.

This form of the law of large numbers is stronger than Bernoulli's: it is not assumed that the variance is a finite quantity. Thus, under rather general conditions and for sufficiently large n, the sample mean $\overline{X}_n$ is close to the true mean of the distribution. Based on this is the moment method.

Let $X_1, \dots, X_n$ be a random sample from a probability distribution with density $f(x, \theta)$.

Here θ is an unknown parameter, or a tuple of unknown parameters, to be estimated.

Along with actual moments $\mu_k = \mathbb{E}[X_i^k]$, we have a sequence of *sample* moments $\mu_k^{(n)} = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$.

By Khinchin's theorem, $\mu_k^{(n)}$ approach μ_k in probability, as $n \to \infty$. The strategy of the moment method is to express θ in terms of moments μ_k of as low order as possible, and then to replace μ_k with $\mu_k^{(n)}$. This line of thought is classical, but the resulting estimates are not necessarily optimal.

EXAMPLE Let $X_1, \dots, X_{10}$ be a random sample from a probability distribution with finite mean μ and variance σ^2, both unknown. Then, using that $\mu_1 = \mathbb{E}[X_i] = \mu$ and $\mu_2 = \mathbb{E}[X_i^2] = \sigma^2 + \mu_1^2$, we obtain

$$\hat{\mu} = \mu_1^{(10)} = \frac{X_1 + \ldots + X_{10}}{10} \quad \text{and} \quad \hat{\sigma}^2 = \mu_2^{(10)} - (\mu_1^{(10)})^2 = \frac{(X_1 - \hat{\mu})^2 + \ldots + (X_{10} - \hat{\mu})^2}{10},$$

estimators for μ and σ^2, respectively.

EXAMPLE Let $X_1, \dots, X_n$ be a sample from a discrete probability distribution with mass function

$$f(k, \theta) = \begin{cases} 1/\theta, & k = 1, 2, \dots, \theta \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mu_1 = \mathbb{E}[X_i] = \dfrac{1 + 2 + \ldots + \theta}{\theta} = \dfrac{\theta + 1}{2}$, and so $\hat{\theta} = 2\mu_1^{(n)} - 1$ is an estimator for θ.

EXAMPLE Let $X_1, \dots, X_n$ be a sample from a uniform probability distribution on $[0, \theta]$.
Then $\mu_1 = \mathbb{E}[X_i] = \displaystyle\int_0^\theta x \, \frac{1}{\theta} \, dx = \frac{\theta}{2}$. Hence $\hat{\theta} = 2\mu_1^{(n)}$.

EXAMPLE Let $X_1, \dots, X_n$ be a sample from a uniform probability distribution on $[\theta_1, \theta_2]$.
Then $\mu_1 = \mathbb{E}[X_i] = \displaystyle\int_{\theta_1}^{\theta_2} x \, \frac{dx}{\theta_2 - \theta_1} = \frac{\theta_1 + \theta_2}{2}$ and $\mu_2 = \mathbb{E}[X_i^2] = \displaystyle\int_{\theta_1}^{\theta_2} x^2 \, \frac{dx}{\theta_2 - \theta_1} = \frac{\theta_1^2 + \theta_1\theta_2 + \theta_2^2}{3}$.
Hence $\theta_1 + \theta_2 = 2\mu_1$ and $\theta_1\theta_2 = 4\mu_1^2 - 3\mu_2$. Solving the quadratic equation, we obtain

$$\theta_1 = \mu_1 - \sigma_n\sqrt{3} \quad \text{and} \quad \theta_2 = \mu_1 + \sigma_n\sqrt{3}.$$

So

$$\hat{\theta}_1 = \mu_1^{(n)} - \hat{\sigma}_n\sqrt{3} \quad \text{and} \quad \hat{\theta}_2 = \mu_1^{(n)} + \hat{\sigma}_n\sqrt{3},$$

where $\hat{\sigma}_n^2 = \mu_2^{(n)} - (\mu_1^{(n)})^2$.