According to WatsonGS84, kriging was developed by Matheron, presumably starting with Krige's paper (report? thesis?) Krige51, and giving it an abstract foundation in probability.

It solves the problem of determining weights λ_p so that the estimate $\hat{f}(x) := \sum_{p \in P} \lambda_p f(p)$ for $f(x)$ is best in some sense. (This description does sound like a very special case of best approximation of linear functionals, as studied so extensively by GolombWeinberger59.) Best here means that $E(\widehat{f}(x) - f(x))^2$ be minimized. Defining

$$C(x, y) := E(f(x)f(y)),$$

this leads to

$$\lambda = (C(x_p, x_q) : p, q \in P)^{-1}(C(x_q, x) : q \in P).$$

(Mutatis mutandis, this is taken from WatsonGS84.) Come to think of it, this is quite general, with E only being required to be such that the inverse here exists. It surely depends on the function class f is thought to come from and just how the corresponding expectation is being defined. (In fact, I don't know what the precise meaning of $E(f(x)f(y))$ might be.) It would be useful to compare this with Golomb-Weinberger.

In the statistical approach, one wants the estimate to be unbiased and optimal, meaning that the expected value of $f(x) - \hat{f}(x)$ should be zero and its variation be minimal. Also, there is a mechanism for allowing for the data, $f(x_p)$, to be noisy. In particular, there is a kriging derivation involving uncorrelated, mean zero and σ^2 variance noise that uses $\hat{f}(x) := \sum_p \mu_p(f(x_p) + noise_p)$ and leads to

$$\mu = (\hat{C}(x_p, x_q) : p, q \in P)^{-1}(\hat{C}(x_q, x) : q \in P)$$

with

$$\hat{C} := \sigma^2 \mathrm{id} + C.$$

However, one can also make some common-sense demands, such as translation invariance:

$$\hat{f}(x + y) = \sum_{p \in P} \lambda_p f(p + y), \quad \forall y,$$

and see where that leads.

There is recent (1996) work by David Levin along exactly these lines.