Error estimation

taken from *Approaching nuclear interactions with lattice QCD*
Marc Illa, arXiv:2109.10068 [hep-lat]

The analysis of statistical uncertainties is one of the key steps in LQCD, since only a finite number of gauge-field configurations are produced to extract observables. Moreover, these configurations are not completely independent, showing correlations among them.

Assuming we have computed N samples of some variable x_i, with $n \in \{1, \dots, N\}$ (the usual starting point is the correlation function in LQCD), the sample mean, variance, and covariance with another variable y_n are given by

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n\,, \quad \sigma^2_{\overline{x}} = \frac{1}{N(N-1)}\sum_{n=1}^{N}(x_n - \overline{x})^2\,, \quad \mathcal{C}(x,y) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \overline{x})(y_n - \overline{y})\,. \tag{1}$$

Correlation functions are usually manipulated and inserted into functions, such as the effective mass. Then, when we look at some function f, the average value, labeled as $f(X)$, can be computed by using $f(\overline{x})$ (and not $\overline{f(x)} = \sum_n f(x_n)/N$, since it would be a biased estimator), and the uncertainty can be obtained via error propagation,

$$\sigma_{f(\overline{x})} = \sigma_{\overline{x}} \left|\frac{df}{dx}\right|_{x=\overline{x}}\,. \tag{2}$$

However, as the number of variables in f increases, and for correlated data, this formula is no longer practical, and alternative methods have to be used. In our case, such methods are resampling methods, the jackknife and bootstrap methods, both widely used in the LQCD community.

Jackknife method

The starting point is to construct the jackknife samples x^J_n, which are defined by taking the average of the variable x without including the nth sample,

$$x^J_n = \frac{1}{N-1}\sum_{m\neq n} x_m = \frac{N}{N-1}\overline{x} - \frac{1}{N-1}x_n\,. \tag{3}$$

Then, the mean value of the function f and the estimation of its uncertainty is

$$f(X) \simeq \overline{f}^J_x = \frac{1}{N}\sum_{n=1}^{N} f(x^J_n)\,, \quad \sigma^2_{f(X)} \simeq \frac{N-1}{N}\sum_{n=1}^{N}\left[f(x^J_n) - \overline{f}^J_x\right]^2\,. \tag{4}$$

An easy check that we can perform is to see that if $f(x) = x$, we recover the expressions from Eq. (1),

$$\begin{aligned}
\overline{x} &= \frac{1}{N}\sum_{n=1}^{N} x^J_n = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{N}{N-1}\overline{x} - \frac{1}{N-1}x_n\right) = \frac{N}{N-1}\overline{x} - \frac{1}{N-1}\overline{x} = \overline{x}\,,\\
\sigma^2_{\overline{x}} &= \frac{N-1}{N}\sum_{n=1}^{N}\left(x^J_n - \overline{x}\right)^2 = \frac{N-1}{N}\sum_{n=1}^{N}\left(\frac{N}{N-1}\overline{x} - \frac{1}{N-1}x_n - \overline{x}\right)^2\\
&= \frac{1}{N(N-1)}\sum_{n=1}^{N}(x_n - \overline{x})^2 = \sigma^2_{\overline{x}}\,.
\end{aligned} \tag{5}$$

A better estimator for $f(X)$ that includes $1/N$ bias correction is given by

$$f(X) \simeq N f(\overline{x}) - (N-1)\overline{f}_x^J \,. \tag{6}$$

Also, this method can be used to compute the covariance matrix, needed when performing χ^2 minimizations for the fitting of the correlation function. It is given by

$$\mathcal{C}^J[f(x), f(y)] = \frac{N-1}{N} \sum_{n=1}^{N} \left[f(x_n^J) - \overline{f}_x^J \right] \left[f(y_n^J) - \overline{f}_y^J \right] . \tag{7}$$

A generalization of the jackknife method is based on not only subtracting one sample from the set, but k samples, ending with $N_k = N/k$ jackknife samples.

Bootstrap method

The bootstrap resampling method, compared to the jackknife one, creates $N_{\rm boot}$ bootstrap samples x_α^B, with $\alpha \in \{1, \ldots, N_{\rm boot}\}$, where each one is obtained from a random selection from the original sample N points (with repetitions allowed),

$$x_\alpha^B = \frac{1}{N} \sum_{n=1}^{N} x_{{\rm rand}(1,N)} = \frac{1}{N} \sum_{n=1}^{N} r_n^\alpha x_n \,, \tag{8}$$

where r_n^α is the number of times x_n appears in the αth bootstrap sample, with the constrain that $\sum_n r_n^\alpha = N$, which follows a binomial distribution,

$$P(r_n^\alpha) = \frac{N!}{r_n^\alpha!(N-r_n^\alpha)!} p^{r_n^\alpha} (1-p)^{N-r_n^\alpha} \,, \tag{9}$$

where $p = 1/N$ is the probability that x_n is chosen. With this method, we can also write how to compute the mean value of the function f and the estimation of its uncertainty,

$$f(X) \simeq \overline{f}_x^B = \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} f(x_\alpha^B) \,, \quad \sigma^2_{f(X)} \simeq \frac{N}{N-1} \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} \left[f(x_\alpha^B) - \overline{f}_x^B \right]^2 . \tag{10}$$

In the case of $f(x) = x$, we recover the known formulas. For example, for the mean value,

$$\overline{x} = \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} x_\alpha^B = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} r_n^\alpha \right) x_n = \frac{1}{N} \sum_{n=1}^{N} x_n = \overline{x} \,, \tag{11}$$

where we have assumed that $N_{\rm boot}$ is very large, so we can use the mean value of a binomial variable, which is Np (with $p = 1/N$, we get $Np = 1$).

A better estimator for $f(X)$, which reduces the bias from order $1/N$ to $1/N^2$, is

$$f(X) \simeq 2 f(\overline{x}) - \overline{f}_x^B \,. \tag{12}$$

The covariance matrix can also be estimated with bootstrap,

$$\mathcal{C}^B[f(x), f(y)] = \frac{N}{N-1} \frac{1}{N_{\rm boot}} \sum_{\alpha=1}^{N_{\rm boot}} \left[f(x_\alpha^B) - \overline{f}_x^B \right] \left[f(y_\alpha^B) - \overline{f}_y^B \right] . \tag{13}$$

What is interesting with the bootstrap resampling method is that we can look at the distribution function of $f(x)$ and compute the errors with the quantiles. Also, the confidence regions of the parameters in some non-linear fits can be extracted easily.

Reference

P. Young, *Jackknife and Bootstrap Resampling Methods in Statistical Analysis to Correct for Bias*, https://young.physics.ucsc.edu/jackboot.pdf