The Bootstrap and Jackknife Methods for Data Analysis

Rainer W. Schiel

Institut für Theoretische Physik, Universität Regensburg

December 21, 2011
A very simple example

- Zero-dimensional data (only one quantity measured)
- from an experiment
- or from a Monte-Carlo simulation
A very simple example

- Plot the Histogram:
A very simple example

Average: $\bar{x} = \frac{1}{N} \sum_i x_i$  \( (= 14.9043 \ldots ) \)

Standard deviation: $\sigma = \sqrt{\frac{1}{N-1} \sum_i (x_i - \bar{x})^2}$  \( (= 3.2298 \ldots ) \)

Standard deviation of the mean: $\sigma_{\text{mean}} = \frac{1}{\sqrt{N}} \sigma$  \( (= 0.1021 \ldots ) \)

Result: $x = \bar{x} \pm \sigma_{\text{mean}}$  \( (= 14.90 \pm 0.10 ) \)
Standard deviation of the mean: an Alternative?

- Repeat the whole experiment (i.e. taking 1000 data points) many (say 2500) times
- Compute the means of each experiment (\(\rightarrow\) 2500 means)

![Histogram with means distribution]

- Standard deviation of these means is \(\sigma_{\text{mean}}\) (compare to 14.90 ± 0.10)
- Method works, but is nonpractical
- \(\Rightarrow\) use re-sampling techniques
The Bootstrap

“A loop sewn at the top rear or sometimes on each side of a boot to facilitate pulling it on.”

It is said that Baron von Münchhausen pulled himself (and his horse) up by his own bootstraps.
The bootstrap re-sampling technique

- Do experiment once (here: 1000 measurements)
- Use this data to “repeat the experiment” over and over again (here: 2500 times)
- “Repeat the experiment” by randomly picking (with replacement) 1000 measurements
- Do this 2500 times
- Compute the mean of each sample
The bootstrap re-sampling technique

- The mean of the means is $\approx$ the mean of the original data set
- The standard deviation of the means of the samples is $\sigma_{\text{mean}}$
Advantages of the Bootstrap method

- No need to propagate errors . . . never again!
  - just compute the final quantity that you’re interested in on all samples
  - the errors are automatically propagated to the final result
- Handle data with skewed distributions
  - the data of the samples is just as skewed as the original data
- Perform more complicated data analysis
  - e.g. consider two noisy data sets \( f_i(x) \) and \( g_i(x) \)
  - you are interested in a fit to \( \frac{f(x)}{g(x)} \)
  - with traditional methods, this can be difficult (division by zero!)
  - no problem with the bootstrap method, though
Disadvantages of the Bootstrap method

- Computationally intensive
  - though usually not a problem with today’s computers
- You might have to write your own code
  - but you might have to do that for other data analysis methods, too
- A reasonably large original data sample is needed (> $O(100)$)
According to www.artofmanliness.com, the Jackknife is a pocket knife that “has a simple hinge at one end, and may have more than one blade.”
The single-elimination jackknife re-sampling technique

- Somewhat similar to the bootstrap re-sampling technique
- Do experiment once (here: 1000 measurements)
- Make 1000 samples of 999 measurements each – with the \( i \)th measurement eliminated in the \( i \)th sample
- Compute the mean (or the wanted quantity) of each sample
The single-elimination jackknife re-sampling technique

- The mean of the means is the mean of the original data set
- The standard deviation of the means is $\sigma_{\text{mean}}/\sqrt{N}$ (where $N$ is the number of measurements (here: 1000))
Bootstrap vs. Jackknife

- The bootstrap method handles skewed distributions better.
- The jackknife method is suitable for smaller original data samples.
Conclusions

- The bootstrap and jackknife methods are powerful tools for data analysis.
- They are very well suited to analyze lattice data.