M1 - Marine Physics 2019–2020

Data Analysis

#4 Confidence interval

1 Confidence interval for the mean

We will construct confidence intervals for the mean of a normally distributed random variable by different methods. We will first use the CLT over one sample, then the the bootstrapping method, and finally check that both methods are valid by comparing them to the results we get from really generating a large number of samples (ideal world).

1. generate a random variable `x`, normally distributed, with `n=50` elements

[Method 1 : CLT]

2. Estimate the standard error for the mean value of `x` by the standard method (using the Central Limit Theorem).
3. Estimate the 95% confidence interval for the mean value of `x` by the standard method (using the Central Limit Theorem). Remember that far small sample size ($n < 100$) one should use the t-distribution (see http ://stockage.univ-brest.fr/ gula/TS1/t-table.pdf).

[Method 2 : Bootstrapping]

4. Construct `m=1000` bootstrap samples from `x` and compute the `m` means (the variable `mu`).
5. Compute the mean, standard deviation and plot the pdf of `mu`. The standard deviation gives you the *bootstrapped standard error* of the mean of the random variable `x`.
6. Compute the 2.5th and 97.5th centiles of `mu`. It gives you the *bootstrapped 95% confidence interval* for the random variable `x`.

[Method 3 : The ideal world]

7. Generate an array $m \times n$ of normally distributed random variables [matlab : `y=randn(n,m);`, python : `y = np.random.randn(n,m)`]. Plot the pdf of the distribution of the `m` means, compute the standard error and the confidence interval at the 95% confidence level for the means.
8. Compare the results obtained by bootstrapping and by using the CLT over one sample to the values you obtained by really computing the means of m samples.

2 Bootstrapped CI for the median

We will now apply the bootstrapping method to estimate the median of a distribution (for which we cannot apply the CLT).

1. generate a random variable `x`, following a `lognormal distribution` [matlab : `lognrnd;`, python : `y = np.random.lognormal`] with mean=1 and standard deviation=1, with `n=20` elements
2. Compute the mean and median of `x`.
3. Estimate standard error and the confidence interval at the 95% confidence level for the mean and median value of `x` by bootstrapping using `m=1000` bootstrap samples.
4. Generate an array $m \times n$ of lognormally distributed random variables [matlab : `y=lognrnd(mu,sigma,n,m);`, python : `y = np.random.lognormal(mu,sigma,(n,m))`]. Plot the pdf of the distribution of the `m` means and medians, compute the standard error and the confidence interval at the 95% confidence level for the means and medians.