Statistical Methods
(summer term 2024)

Bootstrap Monte Carlo

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Quick and dirty confidence intervals: Bootstrap Monte Carlo I

- Sometimes the probability distribution function (PDF) is unknown, making it difficult to perform a reliable Monte Carlo simulation
- Perhaps all we have is a given sample with no information about the underlying PDF
- Is there still a way to derive confidence intervals for a given estimator using Monte Carlo simulation?
- The answer is yes! and the method is called Bootstrap Monte Carlo (or just Bootstrap for short)
- Key Idea: By repeatedly resampling the given data, we can create new datasets that mimic the original sample
- This approach allows us to estimate the variability of the estimator without knowing the true distribution

Quick and dirty confidence intervals: Bootstrap Monte Carlo II

- Assumption: We have a sample that consists of N independent and identically distributed (iid) data "points"
  - a "point" can be complex object (e.g., vector, matrix, ...)
  - iid means that the points are independent of each other and the sequential order does not matter

- Bootstrap procedure: Given the original dataset $D_{(0)}$
  - Generate new datasets $D_{(1)}$, $D_{(2)}$, ... of the same size N by drawing points from $D_{(0)}$ $with\ replacement$
  - Approximately $e^{-1} \approx 37\%$ of the original points will be omitted each time
  - for each dataset calculate the statistic of interest a_i

- Approximate theorem: The bootstrap statistics a_1, a_2, ... are distributed around a_0, similarly to how a_0 is distributed around the true value a_{true}
  - This implies that the variance among the $a_{1,2,3,...}$ provides an estimate of the variance of a_0

- Suggested number of artificial data sets: $\geq N \ln(N)^2$ (as a rule of thumb)

Quick and dirty confidence intervals: Bootstrap Monte Carlo III

- Example: If we are interested in the mean, we calculate the mean for each bootstrap sample and then assess the spread of these means to estimate the confidence interval

- Caveat 1: i The original data set D_0 must be a good representation of the underlying population for the bootstrap estimates to be meaningful. If the original sample is biased or not representative, the bootstrap samples will also be biased

- Caveat 2: The iid (independent and identically distributed) assumption should hold. This means that each data point in the original sample should be drawn independently from the same distribution. Violations of this assumption can lead to incorrect confidence intervals

Example: confidence intervals of a correlation coefficient

In previous exercises you determined the correlation coefficient among the results of students at Heidelberg University. Now, we address a different question: since we only had finite samples, how well can we determine the actual underlying correlation coefficient for the (hypothetical) world-wide population of students taking exams? The R command `cor.test()` provides an answer by giving confidence intervals of the sample correlation coefficient:

```
>cor.test(r[,1],r[,2])
...
95 percent confidence interval:
 0.5613007 0.6923753
sample estimates:
      cor
0.6313255
```

We have not covered the underlying theory yet. However, using the Monte Carlo approach, can you confirm the confidence interval by running a bootstrap Monte Carlo simulation? For simplicity, we'll use one sample, PEP1 vs PEP2. I've provided a modified file, `pep1_ws17-pep2_ss18nozeros.txt`, where zero results have been eliminated.

`cor()`, `quantile()`, `sample()` will be useful here → `bootstrap.ipynb`