Analytics for Observational Data (IT142IU)

Lab 6: Bootstrap and Jackknife methods

## Objectives

* Understanding bootstrap and Jackknife methods
* Doing sampling using these methods to measure statistical parameters.
* Dataset sources:
  + <https://www.kaggle.com/carrie1/ecommerce-data>
  + <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>
* Programming languages: Python/Java
* Ref: Lecture notes

## Tasks

***Part 1. Do bootstrapping.***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | ecommerce-data |
| Choose a random variable:  - The quantity of a product type (StockCode) bought in an invoice  - The total price of an invoice | The total price of an invoice: SumPrice |
| Describe a strategy of bootstrapping to measure the Mean and Std Deviation of the chosen random variable.  Hint: choose a country which has more than 1000 records in the dataset and generate 1000 samples of size 1000. |  |
| Present the distribution of sample means and std deviations using boxplot and bar-chart. Five-number summaries should be shown. |  |
| Estimate the standard errors of the measured parameters |  |
| Identify the confident intervals of measured parameters |  |
| Remark |  |

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | GlobalLandTemperaturesByCountry.csv |
| Choose a random variable | AverageTemperature |
| Describe a strategy of bootstrapping to measure the Mean and Std Deviation of the chosen random variable.  Hint: choose a country which has more than 1000 records in the dataset and generate 1000 samples of size 1000. |  |
| Present the distribution of sample means and std deviations using boxplot and bar-chart. Five-number summaries should be shown. |  |
| Estimate the standard errors of the measured parameters |  |
| Identify the confident intervals of measured parameters |  |
| Remark |  |

***Part 2. Apply Jackknife method***

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | ecommerce-data |
| Choose a random variable:  - The quantity of a product type (StockCode) bought in an invoice  - The total price of an invoice | Same data from bootstrapping: SumPrice |
| Describe the steps of Jackknife method to measure the Mean and Std Deviation of the chosen random variable.  Hint: choose a country which has more than 1000 records in the dataset and generate 1000 samples of size 1000-1. |  |
| Present the distribution of sample means and std deviations using boxplot and bar-chart. Five-number summaries should be shown. |  |
| Estimate the standard errors of the measured parameters |  |
| Identify the confident intervals of measured parameters |  |
| Remark |  |

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | GlobalLandTemperaturesByCountry.csv |
| Choose a random variable | AverageTemperature |
| Describe the steps of Jackknife method to measure the Mean and Std Deviation of the chosen random variable.  Hint: choose a country which has more than 1000 records in the dataset and generate 1000 samples of size 1000-1. | Sorry miss, the data is to big for my PC, and I get crashed all the time while trying to do resampling technique |
| Present the distribution of sample means and std deviations using boxplot and bar-chart. Five-number summaries should be shown. |  |
| Estimate the standard errors of the measured parameters |  |
| Identify the confident intervals of measured parameters |  |
| Remark |  |