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/datasets/fedesoriano/wind-speed-prediction-dataset>
  + <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 | Wind speed prediction |
| Choose a random variable:  - Wind  - Rain | Wind |
| Describe a strategy of bootstrapping to measure the Mean and Std Deviation of the chosen random variable.  Hint: Take a sample of 1000 records in the dataset and generate 1000 samples of size <=1000. The sample sizes are same. | To estimate the statistics of a sample, the bootstrap method creates n subsamples, replacing each one with n equal to the original sample size. The following is a description of the process:  1. Decide how many bootstrap samples to run; for this example, I run 1000 samples.  2. Select a sample size; I used all 1000 records.  3. For every sample of bootstraps:  - Sketch an example using the selected size substituted in.  Determine its standard deviation and mean.  - Determine the sample means and standard deviations aggregate mean. |
| 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 | - Mean: The variable {WIND} has an average value of 9.93945. As a gauge of our level of confidence in this average, the standard error of 0.00511 is used. The true average wind speed should fall between the confidence from 9.61997 to 10.259.  -Standard Deviation: The standard deviation of 5.1155 indicates the degree to which wind speeds deviate from the mean. How much we can infer from a standard deviation of 0.003687 can depend on this measure of variability. What we expect to be the genuine fluctuation of wind speeds falls inside the confidence interval of 4.896 to 5.344.  -In conclusion, these data indicate that the average wind speed is 9.79, with some variance around this value. The range that the confidence intervals provide indicates what the genuine average and variability should be. |

|  |  |
| --- | --- |
| **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 Result:  1000. | Choose Poland  To estimate the statistics of a sample, the bootstrap method creates n subsamples, replacing each one with n equal to the original sample size. The following is a description of the process:  1. Decide how many bootstrap samples to run; for this example, I run 1000 samples.  2. Select a sample size; I used all 1000 records.  3. For every sample of bootstraps:  - Sketch an example using the selected size substituted in.  Determine its standard deviation and mean.  - Determine the sample means and standard deviations aggregate mean. |
| 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 | - Mean: Poland experiences 7.514-degree temperatures on average. We can gauge our level of confidence in this average by looking at the 0.008006 standard error. The range that represents our expectation for the true average temperature is 7 to 8.  - Standard Deviation: The 8.189 value indicates the degree to which the temperatures deviate from the mean. The variability measure's standard error of 0.00355125 indicates the degree of reliability. The confidence interval spanning from 7.978 to 8.416 represents the range in which we anticipate the actual temperature variations to occur.  In conclusion, these data indicate that Poland experiences average temperatures of 7.46 degrees, with some fluctuation from this norm. The range that the confidence intervals provide indicates what the genuine average and variability should be. |

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

|  |  |
| --- | --- |
| **Questions** | **Answers** |
| Dataset | Wind speed prediction |
| Choose a random variable:  - Wind  - Rain | Choosing Wind |
| Describe the steps of Jackknife method to measure the Mean and Std Deviation of the chosen random variable.  Hint: Take a sample of 1000 records in the dataset and generate 1000 samples of size 1000-1. | Explain how to measure the mean and standard deviation of the selected random variable using the Jackknife method.  Advice: Create 1000 samples of size 1000-1 by taking a sample of 1000 records from the dataset. Choose 1000 records at random to create a dataset for sampling (S).  Resample 1: {S\ the initial entry}  Reference 2: {S\ the subsequent item}  Note: The number of samples may be less than 1000, for example, depending on how many data values were extracted from the dataset. Therefore, N resamples of size N-1 are possible to produce.  The leave-one-out technique is used in the Jackknife method to estimate sample statistics. The following is a description of the process:  1. Decide on the sample size you wish to use; in this example, I used 1000 observations.  2. I is the index of the i-th leave-one-out observation in the sample for each i (where I = 1, 2, 3,... n):  - Take off the i-th note.  Determine the standard deviation and mean.  3.The total mean of the computed sample means and standard deviations is the jackknife estimated mean. |
| Present the distribution of sample means and std deviations using boxplot and bar-chart. Five-number summaries should be shown in the box-plot. |  |
| Estimate the standard errors of the measured parameters |  |
| Identify the confident intervals of measured parameters |  |
| Remark | In comparison to the bootstrap approach, the jackknife method exhibits minimal variability in both the mean and standard deviation values. This shows that, in comparison to the bootstrap approach, the jackknife method yields more consistent estimates, although it may be less resistant to outliers or small sample sizes. However, there is strong agreement between the two approaches regarding the WIND variable's central tendency and spread, which gives confidence in the accuracy of these estimations. The particular environment and the analysis's needs, such as whether consistency or robustness is required, may influence which of the two approaches is best. |

|  |  |
| --- | --- |
| **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. | Explain how to measure the mean and standard deviation of the selected random variable using the Jackknife method.  Advice: Create 1000 samples of size 1000-1 by taking a sample of 1000 records from the dataset. Choose 1000 records at random to create a dataset for sampling (S).  Resample 1: {S\ the initial entry}  Reference 2: {S\ the subsequent item}  Note: The number of samples may be less than 1000, for example, depending on how many data values were extracted from the dataset. Therefore, N resamples of size N-1 are possible to produce.  The leave-one-out technique is used in the Jackknife method to estimate sample statistics. The following is a description of the process:  1. Decide on the sample size you wish to use; in this example, I used 1000 observations.  2. I is the index of the i-th leave-one-out observation in the sample for each i (where I = 1, 2, 3,... n):  - Take off the i-th note.  Determine the standard deviation and mean.  3.The total mean of the computed sample means and standard deviations is the jackknife estimated mean. |
| 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 | Both the Jackknife and the Bootstrap approaches yield comparable mean estimates when evaluating the AverageTemperature variable for the United States; the Jackknife mean is 7.74 and the Bootstrap mean is 7.51, demonstrating the robustness of the mean estimate. But the Jackknife approach shows far better accuracy and reduced variability, with a smaller range of estimates and a standard deviation of 4.98 for the means as opposed to 8.1896 for the Bootstrap. Because of this, the Jackknife approach is more dependable and consistent for applications that need stability and low variability, whereas the Bootstrap method is still a versatile and dependable choice for various statistical studies while being more unpredictable. |