IND522
Advanced Statistical Modelling

Fall 2025

Professor Orhan FEYZİOĞLU

Assignment ♯5: Due December 12

1) Generate 1000 random variables with 10 dimensions each using `randn`. Construct a screeplot and find the cumulative percentage of variance. Is there any evidence that one could reduce the dimensionality of these data? If so, what would be a value for k?

2) The **`bank`** data contains two matrices comprised of measurements bank made on genuine money and forged money. Combine these two matrices into one and use PPEDA (not PCA) to discover any clusters or groups in the data. Compare your results with the known groups in the data.

3) Intercity distances for 81 cities of Türkiye is included **`turkiye`** data. Apply the classical multidimensional scaling to reduce the data to two dimensions and construct a scatterplot.

4) Load the data sets in **`posse`**. Apply the PPEDA to each of these data and report your results.

5) Write MATLAB code that implements the parametric bootstrap. Test it using the **`forearm`** data. Assume that the normal distribution is a reasonable model for the data. Use your code to get a bootstrap estimate of the standard error and the bias of the coefficient of skewness and the coefficient of kurtosis. Get a bootstrap percentile interval for the sample central second moment using your parametric bootstrap approach.

6) Load the **`lawpop`** data set. These data contain the average scores on the LSAT (lsat) and the corresponding average undergraduate grade point average (gpa) for the 1973 freshman class at 82 law schools. *Note that these data constitute the entire population.* The data contained in **`law`** comprise a random sample of 15 of these classes. Obtain the true population variances for the lsat and the gpa. Use the sample in **`law`** to estimate the population variance using the sample central second moment. Get bootstrap estimates of the standard error and the bias in your estimate of the variance. Make some comparisons between the known population variance and the estimated variance.

7) Using the **`lawpop`** data, devise a test statistic to test for the significance of the correlation between the LSAT scores and the corresponding grade point averages. Get a random sample from the population, and use that sample to test your hypothesis. Do a Monte Carlo simulation of the Type I and Type II error of the test you devise.

8) The **`remiss`** data set contains the remission times for 42 leukemia patients. Some of the patients were treated with the drug called 6-mercaptopurine (**`mp`**), and the rest were part of the control group (**`control`**). Use the exploratory data analysis techniques to help determine a suitable model (e.g., Weibull, exponential, etc.) for each group. Devise a Monte Carlo hypothesis test to test for the equality of means between the two groups. Use the p-value approach.

9) The **`quakes`** data give the time in days between successive earthquakes. Use the bootstrap to get an appropriate confidence interval for the average time between earthquakes.