Nonparametric Statistics

Name, family name and ID number

2024-02-02 (14 Pluviôse CCXXXII)

Instructions

- For all computations based on permutation/bootstrapping, use $B = 1000$ replicates, and seed = 2024 every time a permutation/bootstrap procedure is run.
- For Full Conformal prediction intervals, use a regular grid, where, for each dimension, you have $N = 50$ equispaced points with lower bound $min(data) - 0.25 * range(data)$ and upper bound $max(data) + 0.25 * range(data)$. Moreover, do not exclude the test point when calculating the conformity measure. Except for the number of points, these are the default conditions of the `ConformalInference` R package.
- Both for confidence and prediction intervals, as well as tests, if not specified otherwise, set $\alpha = 0.1$.
- When reporting univariate confidence/prediction intervals, always provide upper and lower bounds.
- For solving the exam, you must use one of the templates previously provided and available here. Particularly, for each question you are required to report:
  - Synthetic description of assumptions, methods, and algorithms: which methodological procedure you intend to use to answer the question, succinctly describing the main theoretical characteristics of the chosen approach, and why it is suitable for the analytical task at hand. For *e.g.*, if you perform a statistical test, you should mention its assumptions, null and alternative hypotheses, and the test statistic.
  - Results and brief discussion: the actual result of the procedure applied to the data at hand, including any requested comment, output and plot.

Exercise 1

A historian friend of yours, Dr. Roberts, requires your statistical expertise to research French history. She has obtained historical data on the *assignats*[1], government bonds emitted by the Constituent Assembly in December 1789 that became official paper money during the French Revolution.

In particular, she has a data set consisting of discretised observations of time series, where the value of *assignats* at the last day of each month of year II of the French Revolutionary calendar (22 October 1792 to 22 October 1793) is gathered by different reporters. Every row represents the collected measurements on prices of the bonds by the (independent between them) reporters that published them on one of two journals. The value of the *assignats* is measured as a percentage relative to their value at the last day of the *Brumaire* month (the first month of the year under such calendar). For every reporter (row), the following information is available:

- `df1` data frame:

[1] At first, they were bonds that yielded annual 5% interest, whose collateral were the *Biens nationeaux*, that is the properties the government had confiscated from the Catholic Church. See https://en.wikipedia.org/wiki/Assignat

  - `Reporter.ID` the ID of the reporter.
  - `Journal` the journal in which the reporter's observations were published.
  - `Brumaire`, `Frimaire`, *et cetera*: the value of the *assignats* in the last day of such month. Every month is 30 days long in this calendar. Note that the names of the months are chronologically ordered. So the *Brumaire* column has the value of *assignats* the last day the first month, *Frimaire* in the last day of the second month, etc. In other words, they are the discretised observations of a continuous time measured in months.

- `df.tte` data frame:
  - `Journal`: same as before.
  - `time.to.fall`: how long it took the *assignats* to lose 5% of their value according to the data by the reporter in months; `NA` if they never did.

Both dataframes are included in the list object stored in the file `feb2024.rds`, and are available ON THIS LINK

Figure 1: A peasant selling his assignats in exchange for hard cash. Musée Carnavalet, Paris.

1. You firstly explore the data. Plot, for each `Journal`, the (discretised) curves of the reported values of the *assignats* as a function of time, from *Brumaire* to *Vendémmiaire*. Superpose the two sample Tukey depths (compute them treating the available observations as vectors in $\mathbb{R}^{12}$, setting the seed beforehand[2]) of each `Journal` to the plot.

2. To test the hypothesis that, according to *Le Père Duchesne*, there was an **increase** in the median value of *assignats* from the last day of *Nivôse* to the last day of *Vendémmiaire*, perform a sign test.

3. You want to know what the central value of *assignats* one year before the *coup d'état* against Robespierre was. Assuming the symmetry of the distribution and using reports only from *Le Vieux Cordelier*, provide a permutational confidence interval for the value of *assignats* at the last day of *Thermidor*. Set the spacing between grid points at 0.05.

[2]Of course, it would make much more sense to model the observations as functional data, but not everyone followed the 8-CFU version. Moreover, the multivariate depth measures will be very unstable in such high dimensionality.

4. Focus on the value of *assignats* at the last day of *Pluviôse*. You want to test if the mean of this population, according to *Le Vieux Cordelier*, was the same as the available observation from *Brumaire*. Perform a Bootstrap test building a CI of the pertinent statistic. Assume that the natural logarithm of the values of the *assignats* is normally distributed to enhance your Bootstrap algorithm.

Exercise 2

Dr. Roberts now wants to understand the rationale behind the decisions of Talleyrand, a cunning diplomat that made money investing in *assignats*. N.B. use `df.tte` for exercises (1) *et* (2), and `df1` for (3) *et* (4),

Figure 2: The Coup of 9 Thermidor: members of the Convention accuse Robespierre of being a tirant. He was later executed by the guillotine. Source: Wikipedia

1. Using the data in `df.tte`, for each `Journal` obtain the Kaplan-Meier estimator, providing a plot with both of them.
2. Perform a permutational version of the Log-Rank test for the equality of survival curves between both journals.[3]
3. Dr. Roberts knows the first time series in *Le Vieux Cordelier* was reported by Camille Desmoulins (`Reporter.ID`= 18), its founder himself. Using only such row, perform a nonparametric smoothing of the price as a function of time (months) with a gaussian kernel with *bandwidth* = 1.5 months for the historic value of assignats from the last day of *Brumaire* (included) on. Report its plot.
4. On December 22nd, 1793 (*Nivôse* month of the year for which the data are available), a 24-year-old Napoleon Bonaparte won the Siege of Toulon against the British. Provide a Full Conformal (coverage 90%) prediction region for the prediction of the developed model in (3) at **the middle** of the month of such victory for the value of *assignats*.

[3]Hint: you can extract the value of the test statistic with the `chisq` attribute of the `survdiff` function. Then apply the procedure you already know.

Exercise 3

Dr. Roberts wants to revolutionise the way the French Revolution is analysed. It is time for you to show off state-of-the-art statistical procedures. N.B. use `df1`.

1. Some of the reporters are suspected to have been *enragés* supporters (extremists) and their data to have been made up. Define a functional data object, where the abscissa is the time (months) from *Brumaire* on, and the ordinate the corresponding available value of *assignats*. Build a functional Outliergram, plot it and report the indices of the outliers (if there are any).

2. Let us explore the shape of the time series. Using the available observations from spring (*Germinal*, *Floréal* and *Prairial*), obtain the robust estimate of the covariance matrix of the *assignats*'s values with Minimum Covariance Determinant with a reweighting step for each `Journal`. Consider 1000 subsets for initialising the algorithms and set the proportion of the subsets over which the determinants are minimized equal to $\alpha = 0.75$. Report the (reweighted) MCD estimates of location and scatter. Provide the distance-distance plot for *Le Vieux Cordelier* and comment your result.

3. Using the same data as in (2) and the same procedure to estimate the robust covariance matrix and location, perform a permutation test for equality of robust locations between both journals. Use as test statistic the ℓ^∞-norm of the difference of the $\mathbb{R}^3$ vectors.

References

- Benelli, E.; Cantarella, E.; et al. (August 2019). *Rivoluzione e Impero: dalla presa della Bastiglia alla fine di Napoleone.National Geographic*: Speciale Storica, volume number 41.
- Hawtrey, R. G. (1918). *The Collapse of the French Assignats.* The Economic Journal, 28(111), 300–314. https://doi.org/10.2307/2222796
- Levasseur, E. (1894). The assignats: A study in the finances of the French Revolution. Journal of Political Economy, 2(2), 179-202.

Disclaimer

The data were simulated but made coherent with the reality. The questions were made trying not to be inaccurate with respect to the real facts.