**INTRODUCTION**

The project aims to investigate the phenomenon of Italy’s newborns plunge: in particular, it focuses on its evolution in space and time.

**PROBLEM PRESENTATION**

As we will see, newborns are significantly decreasing in Italy in the last 20 years, and this will be a problem in the future of our Nation. Understand why this is happening and having a complete overview of the problem is crucial in order to tackle it properly, as we must do in the next years.

Our government and more in general all the institutions should be aware of the complexity of the phenomenon, starting from analysis of this type.

We will try to explore the cited complexity, starting from a geographical point of view, studying all the provinces’ specificity, and we will try to explore even some covariates which may impact the fertility rates.

**WHY NONPARAMETRIC?**

We have decided to go for a nonparametric approach since the data we’re dealing with are purely functional for the aim we gave them.

Nonparametric smoothing procedures are proven to be better than the parametric ones (de beer), and even for the inference part we want to have the possibility to test for various things and aspect of the data, and even on the various direction we can consider, and so the nonparametric approach gives us this possibility of building the test as we desire.

**INTRO\_A**

The dimension of ages of women is dropped, considering uniquely years and provinces. Since the data regarding the total number of inhabitants are available up to 2019, in this analysis, years are spanned between 2002 and 2019.

The number of newborns is normalized with respect to the total number of inhabitants of the specific province and year.

*Figure 1: Plot intro A [1] - Each curve represents the normalized number of newborns of each province along years.*

The curves in Figure 1 suggest an overall decrease along years. In particular, the difference between the normalized total number of newborns in 2002 and 2019 is strictly positive for every province.

*Figure 2: Plot intro A [2] - Each point represents the delta between 2002 and 2019 for each province.*

To further understand the distribution of the observed changes, the Tukey median of the curves and the Modified Hypograph Index (MHI) are computed.

*Figure 3: Immagine MHI (slide) - MHI index on total newborns divided by province.*

The Tukey median corresponds to the province of Torino, while the maximum and the minimum of the index are obtained for the provinces of Bolzano and Oristano, respectively.

*Figure 4: Plot intro A [3] - Normalized number of newborns of each province along years. Torino (red), Bolzano (green) and Oristano (blue) are highlighted.*

In order to investigate the relations of newborns among northern, central and southern Italy, some permutation tests are performed.

The null hypothesis assumes that the distributions of newborns rates belong to the same population, while the alternative hypothesis is its complementary.

A first one-way anova permutation test, performed only on data of 2019, gives evidence to reject the null hypothesis, meaning that the spatial factor is relevant to the distribution of the curves.

*Figure 5: Plot intro A [4] - Boxplot of the original and the permuted data divided by northern, central and southern Italy.*

*Figure 6: Plot intro A [5] - Empirical cumulative distribution function and histogram of the test statistic of the permutation test.*

Consequently, the one-way anova permutation test is extended to all the years. Again, it gives evidence in favour of the alternative hypothesis, confirming that the spatial factor is actually relevant among all years.

*Figure 7: Plot intro A [6] - Trend of the p-value.*

Lastly, the perspective is shifted on the curves and focused on analyzing differences in newborns trends among northern, central and southern Italy. In this permutation test, the null hypothesis assumes that the Tukey medians of northern, central and southern Italy are not significally different in their trend. The implemented test statistic is the sum of the absolute values of the differences between each median.

*Figure 8: Plot intro A [7] - Normalized number of newborns of each province along years. The medians of northern (blue), central (red) and southern (green) Italy are highlighted.*

The p-value shows that we cannot reject the null hypothesis, thus confirming that the spatial factor is not influent in the differences of medians among northern, central and southern Italy.

*Figure 9: Plot intro A [8] - Empirical cumulative distribution function and histogram of the test statistic of the permutation test.*

**DATASET DESCRIPTION**

The dataset consists of:

* A main dataset which, for each year in 2002-2021 and for each Italian province, presents the number of newborns every 1000 women of the same age, from 17 to 50 years old.
* Covariates: **to be defined.**

**MATHEMATICAL FRAMEWORK**

For each year, and for each province, the fertility rates of the newborns for 1000 women are mathematically speaking a function f:N^34->R, in fact for each year the ratio (*total\_newborns/total\_women \* 1000*) gives a value which is a real number, the starting space (N34) is the space of all possible ages’ groups for the women (less than 17, 18, 19, … , 49, more than 50).

*Even though each record involves only discrete values, these values reflect a smoothing variation in the rates that could be assessed in principle, as often as desired, and is therefore a rate function. Ramsay, Silverman, fda, 2005 springer*

So, according to them, our data are functions f\_i,j :R->R i = 2002, .., 2021; j=…

The recording of these data are not exactly precise, in fact these are projection studies made by istat (reference sui metadata di come vengono raccolti, appunti Fra) and are therefore prone to some error or noise. The actual model is so *n\_ijk = f\_ij(k) + eps\_ijk.*where the error parameter eps is assumed to be unbiased.

As *Joop de Beer* said, in literature usually parametric have been used, but with modern tools, nonparametric splines are more able to describe all kinds of age patterns, and so we decided to smooth these curves in a nonparametric fashion in order to obtain the {f\_ij} and to use them as our raw data.

In order to perform the smoothing of these curves, it is important to note that we are going to analyze not only the curves themselves, but even their derivatives. Indeed we plan to make inference also on both the first and second derivatives of the curves, given that they represent a sort of speed and acceleration in the birth rates, and they may be informative.

Given this fact, and given that in principle the curves are continuous on a compact set {f\_ij} \in C0[17,50] (in fact, before the 17-th year and after the 50-th the rates are exactly zero) and without loosing generality we reduce the domain of our function to C5[17,50], in order to be able to perform a smoothing with penalization on the fifth derivative, which result in a degree-5 spline basis thanks to the theorem.

The last thing we had to figure out before the actual smoothing was the number of basis, and we decided to select the number of basis via a generalized cross-validation approach, where we summed the GCV obtained for each curve in the dataset (for each year and for each province). The result is 9 basis.

The raw data present the typical shape of the fertility rates, with a peak around 30 years old and the tendency to the zero in the neighbourhood of 17 and 50 years old. Given the mathematical framework of the FDA, we smoothed these curves with the parameters obtained in the previous section. An important thing to notice is the non reliability of the values at the boundary of the domain, indeed being those values predicted and cumulative (the actual values are below 17 years old and above 50), in the neighborhood of those points the values are not informative.

Particular attention is given to the second derivative of these curves, in fact this derivative can be interpreted as acceleration of the curves, and in this context they represent the trend of the newborns in the following year, indeed we will focus in particular on the first maximum of the second derivative, which represent the age at which the majority of women start to have children.

Analogously, we tried to give an interpretation of surface to the rate in time, resulting in a surface for each province.

Each province has a function g\_i :N34xN20 -> R, and we want to smooth them and to consider them as they are g\_i:R2->R.

For the same argoument as before, there are in C5[17,50]x[2002,2021], and we want to smooth them with the tensur product splines, in order to restore some relationship between them and the respective univariate smoothing splines. Moreover, time and age are two in principle independent field, and the tensor product spline are the best tool to handle this situation.

*Polacco dice GLAM, da vedere.*

**INTRO**

The main dataset has three dimensions: ages of women, years and provinces.

Note that for the introduction we had reduced the space removing the women’s age. Analogously, we want now to reduce the space integrating out one-at-a-time the other two variables and performing exploratory analysis in this reduced space.

**INTRO\_B**

In this first exploratory analysis, we integrate out the effect of the geographic region where the rates come from, and we focus our attention on the curves of the rates in the 20 years of analysis. To do this, a new dataset was downloaded containing the rate extracted in the same way as those for the provinces, but for the entire nation, and these data are divided in the various years.

An introductive quantitative analysis can be carried out through the depth measures, indeed having 20 lines, one for each year, the MEI represent the portion of lines which is below that specific one, so it is an overall measure of how the rates are above or below with respect to the other curves in the 20 years. It is an overall measure, interpreted as a trend, of how these rates are varying in the years, confirming the trend of decreasing of these rates, noting that this result is correlated but not the same result obtained considering the overall newborns, indeed these are rates, which are decreasing but in a different manner with respect to the total babies born in a specific year, which are the result of these rates integrated with respect to the total population in that year.

**INTRO\_C**

The framework considers ages of women and provinces. In order to have a more complete view, a weighted average of the number of newborns in the last three years - 2019, 2020 and 2021 - is taken into account, with weights of 1/6, 1/3 and 1/2, respectively.

*Figure 1: Plot intro C [1] - Weighted average of the number of newborns against women’s ages, for each province.*

Some depth measures are computed. The Tukey median corresponds to the province of Torino, the maximum and the minimum of the Modified Hypograph Index to Bolzano and Oristano, respectively.

*Figure 2: Plot intro C [2] - Weighted average of the number of newborns of each province against women’s ages. Torino (red), Bolzano (green) and Oristano (blue) are highlighted.*

A permutation test on the Tukey medians of northern, central and southern Italy is computed. The null hypothesis assumes that the differences between medians summed are negligible. The implemented test statistics is the sum of the products of the difference between the medians. The test allows us to accept the null hypothesis.

*Figure 3: Plot intro C [3] - Weighted average of the number of newborns of each province against women’s ages. The medians of northern (blue), central (red) and southern (green) Italy are highlighted.*

*Figure 4: Plot intro C [4] - Empirical cumulative distribution function and histogram of the test statistic of the permutation test.*

The same approach is used to perform the same type of test on the derivatives of the curves.

For the first derivative the difference in the province is again negligible, leading to the acceptance of the null hypothesis, with a pvalue of 0.89%.

Even for the second derivative we’re not able to reject the null hypothesis with the differences between median, so the differences in the geographic location cannot be restored from these simple analysis, leading us to much more complex studies.