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

Spiegare come mai vogliamo capire il problema, a chi può interessare, da che prospettiva vogliamo capirlo (fertility rates e non nati totali), come mai sono importanti differenze territoriali, sono davvero territoriali o economiche?

WHY NONPARAMETRIC?

We have decided to go for a nonparametric approach since the data we’re dealing with are purely functional (**da specificare gli spazi coinvolti**).

Da specificare nel framework che usiamo nonparametric estimations perché sono meglio, nonparametric tests per inferenza.

MATHEMATICAL FRAMEWORK

**Da decidere**

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: **da completare**

INTRO

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

It has been decided to perform a first exploratory analysis, taking into account two of the three dimensions each time.

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.*

INTRO\_B

In this second exploratory analysis, we integrate out the effect of the geographic region where the rates are 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.

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 neighborhood 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 they can be interpreted as the acceleration of these curves, and in this context they represent the age in which females start having a considerable number of children (**mamma mia terribile da riscrivere ma sono stanco non mi viene meglio di così**).

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, conferming 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 borned 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 are equal. 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.*