# Exploratory Data Analysis

During the data story part, we saw that it seems there is a correlation between the number of doctors and the population in France. According to the scatter plot showing the population depending on the doctor number and the heatmaps showing variations in population and number of doctors throughout France, we saw there is a higher number of doctors in location with a large population. Let us now take a closer look at the relationship between these two variables from a statistical point of view and regarding the location. To do this, we will perform first a correlation test between these two variables. Then, we will do a one-way ANOVA between the ratio of the number of doctors according to the population in northern France and southern France. Finally, we will do the same thing depending on the regions of France. To do this, we will use the dataset Population_Doctor_by_zipcode, which includes the number of doctors, the population, the ratio between both variables, the regions and the north/south class grouped by zipcode. We took a threshold $\alpha$=0.05.

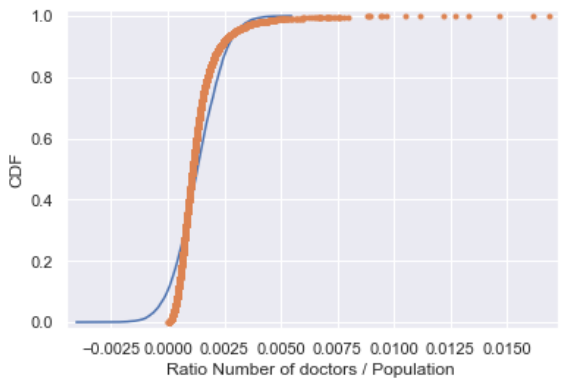
## Hypothesis test correlation between the number of doctors and the population

Null Hypothesis: the two variables Number of Doctors and Population are completely uncorrelated.

We have simulated the data assuming the null hypothesis is true. We had using Pearson correlation, p as the statistic test. Then, using bootstrapping resampling, we had computed p-value as fraction of replicates that have p at least as large as observed. The procedure done, and in all 10000 of the replicates under the null hypothesis not one had a Pearson correlation coefficient as high as the observed value of 0.65. The p-value is very small, so we have rejected the null hypothesis. There is a significant correlation between the number of Doctors and the Population. There is a higher Number of Doctors for the more populated territories.

## One-Way ANOVA between the ratio number of doctors / population on the North and the ratio on the South

First, let's check the normality of the values by plotting the cumulative distribution function. According to the CDF plot, the data seem normally distributed.



Here a summary of the data using rp.summary_cont().

| North_South | N | Mean | SD | SE | 95% Conf. | Interval |
|---|---|---|---|---|---|---|
| North | 3228 | 0.001215 | 0.001010 | 0.000018 | 0.001180 | 0.001250 |
| South | 2364 | 0.001458 | 0.001092 | 0.000022 | 0.001414 | 0.001502 |

Then, using stats.f_oneway(), we calculated the ANOVA between the ratio Doctors number / Population on the North of the France and the ratio on the South of the France and we obtained:

```
F_onewayResult(statistic=73.51922931198455, pvalue=1.2747156042454852e-17)
```

There is a significant difference between the ratio on the North and the ration on the South of the France. The ratio Doctors number / Population is higher on the South of the France.

## One-Way ANOVA between the ratios number of doctors / population depending on the location (Régions)

Here a summary of the data:

| NOM_REG | N | Mean | SD | SE | 95% Conf. | Interval |
|---|---|---|---|---|---|---|
| AUVERGNE-RHONE-ALPES | 702 | 0.001450 | 0.001134 | 0.000043 | 0.001366 | 0.001534 |
| BOURGOGNE-FRANCHE-COMTE | 349 | 0.001095 | 0.000703 | 0.000038 | 0.001021 | 0.001169 |
| BRETAGNE | 308 | 0.001303 | 0.000819 | 0.000047 | 0.001211 | 0.001394 |
| CENTRE-VAL DE LOIRE | 235 | 0.001114 | 0.001352 | 0.000088 | 0.000941 | 0.001287 |
| CORSE | 47 | 0.001375 | 0.000949 | 0.000138 | 0.001101 | 0.001650 |
| GRAND EST | 607 | 0.001230 | 0.000887 | 0.000036 | 0.001159 | 0.001300 |
| HAUTS-DE-FRANCE | 592 | 0.001255 | 0.000881 | 0.000036 | 0.001184 | 0.001326 |
| ILE-DE-FRANCE | 492 | 0.001497 | 0.001366 | 0.000062 | 0.001376 | 0.001618 |
| NORMANDIE | 346 | 0.001025 | 0.001127 | 0.000061 | 0.000906 | 0.001144 |
| NOUVELLE-AQUITAINE | 627 | 0.001340 | 0.000955 | 0.000038 | 0.001266 | 0.001415 |
| OCCITANIE | 613 | 0.001527 | 0.001141 | 0.000046 | 0.001437 | 0.001618 |
| PAYS DE LA LOIRE | 299 | 0.000989 | 0.000635 | 0.000037 | 0.000917 | 0.001061 |
| PROVENCE-ALPES-COTE D'AZUR | 375 | 0.001564 | 0.001146 | 0.000059 | 0.001448 | 0.001680 |

Then we calculated the ANOVA between the ratios Doctors number / Population depending on the region of the France and we obtained:

```
F_onewayResult(statistic=13.474806947107517, pvalue=6.513761694105628e-28)
```

There is a significant difference between the ratios depending on the region of the France. Then we have made some post hoc comparison using Bonferroni Correction to analyse if there is a significant

difference between the ratios of specific regions. For example, let's see if there are significant differences between the ratios in Ile-de-France where the capital of France is located and other regions using stats.ttest_ind(). We obtained:

- Bonferroni Correction between Ile-de-France and Bretagne: Ttest_indResult(statistic=2.25912195618384, pvalue=0.02414486706798545)
- Bonferroni Correction between Ile-de-France and Provence-Alpes-Côte-d'Azur:Ttest_indResult(statistic=-0.7638462263829737, pvalue=0.44516718045069736)
- Bonferroni Correction between Ile-de-France and Nouvelle-Aquitaine: Ttest_indResult(statistic=2.257721053343628, pvalue=0.024155482253022397)
- Bonferroni Correction between Ile-de-France and Occitanie: Ttest_indResult(statistic=-0.3968092790129793, pvalue=0.6915848692456656)

We can see there are significant differences between the ratio of the Ile-de-France and the ratio of the Nouvelle-Aquitaine or the ratio of the Bretagne. And there is no significant difference between the ratio of the Ile-de-France and the ratio of the Occitanie and the ratio of the Provence-Alpes-Côte-d'Azur. Then the ratio Doctors number / Population is very different according to the France regions and it's not depending only of the location of the region (North vs. South). Indeed, Ile-de-France and Bretagne are located on the North and have significantly different ratios and Occitanie or Provence-Alpes-Côte-d'Azur are on the South and do not have significantly different ratios with that of the Ile-de-France. They could be interesting features to predict the number of doctors in the next part.