P&S research project: interim report

Team: Serhii Matsyshyn, Serhii Ivanov

Dataset:

We will use publicly available data from the United Nations Department of Economic and Social Affairs. They provide a wide range of datasets concerning almost all aspects of demography.

We will combine two datasets ("Population by age, sex, and urban/rural residence" and "Deaths by age, sex, and urban/rural residence") to have a dataset consisting of countries, census years, urban/rural residence, gender, age groups, the number of people in each group, and the number of deaths for each age group.

Using such a wide dataset allows us to obtain new intermediate parameters (for example, the percentage of the urban/rural population grouped by other demographic parameters). That is, in the final result, this enriched dataset can be supplemented by us with new parameters, which can help us find interesting dependencies.

In the resulting dataset, we will have up to 172 countries, with demographic information collected from 1948 year to the 2022 year.

Hypothesis (1): the birth rate changes depending on the population ageing

Task: check this assumption on some country (or a group of countries) where there is a problem of an "aging" population.

Parameters that will be analyzed: age structure of the population in the country under study (change in the age structure of the population over time), birth rate (change in birth rate over time).

Research directions: to compare the impact of changes in the age structure of the population on changes in birth rates; determine average values by year; trace the linear relationship between these characteristics (linear regression analysis); visualization of research data.

Possible tests to be conducted:

- Linear regression analysis: This can be used to examine the relationship between the age structure of the population (of a country) and the birth rate over time. We can plot the data on a plot and fit a linear regression line to see if there is a positive or negative relationship between the two variables.

- T-test: A t-test could be used to determine whether there is a significant difference in the mean birth rate between two different age groups or between two different time periods. For example, we can compare the mean birth rate for individuals in the 20-30 age group to the mean birth rate for individuals in the 30-40 age group to see if there is a difference in birth rates between these two groups.

- Chi-square test: This test could be used to examine the relationship between the age structure of the population and the birth rate. For example, we can compare the observed frequency of births in different age groups to the expected frequency of births, and use the chi-square test to determine whether there is a significant difference between the two.

Hypothesis (2): the growth of the urban population has affected the average life expectancy of the population.

Task: to test this hypothesis on the example of one country (or a group of countries).

Parameters that will be analyzed: indicators of the place of residence (village/city); average life expectancy.

Research directions: analysis of changes in the urban population over time; analysis of changes in average life expectancy over time; to compare the dependence between time changes of the investigated indicators; trace the linear relationship between these characteristics (linear regression analysis); visualization of research data.

Tests to be conducted:

- Linear regression analysis: This can be used to examine the relationship between the growth of the urban population and the average life expectancy over time. We can plot the data on a plot and fit a linear regression line to see if there is a positive or negative relationship between the two variables.

- T-test: A t-test could be used to determine whether there is a significant difference in the mean life expectancy between urban and rural populations at a given point in time.

- Chi-square test: This test could be used to examine the relationship between the place of residence (urban vs. rural) and life expectancy. For example, we can compare the observed frequency of individuals with different life expectancies in urban and rural areas to the expected frequency, and use the chi-square test to determine whether there is a significant difference between the two.