TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Dr. Crystal Wiedner
Dr. Marlies Hafer
Dr. Rouven Michels

Author: Harsh Ganpatbhai Rana (229989)

November 3, 2024

Contents

1 Introduction

Demographic data provides essential information about population characteristics, helping governments and organizations create policies and services tailored to different communities' needs. The International Database (IDB, 2024) from the U.S. Census Bureau is a key source for this type of data, covering over 227 countries and providing detailed statistics. This dataset includes variables such as life expectancy at birth and under-five mortality rates, grouped by gender and divided across various regions. With data starting from 1950 and projections extending to 2100, (IDB, 2024) data is crucial for understanding global population trends and planning for the future.

This report aims to provide a descriptive analysis of life expectancy and under-five mortality rates for 227 countries, using data from the years 2004 and 2024. By comparing these variables over a 20-year span, the goal is to understand how they have changed over time, focusing on differences by gender and region. One objective is to see if these variables remain relatively consistent within certain subregions but show variation between different regions. Additionally, we look for any correlations between life expectancy and under-five mortality rates, which could reveal important relationships relevant to policy and planning.

Several statistical methods are applied to address these questions. Frequency distributions show the spread of each variable in 2004 and 2024, while measures like the mean and median provide an idea of typical values, and standard deviation shows variability within regions. The Pearson correlation coefficient is used to find possible relationships between life expectancy and mortality rates. Various charts—such as histograms, scatter plots, and box plots—are also used to make the data easier to interpret and to highlight patterns across regions, genders, and over time.

This report is organized as follows, Section 2 explains the problem statement, describing the dataset, its source, and the goals of the analysis. Section 3 covers the statistical methods and tools used, including formulas and software. Section 4 presents the results of the analysis, with visualizations to illustrate key findings. Finally, Section 5 concludes with a summary of the main results and suggestions for future research.

2 Problem statement

This section provides a detailed description of the dataset, including its source, data collection methods, quality considerations, and the specific objectives of this analysis.

2.1 Source and Quality of Data

The dataset used in this analysis is a subset from the International Database IDB (2024), a comprehensive repository maintained by the U.S. Census Bureau. The (IDB, 2024) contains demographic data on more than 227 countries, covering population indicators from 1950 through 2100. The database compiles data from various sources, including national censuses, surveys, and administrative records. When national data is unavailable or outdated, the Census Bureau uses statistical techniques to estimate values, ensuring consistent and up-to-date information across regions.

This specific dataset includes life expectancy at birth and under-five mortality rates for 227 countries for the years 2004 and 2024. The data is stratified by gender and organized by five major regions and 21 subregions. This dataset serves as a valuable resource for examining demographic trends over a 20-year period, allowing for in-depth analysis of population health and social development.

Variable Name	**Data Type**
Under5_Mortality_Both	Numeric (Continuous)
Under5_Mortality_Males	Numeric (Continuous)
Under5_Mortality_Females	Numeric (Continuous)
Life_Expectancy_Both	Numeric (Continuous)
Life_Expectancy_Males	Numeric (Continuous)
Life_Expectancy_Females	Numeric (Continuous)
Region	Nominal
Subregion	Nominal
Country	Nominal
Year	Numeric (Discrete)

Table 1: Description of Variables

The variable "Under5_Mortality" represents the number of children who die before reaching the age of 5 per 1,000 live births. "Life_Expectancy" measures the average number of years a newborn is expected to live if current mortality rates remain constant, with values available for both sexes combined, as well as separately for males and females.

The "Country" variable lists the names of the 227 countries included in the dataset, while "Region" and "Subregion" indicate the geographical classifications (continent and subregion) for each country. "Year" specifies the specific year for each observation, allowing for comparisons over time.

This dataset is of high quality, with minimal missing values. Missing data for seven variables was addressed through mean imputation, except for missing values in "Region" and "Subregion," which were removed. As a result, the dataset used for analysis is well-suited for examining demographic trends and ensuring reliable results.

2.2 Objectives

The purpose of this analysis is to examine how life expectancy at birth and under-five mortality rates have evolved over the last 20 years, focusing on patterns by sex and region. The key objectives are as follows, to describe the frequency distributions of the life expectancy and under-five mortality variables, examining differences by sex and region; to analyze whether these variables show relative consistency (homogeneity) within subregions and greater variation (heterogeneity) between subregions, by assessing the spread within specific regions and comparing central values across regions; to investigate any bivariate correlations between life expectancy and under-five mortality, exploring if higher life expectancy corresponds to lower under-five mortality; and to observe changes in these variables between 2004 and 2024, identifying any notable shifts over the 20-year period. This analysis will employ descriptive statistics and visualization techniques to achieve these objectives, contributing to an understanding of global demographic trends and their implications for public health and policy.

3 Statistical methods

3.1 Descriptive statistics

This section covers the main descriptive statistics used to analyze the dataset, including measures of central tendency, variability, and relationships between variables.

3.1.1 Mean

The mean, or average, is a central measure that represents the typical value in a dataset and is useful for summarizing data into a single, representative figure that describes the "center" of the data. To calculate the mean, we add up all the values in a dataset and then divide by the total number of observations. The mean, often denoted by the symbol μ, is calculated as follows

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

where N is the total number of observations, and $x_1, x_2, \ldots, x_N$ are the individual data values in the dataset. This formula provides the arithmetic mean, which is sensitive to every value in the dataset. Because each value contributes equally, the mean can be influenced by extreme values (outliers), making it higher or lower than most values in the dataset. In this analysis, the mean is used to compare average life expectancy and under-five mortality rates across countries and years, providing a foundational understanding of central tendencies in the demographic data. (Black, 2009, pg. 49-50)

3.1.2 Median

The median is another measure of central tendency that represents the middle value of a dataset when it is ordered from smallest to largest. Unlike the mean, the median is not affected by extreme values, making it a better indicator of the "typical" value when the dataset contains outliers or is skewed. To find the median, we arrange all data values in ascending order. If the number of values (n) is odd, the median is the middle value. If n is even, the median is calculated as the average of the two middle values. The median is represented by $\tilde{x}$, and it is calculated as follows

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

where $x_1, x_2, \ldots, x_n$ are the ordered values, and $\tilde{x}$ denotes the median. The median helps provide a realistic "center" for the data, especially in cases where the data distribution is skewed. It is particularly useful in our demographic analysis to understand the central life expectancy or mortality rates without distortion from extreme values. (Hay-Jahans, 2017, pg.75-76)

3.1.3 Variance and standard deviation

Variance and standard deviation are measures of variability that show how much the values in a dataset differ from the mean. The variance (σ^2) measures the average of the squared differences between each value and the mean, which helps us understand how spread out the values are. A higher variance indicates that the values are more spread out, while a lower variance suggests they are closer to the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

where $x_1, x_2, \ldots, x_N$ are the individual data points, μ is the mean of the dataset, and N is the total number of data points. The standard deviation, denoted by σ, is the square root of the variance

$$\sigma = \sqrt{\sigma^2}$$

The standard deviation is in the same units as the original data, making it easier to interpret than the variance. Both variance and standard deviation are helpful for examining how much variability exists in life expectancy or mortality rates across different countries or regions. These measures allow us to assess whether there is a consistent pattern or significant variation within the data, which can point to differences in demographic trends across areas or time periods. (Black, 2009)

3.1.4 Five-Number summary

The five-number summary provides a quick snapshot of the data distribution and includes the minimum (the lowest observed value), first quartile (Q1) (the 25th percentile, where 25% of the data lies below this value), median (the 50th percentile, or middle value), third quartile (Q3) (the 75th percentile, where 75% of the data lies below this value), and maximum (the highest observed value). Quartiles help to divide the data into four equal parts. The first quartile (Q1) marks the point below which 25% of values fall, and the third quartile (Q3) is the point below which 75% of values fall. The inter-quartile range (IQR), defined as $Q3 - Q1$, measures the spread of the middle 50% of the data.

$$\text{IQR} = Q3 - Q1$$

Values lying more than 1.5 times the IQR below Q1 or above Q3 are considered outliers, which might suggest unusual demographic trends or exceptional cases in our data. This summary gives a clear picture of the data spread and highlights any extreme values, aiding in understanding overall patterns in life expectancy and under-five mortality. (Hay-Jahans, 2017, pg.85)

3.1.5 Correlation

The correlation coefficient measures the strength and direction of the relationship between two variables. It helps us understand how two variables change together. A positive correlation means that as one variable increases, the other does too, while a negative correlation indicates that as one increases, the other decreases. Correlation values range from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

This measure is particularly useful for our analysis as it allows us to examine potential connections between life expectancy and under-five mortality rates. For example, a negative correlation could suggest that as life expectancy rises, under-five mortality tends to decrease, or vice versa. (Black, 2009, pg. 466)

3.1.6 Pearson correlation

The Pearson correlation coefficient, denoted as r, is a specific measure for assessing the strength and direction of a linear relationship between two continuous variables. For two variables X and Y, with observations x_i and y_i, means $\bar{x}$ and $\bar{y}$, and variances σ_X^2 and σ_Y^2, the formula is

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

where r values close to 1 or -1 indicate a strong linear relationship, while values near 0 suggest little to no linear relationship. The Pearson correlation helps to identify whether high values of one variable (e.g., life expectancy) tend to be associated with high or low values of another variable (e.g., mortality rate). This method is especially helpful for understanding relationships between life expectancy and under-five mortality rates, providing insights into how these indicators might be connected across regions and over time. However, it is important to note that a zero correlation does not imply a complete lack of relationship, as a non-linear association may still exist. (Hay-Jahans, 2017, pg. 321-322)

3.2 Visualization techniques

Visualizing data is essential for identifying patterns, trends, and outliers. This section describes the main visualization techniques used in this report: histograms and boxplots.

3.2.1 Histogram

A histogram is a graphical representation of the distribution of a numerical dataset. The dataset is divided into intervals, or "bins," and each bin counts the number of data points that fall within it. Each bin is represented as a bar, where the bar's height reflects the count (frequency) of values in that interval, and the area of the bar (width multiplied by height) approximately represents the relative frequency of each category. In a histogram, the choice of bin width is essential, as it affects the histogram's appearance and can reveal or obscure data patterns. The Freedman-Diaconis rule, used in this report, is a popular method for selecting bin width in a way that balances detail with clarity. The number of bins, k, is calculated as follows

$$k = \frac{\max(x) - \min(x)}{2 \cdot \frac{\mathrm{IQR}}{\sqrt[3]{n}}}$$

where $\max(x)$ and $\min(x)$ are the maximum and minimum values in the dataset, n is the sample size, and IQR stands for the inter-quartile range. The Freedman-Diaconis formula adjusts the bin width based on data variability, which helps to ensure that the histogram accurately reflects the data distribution, especially when there are outliers or a skewed distribution. (Hay-Jahans, 2017, pg.131)

3.2.2 Boxplot

A boxplot, also known as a box-and-whisker plot, is a visualization tool that provides a summary of the distribution, spread, and skewness of numerical data. A boxplot displays the median, interquartile range (IQR), and potential outliers in the dataset. The central box represents the IQR, spanning from the first quartile (Q1) to the third quartile (Q3), with a line inside the box representing the median. The "whiskers" extend from the box to show the range of the data, typically reaching the smallest and largest values within 1.5 times the IQR from Q1 and Q3. Points outside this range are plotted individually as potential outliers.

The boxplot's structure allows for easy identification of data symmetry and skewness. If the right whisker is longer than the left, or if the median is closer to Q1, the data is right-skewed. If the left whisker is longer than the right, or if the median is closer to Q3, the data is left-skewed. If both whiskers are approximately equal and the median is centered within the box, the data is considered symmetric. Boxplots are particularly useful for identifying outliers and comparing distributions across multiple groups, such as life expectancy rates across regions. (Hay-Jahans, 2017, pg. 137-138)

3.3 Software

For this analysis, Python version 3.9.0 Python Software Foundation (2020) was used. Python provides an extensive range of libraries that are essential for data analysis and visualization. The key libraries used in this analysis includes Pandas for data manipulation and analysis, NumPy for numerical operations, and Matplotlib and Seaborn for data visualization.

4 Statistical analysis

4.1 Descriptive Analysis

This section analyzes the distribution of the following six variables across 227 unique countries in 2024 such as Under-5 Mortality Rate (Both Sexes), Under-5 Mortality Rate (Males), Under-5 Mortality Rate (Females), Life Expectancy at Birth (Both Sexes), Life Expectancy at Birth (Males), and Life Expectancy at Birth (Females).

Figure 1 presents histograms of life expectancies for males, females, and both sexes. The histograms reveal that life expectancy data is left-skewed, with most countries showing higher life expectancies clustered closer to the upper range. The x-axis represents life expectancy in years, while the y-axis shows the frequency of countries within each range. Notably, there are only a few countries with life expectancy below 60 years, indicating overall improvements in global health.

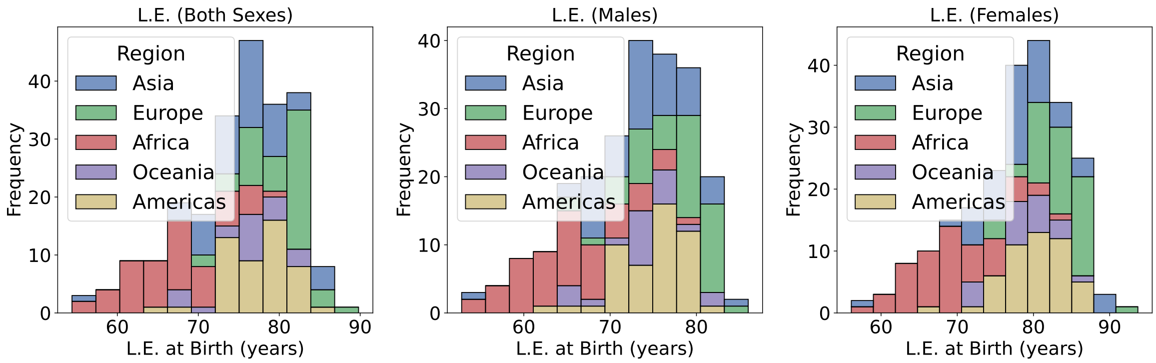

Figure 1: Histogram of Life Expectancy at Birth for Males, Females, and Both Sexes by Region

In Figure 2, histograms for under-five mortality rates for both sexes, males, and females are displayed. This data, unlike life expectancy, shows a right-skewed distribution, with the majority of under-five mortality rates ranging from 2 to 12 per 1,000 live births. The mean under-five mortality rate is highest for males, at approximately 27.38, compared to females and both sexes combined. This right-skewed pattern suggests that while most countries have low child mortality, there are still some with considerably higher rates, likely influenced by factors such as healthcare access and socioeconomic conditions. In these histograms, the x-axis represents mortality rate bins of width 10, while the y-axis denotes frequency.

According to the dataset, the average difference in life expectancy between males and females is approximately 5 years, with females generally showing higher life expectancy. This gender gap in life expectancy is visualized in Appendix Figure 7, which shows the frequency distribution of life expectancy differences between sexes. This difference is symmetrically distributed, with a positive mean difference, reflecting that, on average, females have a higher life expectancy at birth.

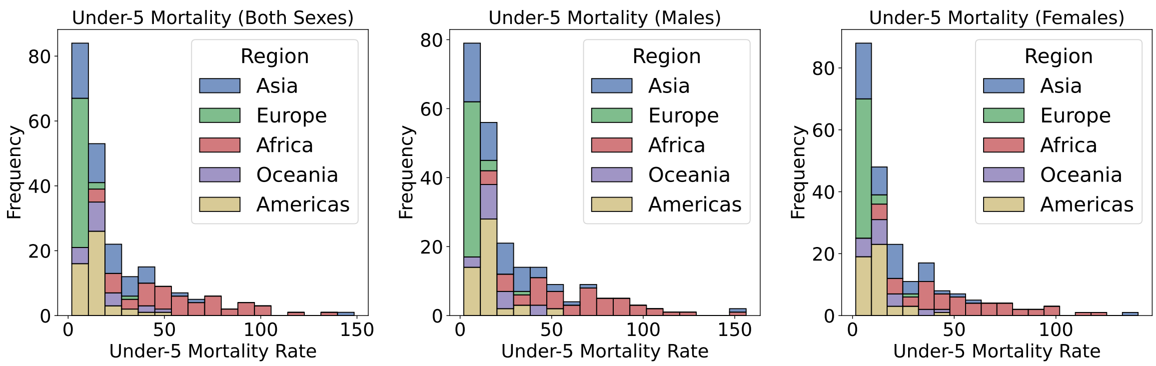

Figure 2: Histogram of Under-5 Mortality Rates for Males, Females, and Both Sexes by Region

4.2 Variable's Value Comparison

This section presents a comparative analysis of life expectancy and under-five mortality rates by subregions within the Asia region for the year 2024. Boxplots are used to illustrate the variation in these variables across Eastern Asia, South-Central Asia, South-Eastern Asia, and Western Asia, with separate plots for males, females, and both sexes combined. The interquartile range (IQR) of each boxplot provides insight into the variability within each subregion. When the IQR is shorter, we assume greater homogeneity within the subregion, indicating that life expectancy or under-five mortality is relatively consistent. Conversely, wider IQRs suggest more heterogeneity within a subregion.

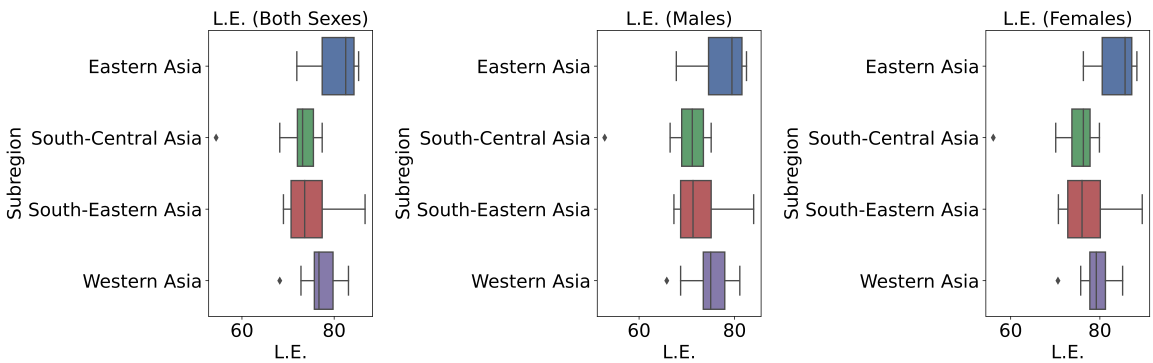

Figure 3: Boxplot for Asia Region (Life Expectancy at Birth for Males, Females, and Both Sexes)

Figure 3 displays the boxplots for life expectancy at birth for males, females, and both sexes in the Asia region. The Western Asia subregion exhibits the highest homogeneity, as indicated by its smaller interquartile range across all categories, reflecting relatively consistent life expectancy values. In contrast, the Eastern Asia subregion has a notably larger IQR, suggesting greater variability in life expectancy among countries within this subregion. Comparing the median life expectancies of South-East Asia and South-Central Asia, both subregions show closely aligned median values, implying low heterogeneity between them. Outliers are present in each of the subregions, particularly within South-Central Asia and Western Asia.

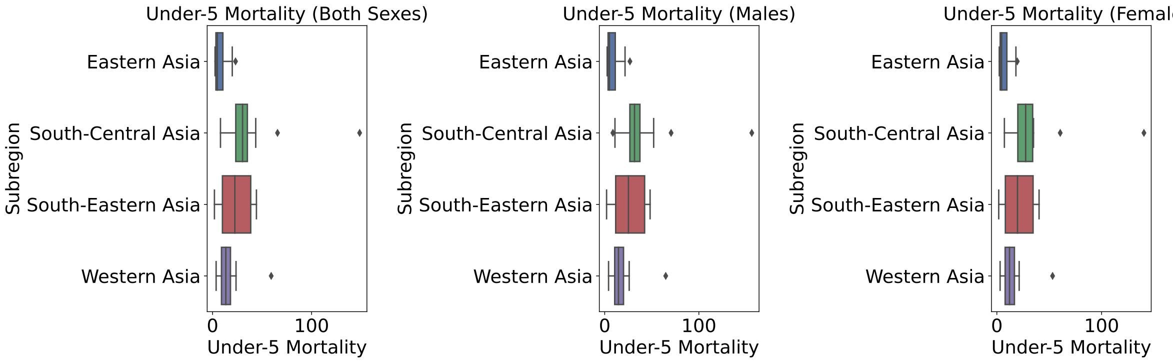

Figure 4: Boxplot for Asia Region (Under-5 Mortality Rates for Males, Females, and Both Sexes)

In Figure 4, boxplots of under-five mortality rates for both sexes, males, and females are shown for the Asia region. For both sexes combined, East Asia and West Asia show the highest level of homogeneity, indicated by their very small IQRs. This suggests a similar range of under-five mortality rates across countries within these subregions. The South-Central Asia subregion also demonstrates relatively high homogeneity, though with a slightly broader IQR compared to East Asia and West Asia. South-East Asia, however, exhibits the least homogeneity, as evidenced by a wider IQR. This pattern is consistent across both sexes, as well as for males and females separately. Outliers are also observed in the under-five mortality rates, particularly within South-Central Asia and South-East Asia.

4.3 Analysis of Correlation

In this section, we examine the correlation coefficients among the key continuous variables, Life Expectancy (for both sexes, males, and females) and Under-5 Mortality (for both sexes, males, and females) for the year 2024. The following abbreviations are used for readability: "L.E." stands for Life Expectancy, and "UM" represents Under-5 Mortality. Table 2 below provides the Pearson correlation coefficients among these variables.

	L.E. Both	**L.E. Males**	**L.E. Females**	**UM Both**	**UM Males**	**UM Females**
L.E. Both	1.00	0.99	0.99	-0.89	-0.89	-0.89
L.E. Males	0.99	1.00	0.97	-0.87	-0.87	-0.86
L.E. Females	0.99	0.97	1.00	-0.90	-0.90	-0.90
UM Both	-0.89	-0.87	-0.90	1.00	1.00	1.00
UM Males	-0.89	-0.87	-0.90	1.00	1.00	0.99
UM Females	-0.89	-0.86	-0.90	1.00	0.99	1.00

Table 2: Pearson Correlation Coefficients among Life Expectancy and Under-5 Mortality Variables

The correlation table reveals several key patterns in the relationships among life expectancy and under-five mortality rates.

Firstly, there is a high positive correlation among the life expectancy variables for both sexes, males, and females. This indicates that life expectancy for both sexes is highly aligned with life expectancy for males and females individually. The strong positive correlation suggests that as life expectancy increases for one group (e.g., males), it also increases proportionally for the other groups (e.g., females and both sexes). This consistency highlights the uniformity in life expectancy trends across genders.

Secondly, we observe a high negative correlation between life expectancy (for both sexes, males, and females) and under-five mortality rates (for both sexes, males, and females). This negative correlation indicates an inverse relationship: as life expectancy rises, under-five mortality rates decrease proportionally. This relationship aligns with public health expectations, as improvements in healthcare, living standards, and socio-economic conditions tend to increase life expectancy while reducing early childhood mortality.

Lastly, the under-five mortality rates for both sexes, males, and females are also highly positively correlated with each other. This implies that under-five mortality rates are consistent across genders, with changes in mortality rates for one group (e.g., males) reflected similarly in the other groups (e.g., females and both sexes). This consistency suggests that underlying factors affecting under-five mortality, such as healthcare access and environmental conditions, impact all groups similarly.

4.4 Comparison over the Last 20 Years

In this section, we examine changes in variable values between the years 2004 and 2024. Box plots are used to highlight these differences, illustrating shifts in median values and variability over this 20-year period. Figure 5 shows the increase in life expectancy at birth for males, females, and both sexes, while Figure 6 displays the corresponding trends in under-five mortality rates.

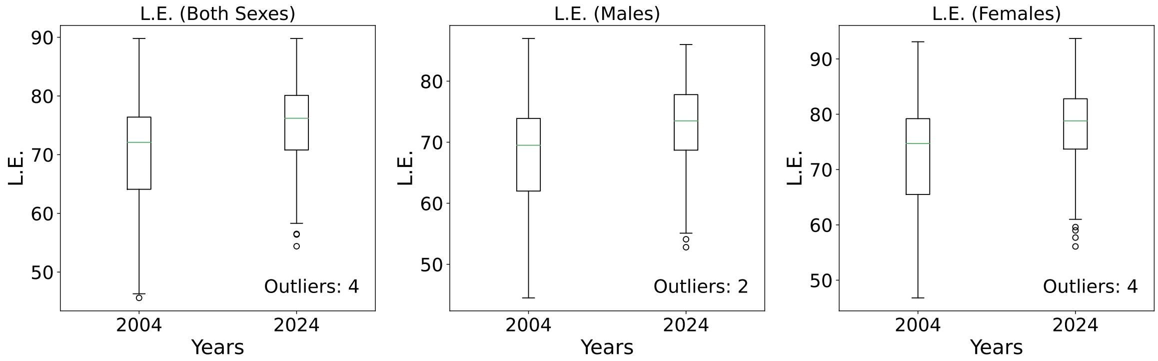

Figure 5: Boxplot showing the difference in life expectancy at birth for males, females, and both sexes between 2004 and 2024

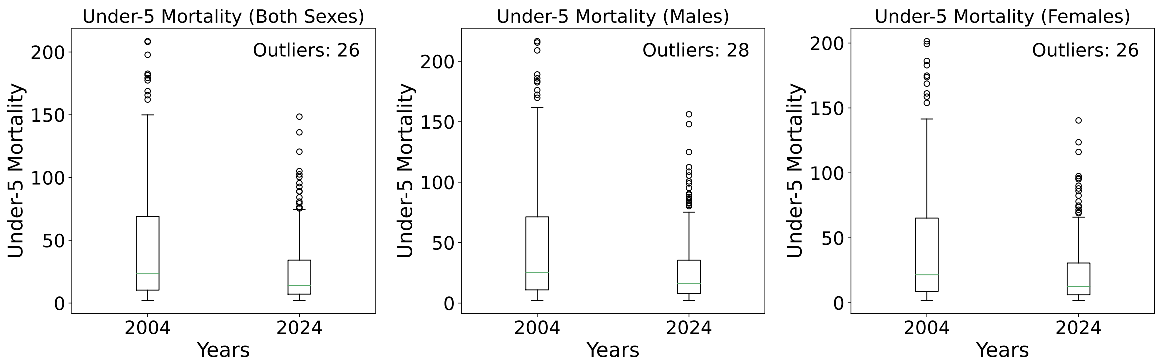

Figure 6: Boxplot showing the difference in under-five mortality rates for males, females, and both sexes between 2004 and 2024

Figure 5 illustrates the differences in life expectancy at birth for males, females, and both sexes between 2004 and 2024. The x-axis represents the years 2004 and 2024, while the y-axis displays the life expectancy values. Across all categories, an increase in the median life expectancy is observed, along with a reduction in variability, indicating greater consistency in life expectancy values among countries. In 2024, three outliers are identified for life expectancy in both sexes, and two outliers are observed for both males and females, indicating some countries with significantly different life expectancies from the norm.

In Figure 6, the box plots illustrate under-five mortality rates for both sexes, males, and females. A notable decrease in the median under-five mortality rate from 2004 to 2024 suggests improvements in child health worldwide. This trend is consistent across both sexes, as well as individually for males and females. Additionally, variability has decreased, pointing to a more uniform reduction in mortality rates across regions. Despite these improvements, there are several outliers in the 2024 data such a 22 for both sexes, 25 for males, and 22 for females. These outliers reflect countries with persistently high under-five mortality rates, contrasting with the overall downward trend.

Comparing the two figures, the majority of life expectancy values are concentrated in the upper regions of the box plots, while under-five mortality rates cluster in the lower sections, reinforcing the positive demographic shifts over the past two decades. The increased median life expectancy and reduced under-five mortality rates underscore the global advancements in health and socioeconomic conditions during this period.

5 Summary

The dataset analyzed in this report was sourced from the International Data Base (IDB, 2024). It contains observations on key demographic variables, including under-five mortality rates and life expectancy for males, females, and both sexes, from 227 countries across 5 major regions and 21 subregions. To understand regional trends in life expectancy and mortality rates over the last 20 years (2004-2024), we applied descriptive statistical techniques such as histograms, correlation analysis, and box plots.

To compare life expectancies by gender, we examined the frequency distribution of the continuous variables. Our analysis shows that, on average, females have a higher life expectancy than males. Further analysis revealed a stronger linear relationship between under-five mortality rates and life expectancy at birth for females compared to that for males. For most countries, under-five mortality rates are now near zero.

In terms of homogeneity and heterogeneity within Asian subregions, South-Central Asia and South-East Asia showed greater similarity in life expectancy values, while East Asia and West Asia were more homogeneous with respect to under-five mortality rates. Regarding correlations, we observed a strong negative relationship between life expectancy and under-five mortality rates, indicating that as life expectancy increases, under-five mortality rates tend to decrease. Over the past two decades, life expectancies have increased, while under-five mortality rates and their variability have declined.

To gain a deeper understanding of the factors affecting life expectancy and mortality rates across different regions, future research could incorporate additional socioeconomic variables, such as per capita GDP or employment rates, to assess their potential influence on health and demographic trends.

Bibliography

Ken Black. *Business Statistics: For Contemporary Decision Making, Sixth Edition*. John Wiley & Sons, Inc., 2009. ISBN 9780470409015.

Hay-Jahans. *An R Companion to Linear Statistical Models*. CRC Press LLC, 2017, 2017. ISBN 9781138116030.

IDB. International database. `https://www.census.gov/data-tools/demo/idb/#/country?COUNTRY_YEAR=2022&COUNTRY_YR_ANIM=2024`, 2024. Accessed: 2024-11-03.

Python Software Foundation. Python, 2020. URL `https://www.python.org/`.

Appendix

Additional Figures

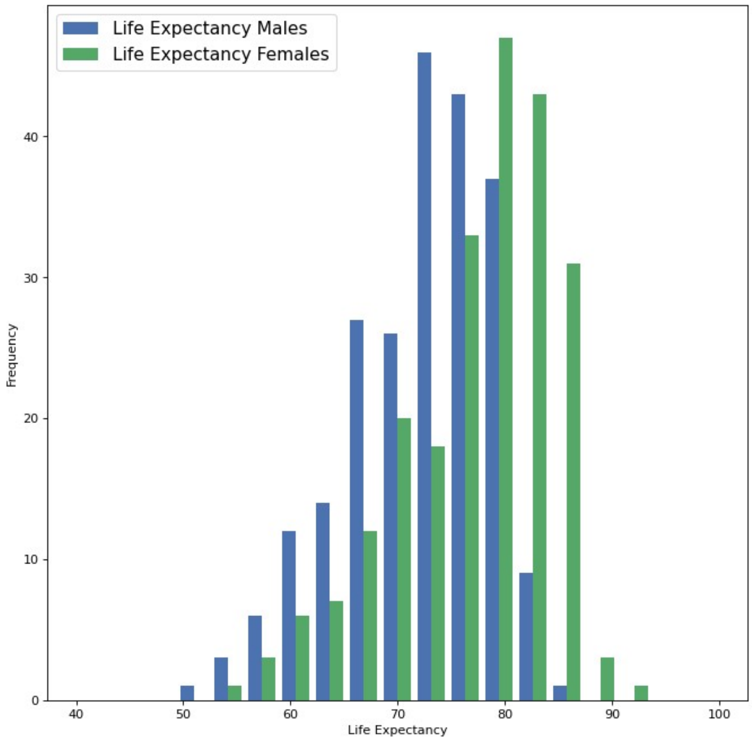

Figure 7: Histogram for comparison of life expectancy of males and females in the year 2024