Exploratory data analisys of Tobacco usage from 2010 to 2020

[*] Essay for the course of Data Science for health system

Marco Venturi
University of Perugia
Perugia, Italy
marco.venturi1@studenti.unipg.it

***Abstract*—This paper aims to study the trend of tobacco usage from 2010 to 2020 using data from the World Health Organization. The analysis provides insights into tobacco usage in specific geographic areas in 2020, with a closer focus on differences in tobacco consumption during different years for both sexes and single sexes. This preliminary analysis forms the basis for statistical tests used to check if the trend has changed in the last decade for both female and male populations. All the graphics and statistical tests are conducted using the R programming language. The study shows that the trends for females and males are decreasing compared to past years.**

***Index Terms*—tobacco, WHO, monitoring**

I. Introduction

The tobacco epidemic is one of the biggest public health threats the world has ever faced, killing over 8 million people a year around the world. More than 7 million of those deaths are the result of direct tobacco use while around 1.3 million are the result of non-smokers being exposed to second-hand smoke [1]. The study proposes to analyze the consumption of tobacco and tobacco products in the world in the period between 2000 and 2020 by examining the difference in incidence between different geographical areas and sexes.

II. Dataset

The dataset analyzed is called 'Non-age-standardized estimates of current tobacco use, tobacco smoking, and cigarette smoking (Tobacco Control: Monitor).' These data are provided by the World Health Organization [2]. This dataset contains information about the consumption of tobacco products, including cigarettes, pipes, cigars, cigarillos, waterpipes (hookah, shisha), bidis, kretek, heated tobacco products, and all forms of smokeless (oral and nasal) tobacco. Tobacco products exclude items that do not contain tobacco, such as electronic nicotine delivery systems (ENDS), including e-cigarettes, 'e-cigars', 'e-hookahs', JUUL, and 'e-pipes'.The dataset provides information about the percentage of the population in each country that consumes tobacco products, as well as the percentage for each sex. These percentage values are derived from the population aged 15 years and over who currently use any tobacco product, whether smoked or smokeless. These values are estimated using a statistical model based on Bayesian negative binomial meta-regression [3]. The original dataset consists of 13,284 samples, each with 34 variables. However, not all variables were useful for the current analysis, and many variables were discarded.

III. Modeling the dataset

As a first step, the dataset is cleaned to remove NaN variables and all columns that do not provide significant information, such as indicators and data types (i.e., 'ValueType,' 'Location.Type,' etc.). The resulting dataset is composed of the following

- ParentLocation:
  Contains the name of the geographic area where the sample was collected. The geographic areas include South-East Asia, Europe, Africa, Eastern Mediterranean, Western Pacific, and the Americas.

- Location:
  Contains the name of the country where the samples were collected. There are 164 different countries within the dataset.

- Period:
  Contains the year when the sample was collected. The samples were collected in the 2000s, 2005s, 2010s, 2015s, 2018s, 2019s, 2020s, while 2023s and 2025s are estimates.

- Dim1:
  Contains the sex of the sample. The sex features can be male, female, or both sexes.

- Value:
  Contains the percentage of people that consume tobacco and tobacco products, along with confidence intervals for the sample.

After the feature selection step, the selected variables are renamed as follows: 'geo_region,' 'state,' 'year,' 'sex,' and 'value.' Subsequent operations are performed to handle additional or redundant information. One additional piece of information is the presence of the confidence interval within the 'value' variable, which is removed. Another additional

piece of information is the presence of samples collected in the years 2023 and 2025; these samples are removed because they are estimates and not actual data. The redundant information is the presence of three measurements for each state in each year, and these samples are then grouped into a single sample representing the mean between each state's measurements for each year.

IV. Exploratory analysis

This section describes the main findings obtained from the EDA by focusing on the distribution of consumers in various geographical areas and with respect to sex as well. The variation of values over the years is also observed.

A. Value distribution

At the beginning of the analysis, only samples with a sex value equal to 'both sexes' collected in the 2020s are examined to have a comprehensive look at the general frequency distribution of the latest values. The histogram presented in Figure 1 represents this distribution.

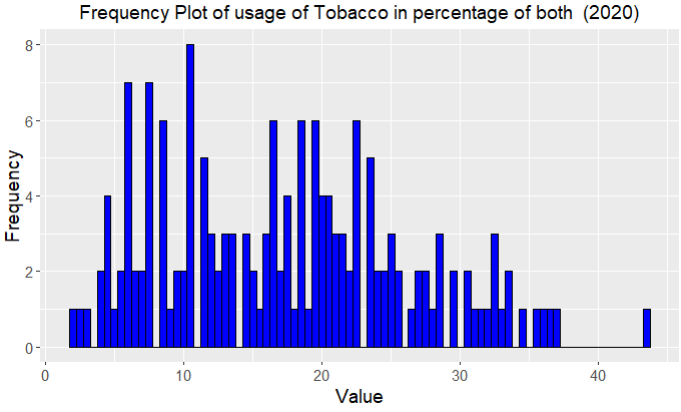

Fig. 1. Frequency plot of bothsex values from 2020.

Looking only at the distribution of values for both sexes, the values appear to be distributed as a normal distribution with a peak around 10 percent and a slightly higher concentration between 5 and 10 percent. This observation makes one wonder if this representation of normality is observed because the means of the male and female distributions are close to a normal distribution or if both of them adhere to normality.

From Figure 2, we can see that the density distribution of males seems to be close to normal, while female subjects do not appear to adhere to normality and exhibit a different shape. These density behaviors could indicate that the distribution of values for both sexes is influenced by the means of males and females. However, this should be statistically verified.

B. Distribution of values over geographic ares

In Figure 3, the distributions of the six geographic areas with respect to both sexes during 2020 are represented. All areas seem to exhibit a concentration of samples that differs among them, with the highest value belonging to the Western Pacific area, while South-East Asia and the Western Pacific

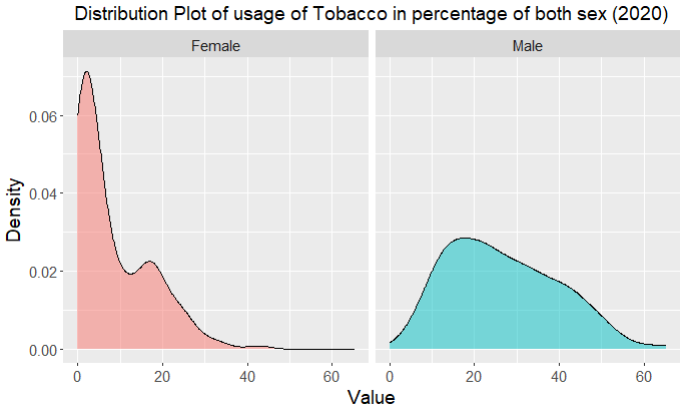

Fig. 2. Density plot of male and female values from 2020.

are the areas with a more uniform density of samples. For a deeper understanding, it would be important to show how the geographic area values differ when they are divided by sex.

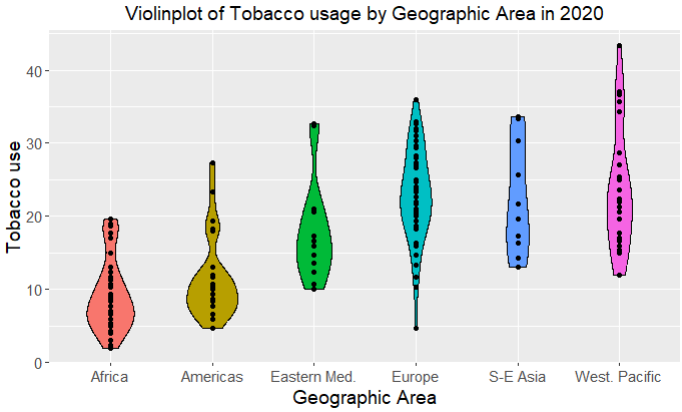

Fig. 3. violin plot of both sex values divided by geographic areas from 2020.

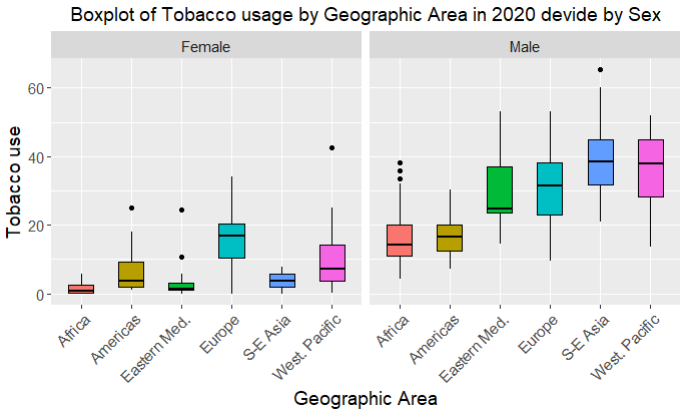

Fig. 4. Violin plot geographic area with male and female values from 2020.

Observing Figure 4, it is clear that Figure 3 conveys different information regarding the geographic area with the highest value. In fact, the geographic area with the highest

percentage of tobacco consumers is a state in South-East Asia, representing male consumers. This assumes that the mean of South-East Asia and the Western Pacific area in male observations is close to each other. When looking at the distribution of females across areas, it seems that the highest value is present in Europe or the Western Pacific area. Further investigation is needed to understand if the highest sample in the Western Pacific area can be seen as an outlier or a valid measurement.

C. Mean value over geographic areas

It is interesting to understand how consumers, on average, are spread across the geographic areas, and Figure 5 shows this information with respect to 2020. The means seem to be distributed in similar values in Europe, South-East Asia, and the Western Pacific areas, while lower mean values are present in Africa, the Americas, and the Eastern Mediterranean, with different mean values among them. This information should be statistically verified.

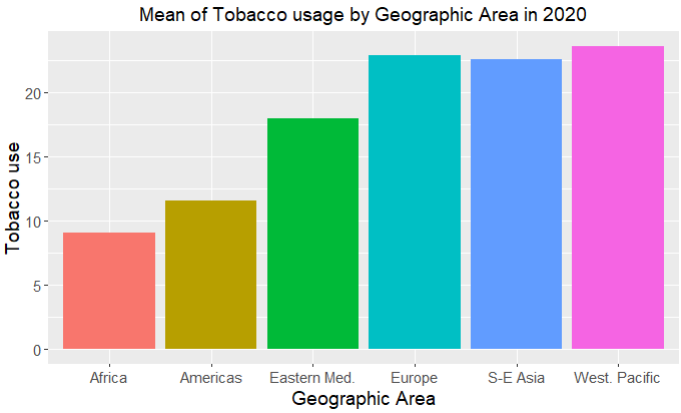

Fig. 5. Histogram plot of both sex values respect to geographic areas in 2020.

In Figure 6, it is shown how the means differ in geographic areas with respect to the means of both sexes in those geographic areas. On the female side, the figure suggests that European females are, on average, major tobacco consumers compared to other areas, while South-East Asia males are the highest consumers on average. It could be interesting to verify these hypotheses using a Friedman test to determine if the groups/geographic areas differ in mean among themselves. It is worth noting that the normality of the subjects/countries has already been rejected in the statistical tests conducted in section five.

D. Mean value over years by geographic areas

Up to this point, the visual information has shown how the data behaved in 2020. In this section, we visualize how the mean values have changed over the years. Figure 7 shows how the mean values of each geographic area changed between 2000 and 2020 for both sexes.

It appears that all areas have a decreasing trend in mean, with a significant change in the Americas and a less pronounced decrease in the Eastern Mediterranean compared to

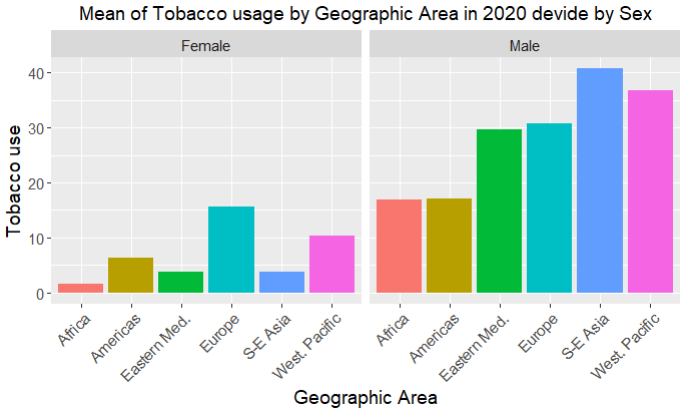

Fig. 6. Histogram plot of male and female values respect to geographic areas in 2020.

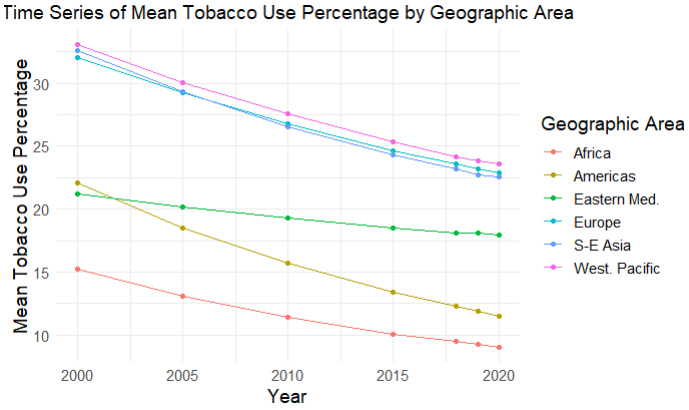

Fig. 7. Time series plot of both sex mean values from 2000 to 2020.

other areas. It is interesting to observe how these general trends are influenced by the male and female trends.

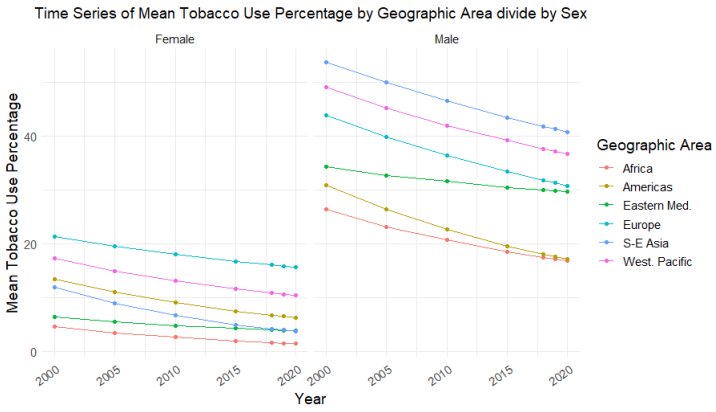

Fig. 8. Time series plot of female and male mean values from 2000 to 2020.

In Figure 8, the trend in the mean values for females and males between 2000 and 2020 is shown. From the male data perspective, it appears that there is a decreasing trend in all geographic areas, with a more significant decrease in

the Americas and a less significant decrease in the Eastern Mediterranean. In contrast, from the female data perspective, there seems to be a general decreasing trend that is lower than the male one, with South-East Asia appear to experience the most significant decrease.

E. Recap of Exploratory Data Analysis

The visual prospective suggest to us that the female and male distribution in 2020 is drastically different and this result influence the mean value between them. This suggestion tell us that look to the data using the percentage from both sexes doesn't give to us the ideal behavior of the data. A similar suggestion is when it's focus the distribution on male and female respect to geographic areas, in all areas the female data have a lower value interval respect to the male one and female have also a lower mean value. This observations in 2020 are not extended to previous years, but focusing on the trend over the years the visualizations suggest that the general trend is influence in the past years too from the mean of both sexes, because the behaviour of the two sexes are different during these periods. Starting from the conclusions drawn from the analysis presented in this study, it was decided to observe whether the averages of male data and female data really changed during the decade 2010 to 2020.

V. Statistical tests

This section describes how the null hypothesis "the mean of males/females using tobacco is the same over the last five recorded years" is analyzed with the aim to accept or reject the null hypothesis. First, the normality of the male and female samples from the last five years is analyzed, and it is checked if normality, homoscedasticity, and sphericity are satisfied. Then, Friedman [4] and repeated measures Anova [5] tests are applied to the hypothesis. Both tests are conducted on male subgroup data, while only the Friedman test is applied to female subgroup data.

A. Normality test and Anova requirements

Before the application of the Normality test, quantile plots from each group are visualized for a preliminary observation of the behavior of data. Quantile plots are similar in this case if they belong to the same subgroup; for this reason, in this essay, only one representative plot is shown for each subgroup. The quantile plot tell us:

- Quantile plots for the male subgroup shown in Figure 9 indicate the possible presence of normality condition in all years because they seem to have samples that follow the line pattern. A Shapiro-Wilk test is needed to check this assumption.
- Quantile plots for the female subgroup shown in Figure 10 indicate that the normality condition is not satisfied in all years because they seem to have samples that do not follow the line pattern. A Shapiro-Wilk test is needed to check this assumption.

The Shapiro-Wilk test is used to check normality in male and female samples. From Table 1, the result indicates that the

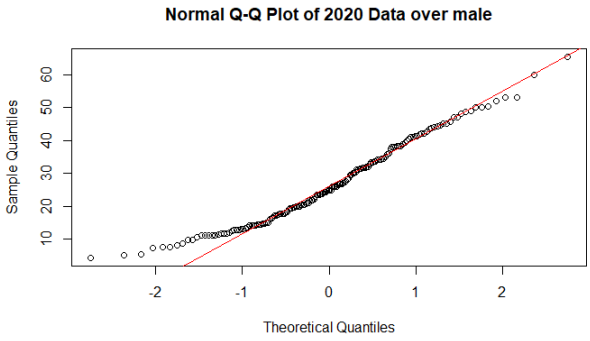

Fig. 9. Quantile plot of male in 2020

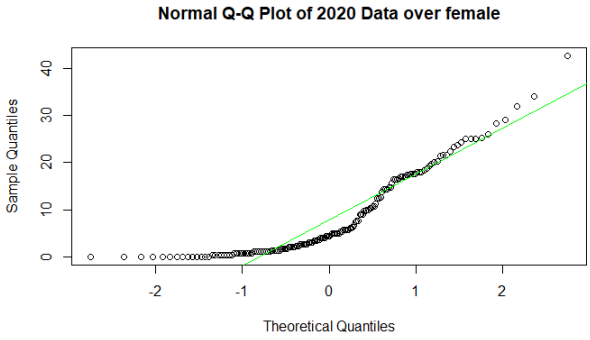

Fig. 10. Quantile plot of female in 2020

TABLE I
SHAPIRO-WILK FOR MALE IN EACH YEAR

Year	W statistics	p-value
2010	0.98049	0.02058
2015	0.97356	0.003125
2018	0.96975	0.00118
2019	0.96925	0.001042
2020	0.96742	0.0006639

TABLE II
SHAPIRO-WILK FOR FEMALE IN EACH YEAR

Year	W statistics	p-value
2010	0.86804	$< 8.0 \times 10^{-11}$
2015	0.85454	$< 1.9 \times 10^{-11}$
2018	0.84469	$< 6.6 \times 10^{-12}$
2019	0.84195	$< 5.0 \times 10^{-12}$
2020	0.83917	$< 3.8 \times 10^{-12}$

null hypothesis is rejected for all five groups with a p-value lower than 0.05. The same outcome can be observed in Table 2 from the female groups, with a stronger rejection compared to the male group because the p-value of each group is strictly lower than the 0.05 threshold. In conclusion, Normality is rejected in both macro groups. In Figure 2, the female density distribution has a shape completely different from a normal distribution, and Figure 10 shows a different distribution of data points over the theoretic quantiles, so it is reasonable that normality is rejected in the female group, but the situation is somewhat different for males. The male density distribution has a shape close to a normal distribution, and the quantiles plot in Figure 9 follows a pattern close to the normal pattern of the theoretic distribution.

TABLE III
BARTLETT TEST OF HOMOGENEITY OF VARIANCES

Sex	Degree of freedom	K-square statistics	p-value
male	4	0.055455	0.9996
female	4	3.431	0.4885

TABLE IV
SPHERICITY TEST

Sex	Epsilon
male	0.9860699
female	0.9866401

Table 3 shows the result of the Bartlett test to check the presence of homoscedasticity between groups. In each group, the null hypothesis is accepted because the p-value is greater than the threshold of 0.05. While the sphericity test results in Table 4 display the epsilon values for both macro groups, and the results indicate that in both cases, sphericity doesn't significantly affect the results because they have a low deviation. Epsilon values are not significantly distant from 1.0. Based on these results, the main hypothesis test is structured using the non-parametric Friedman test in both sexes, but with an additional approach in the male groups. The male group, as indicated above, has a quantile and density plot that suggest possible normality. Considering that there are enough samples to reflect the applicability of the law of large numbers and with the added presence of homoscedasticity and sphericity, it may be reasonable to apply the main hypothesis Anova test as well and compare the results of both tests to gain a broader perspective on the case.

B. Hypothesis Test

In this sub-section, we observe the results of ANOVA repeated measure test and Friedman Test for male groups and only the Friedman test for female groups. In case of the rejection of the null hypothesis, a post-hoc analysis using a pairwise test with the Bonferroni correction and the Benjamini-Hochberg correction is applied

In Table 5, the results of the ANOVA, F-subjects, and Friedman tests are presented. For both sexes, the null hypothesis

TABLE V
HYPOTHESIS TEST RESULTS IN BOTH GROUPS

Sex	Test	p-value
male	Anova	$< 2 \times 10^{-16}$
male	F_subjects	0
male	Friedman	$< 2.2 \times 10^{-16}$
female	Friedman	$< 2.2 \times 10^{-16}$

is rejected because the p-value for each sex is lower than the threshold. It is possible to observe, regarding male samples, that the ANOVA and Friedman tests yield similar p-values, indicating closely matched results. This condition implies that using ANOVA as the statistical test for male samples is a suitable choice and allows for the application of another test, the F-subjects test. The null hypothesis for the F-subjects test is that "the means in the different subjects are the same." The result for F-subjects is 0, which is logical due to the following reasons:

- a large number of degree of freedom, 163
- a number of sample equal to 820 observation
- the variability between subject of 825.3
- a sum of residual square of 1.9

These parameters generate a cumulative distribution function of the F-distribution that has the F-sub value concentrated around 1. Considering that the F-sub p-value is calculated as 1 minus the cumulative distribution of F, it is reasonable to obtain this result.

TABLE VI
MANN-WHITNEY WITH BONFERRONI CORRECTION OF MALE GROUPS

Years 1	Year 2	p-value
2015	**2010**	$< 2 \times 10^{-16}$
2018	**2010**	$< 2 \times 10^{-16}$
2018	**2015**	$< 2 \times 10^{-16}$
2019	**2010**	$< 2 \times 10^{-16}$
2019	**2015**	$< 2 \times 10^{-16}$
2019	**2018**	$< 2.8 \times 10^{-15}$
2020	**2010**	$< 2 \times 10^{-16}$
2020	**2015**	$< 2 \times 10^{-16}$
2020	**2018**	$< 2 \times 10^{-16}$
2020	**2019**	$< 2 \times 10^{-16}$

TABLE VII
MANN-WHITNEY WITH BENJAMINI-HOCHBERG CORRECTION OF MALE GROUPS

Years 1	Year 2	p-value
2015	**2010**	$< 2 \times 10^{-16}$
2018	**2010**	$< 2 \times 10^{-16}$
2018	**2015**	$< 2 \times 10^{-16}$
2019	**2010**	$< 2 \times 10^{-16}$
2019	**2015**	$< 2 \times 10^{-16}$
2019	**2018**	$< 2.8 \times 10^{-15}$
2020	**2010**	$< 2 \times 10^{-16}$
2020	**2015**	$< 2 \times 10^{-16}$
2020	**2018**	$< 2 \times 10^{-16}$
2020	**2019**	$< 2 \times 10^{-16}$

TABLE VIII
MANN-WHITNEY WITH BONFERRONI CORRECTION OF FEMALE GROUPS

Years 1	Year 2	p-value
2015	**2010**	$< 2 \times 10^{-16}$
2018	**2010**	$< 2 \times 10^{-16}$
2018	**2015**	$< 1.1 \times 10^{-15}$
2019	**2010**	$< 2 \times 10^{-16}$
2019	**2015**	$< 2 \times 10^{-16}$
2019	**2018**	$< 1.9 \times 10^{-10}$
2020	**2010**	$< 2 \times 10^{-16}$
2020	**2015**	$< 2 \times 10^{-16}$
2020	**2018**	$< 6.1 \times 10^{-15}$
2020	**2019**	$< 1.2 \times 10^{-8}$

TABLE IX
MANN-WHITNEY WITH BENJAMINI-HOCHBERG CORRECTION OF FEMALE GROUPS

Years 1	Year 2	p-value
2015	**2010**	$< 2 \times 10^{-16}$
2018	**2010**	$< 2 \times 10^{-16}$
2018	**2015**	$< 2 \times 10^{-16}$
2019	**2010**	$< 2 \times 10^{-16}$
2019	**2015**	$< 2 \times 10^{-16}$
2019	**2018**	$< 2.2 \times 10^{-11}$
2020	**2010**	$< 2 \times 10^{-16}$
2020	**2015**	$< 2 \times 10^{-16}$
2020	**2018**	$< 7.6 \times 10^{-16}$
2020	**2019**	$< 1.2 \times 10^{-9}$

After analyzing the null hypothesis regarding the means of the groups and calculating the F-sub test, it is interesting to perform a post-hoc test to determine how the means of each group differ when compared in pairs. The post-hoc test applied to all pairs of years is the non-parametric Mann-Whitney test. Initially, the Bonferroni correction was applied, and the results can be seen in Table 6 and Table 8. Subsequently, the Benjamini-Hochberg correction was used, and the corresponding results are presented in Table 7 and Table 9. The use of these corrections was necessary to prevent the probability of a type I error from becoming excessively high and unduly influencing the results. Both corrections were employed to identify any potential inconsistencies between them. Table 6 and Table 7 display the results of the Mann-Whitney test with the two different corrections. The findings indicate that there are significant differences between all years, and this result is also evident in Table 8 and Table 9 when comparing the female groups.

VI. CONCLUSION

The results of this paper indicate that males are higher consumers of tobacco products compared to females, with a higher mean percentage in all geographic areas. While males have a wider density compared to females, females present a higher concentration of values around 5 percent. These insights suggest that the next awareness campaign should target men rather than women. When looking at the frequency distribution of each gender in relation to geographic areas, it is observable that the Eastern Mediterranean, South-East Asia, Europe, and the Western Pacific have a higher distribution than the Africa and Americas areas with respect to male samples. However, in the female distribution, only Europe has a mean of over 10 percent of tobacco consumers. These findings can be used to suggest a higher investment in awareness campaigns in the areas with the highest consumption, with a campaign focus on women as well in Europe. Overall, it is confirmed that the mean consumption of tobacco products in the last 10 years is decreasing in all studied areas, with a more significant decline in males than females. Females appear to have a more linear behavior compared to males. In further studies, it could be interesting to focus on countries within the same area to understand if all countries have similar means or if there are significant differences between them. Additional information could be collected regarding age ranges, where there should be several intervals indicating the percentage of total smokers within each observed range. This information could lead to the discovery of further insights about tobacco consumption in relation to the age of the consumers.

REFERENCES

[1] "Tobacco", who, https://www.who.int/news-room/fact-sheets/detail/tobacco (accessed Sept. 20, 2023)

[2] "Non-age-standardized estimates of current tobacco use, tobacco smoking and cigarette smoking (Tobacco control: Monitor)", who, https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-tobacco-control-monitor-current-tobaccouse-tobaccosmoking-cigarrettesmoking-nonagestd-tobnonagestdcurr (accessed Sept. 8, 2023)

[3] Ver Bilano, Stuart Gilmour, Trevor Moffi et, Edouard Tursan d'Espaignet, Gretchen A Stevens, Alison Commar, Frank Tuyl, Irene Hudson, Kenji Shibuya , "Global trends and projections for tobacco use, 1990–2025 an analysis of smoking indicators from the WHO Comprehensive Information Systems for Tobacco Control", Lancet 2015; 385: 966–76.

[4] Sheldon, M.R., Fillyaw, M.J. and Thompson, W.D. (1996), "The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs", Physiotherapy Research International, 1996; 1: 221-228.

[5] Eunsik Park, Meehye Cho, Chang-Seok Ki, "Correct Use of Repeated Measures Analysis of Variance", Korean J Lab Med 2009;29:1-9.