Has the Number of Smokers in England Decreased in the Last 20 Years?

Kreativstorm Data Analysis Course

Group: B.2

Full Names (alphabetical order):
Antonios Mitsis
Everalda Solange
Gayathri Anantha
Valeria Adlard (Tokareva)

Emails:
anmitsis@hotmail.com
everaldasolange@gmail.com
angayathri@gmail.com
neitling.corp@gmail.com

Tutor: Renato Pavleković

The research topic of "has the number of smokers in England decreased in the last 20 years" will be of interest to various groups due to its potential implications for public health, policy-making, and understanding societal trends. Some of the main groups interested in this research are listed below:

- **Government and Policy Makers**
  Governments and policy makers are interested in promoting the health and well-being of their citizens. A decline in smoking rates could lead to fewer healthcare costs associated with smoking-related illnesses.
- **Public Health Organizations**
  Public health agencies and organizations are concerned with the well-being of the population. A decrease in smoking rates would be of interest to them as it could indicate progress in reducing a major risk factor for various diseases like lung cancer, heart disease, and respiratory conditions. It might inform them about the effectiveness of their anti-smoking campaigns and interventions.
- **Tobacco Industries**
  Tobacco industries might have a different interest on the topic as a decrease in the smokers of England could impact their market and profits. It's important to recognize that the interests of tobacco industries might not always align with those of public health and other organizations.
- **Economic Analysts**
  Economic analysts might be interested in the research due to its potential economic impact.
- **General Public**
  The general public is always interested in matters related to health and well-being. Understanding whether smoking rates have decreased could impact individual behaviors and attitudes towards smoking. It could also shape perceptions of the effectiveness of public health campaigns.

These are some, but not all, groups that find interest in the research topic of a decreasing ration of smokers.

Searching through the case study folders we find interest in many different files. Firstly, we have the two datasets, for the 1999 and 2019 researches by the “HSE_1999” and “HSE_2019” that are in. dta format. Then we have the two corresponding dataset documentation in .pdf format, that provide useful information about the variables of the dataset as well as general information about the researches. The datasets contain information about various topics such as drinking, general health etc. but we are only interested in the smoking part. There are also two word files that contain all of the variables, their type and the meaning of the values they take.

In general, we will only be needing the positive values of the variables as the negatives represent different reasons for unanswered questions, so we will be treating all of them as NA’s. Also, since the datasets are quite large and contain more than 1,800 variables and 10,000 observations it is important to remove the columns that are not about smoking so we can search through the dataset faster and more efficiently. This is not that important however, since we have the dataset documentation that helps us navigate without getting lost.

The idea is to select some variables that contain information about the smoking status of the population from both time periods and compare the smoking ratios, trying to find a significance difference in the number of smokers. The variables we have chosen to use are “cignow” from the 1999 dataset and “cignow_19” from the 2019 dataset. It important to select variables that represent the same thing in order to compare their values. The variables mentioned before will be listed from the data dictionaries that are provided, mentioning their type and levels.

Pos. = 1642 **Variable =** cignow **Variable label =** Whether smoke cigarettes nowadays
This variable is *numeric*, the SPSS measurement level is *nominal.*
SPSS user missing values = -99 thru -1

Value label information for cignow
Value = -9 Label = No answer/refused
Value = -8 Label = Don't know
Value = -6 Label = Schedule not obtained
Value = -2 Label = Schedule not applicable
Value = -1 Label = Item not applicable
Value = 1 Label = Yes
Value = 2 Label = No

Pos. = 921 **Variable =** cignow_19 **Variable label =** Whether smoke cigarettes nowadays (capi+casi)
This variable is *numeric*, the SPSS measurement level is *NOMINAL*
SPSS user missing values = -1.7976931348623155e+308 thru -1.0

Value label information for cignow_19
Value = 1.0 Label = Yes
Value = 2.0 Label = No
Value = -9.0 Label = Refused
Value = -8.0 Label = Don't know
Value = -1.0 Label = Not applicable

After inspecting the variables, we notice that we only deal with categorical variables, therefore the only statistical test we can perform is the chi-square test for proportions. This test will provide useful information about the change in proportions of smokers to non-smokers. Our goal is to determine whether there is a strong relationship between smokers and time period and if there is we can compare the actual values of smokers in the two datasets to decide if there is a decrease, increase or no change in the number of smokers.

There are four assumptions for the chi-square test, including:

1. **Both variables are categorical:** This is true, all variables are categorical with two levels, smoker and non-smoker. (We have removed all the non-positive values considering them NA's)
2. **Independent Observations:** Since our data are taken from a survey, we can assume that each observation is independent and randomly sampled.
3. **Cells in the contingency table are mutually exclusive:** Each individual contributes once to the data as it is assumed that the research was conducted properly.
4. **Expected values of cells in contingency table should be greater than 5:** We can first create the contingency table that we will use in the chi-square test.

Contingency Table	Old	New	Sum	
Smokers	1519	1254	2773	
Non-Smokers	1344	3063	4407	
Sum	2863	4317	7180	**Total**

To calculate the expected values, we simply multiply the row sum times the column sum divided by the total for each cell.

Expected Values	Old	New
Smokers	1,105.72	1,667.28
Non-Smokers	1,757.28	2,649.72

Since all of the assumptions are covered, The results from the test can be trusted and are reliable. The null hypothesis of our chi-square test is that the proportion of smokers to non-smokers in the 1999 dataset is the same as in that of the 2019 dataset. The alternative hypothesis is that these proportions are significantly different, which means that the time period of 20 years has had a significant effect on the number of smokers, increased or decreased.

After performing the chi-square test there is significant association in smoking status and time period, $\chi^2(1) = 417.55$ and p-value = $2.2 \cdot 10^{-16} < 0.05$. This seems to represent the fact that the ratio of smokers to non-smokers in 1999 is 1.8 times higher than that of the 2019 dataset. Specifically, 53.05% of the population sample were smokers in 1999, while only 29.04% were smokers in the 2019 dataset. This along with the statistically significant difference in proportions from the chi-square test is enough evidence to suggest that the number of smokers in England has decreased over the past 20 years.

Taking these results into consideration, we can state that the influence of various health organizations on the population has increased and the policy in the field of public health is being carried out successfully. Nevertheless, this probably led to some transformations in the policy of tobacco companies and changes in economic strategies. All in all, this trend showcases the potential for collective efforts to drive positive changes in society's overall well-being.