**Daily Cigarette Consumption Analysis**

**1. Project Overview**

In this project, the variable cigs\_per\_day (number of cigarettes smoked per day) was selected from a health dataset available on Kaggle. The aim is to analyze this variable using basic statistical methods. The dataset includes information about individuals’ health and lifestyle. The chosen variable is suitable for performing statistical analysis on a single numeric column.

**2. Project Goals**

* To calculate descriptive statistics such as mean, median, variance, standard deviation, and standard error
* To visualize the data using a histogram and a boxplot to understand distribution and detect outliers
* To construct 95% confidence intervals for the mean and variance
* To perform a hypothesis test about the population mean
* To calculate the minimum required sample size with a ±0.1 margin of error at a 90% confidence level

**3. Data Description**

The dataset used in this project was taken from Kaggle. It includes health-related information about individuals, such as age, gender, heart rate, blood pressure, and cholesterol levels. The selected numerical variable is cigs\_per\_day (number of cigarettes smoked per day).

The cigs\_per\_day column shows how many cigarettes a person smokes in a day. This variable was chosen because it directly provides information about smoking habits and allows us to examine people's daily habits using numerical data.

**4. Descriptive Statistics**

Some basic statistical calculations were made for the cigs\_per\_day variable:

* Mean: 9.17
* Median: 0.00
* Variance: 144.91
* Standard Deviation: 12.01
* Standard Error: 0.1931

The mean and median are very close to each other, which shows that the data is generally balanced. The variance and standard deviation indicate that the number **of** cigarettes people smoke per day does not vary much and the values are close to each other.

**5. Data Visualization**

**Histogram Interpretation:**According to the histogram (Figure 1), most people smoke between 0 and 5 cigarettes per day.  
The distribution appears right-skewed because a few people smoke 40 or more cigarettes, which pulls the data to the right.  
This shows that low cigarette consumption is common, but there are also some individuals who smoke heavily.

**Boxplot Interpretation:**In the boxplot (Figure 2), the median value is around 2 to 4 cigarettes.  
Some people smoke 50, 60, or even 70 cigarettes a day, and these are considered outliers since they fall outside the normal range.  
This plot shows both the average smoking level and the extreme outlier values.

A graph of cigarettes in a number of cigarettes per day

Description automatically generatedA graph with a blue rectangle

Description automatically generated

*(Figure 1)*   *(Figure 2)*

**6. Confidence Intervals**

In this section, 95% confidence intervals were calculated for both the mean and the variance of daily cigarette consumption.

According to the results:

* The mean number of cigarettes smoked per day is estimated to be between 8.79 and 9.55 with 95% confidence.
* The variance is estimated to lie between 138.68 and 151.58.

These intervals suggest that the true population values are likely to fall within these ranges. Since the sample size is large (n > 30), the use of normal and chi-square distribution assumptions is considered appropriate.

**7. Sample Size Estimation**

In this section, we calculated the minimum sample size required to estimate the population mean with a 90% confidence level and a margin of error of ±0.1.

Accurate estimation depends on having enough data. The following values were used in the calculation:

* Confidence level: 90%
* Z-score: 1.645 (from standard normal distribution)
* Margin of error (E): ±0.1
* Standard deviation (from our dataset): 12.01

Based on this, the required sample size was found to be at least 393 individuals.

Since our dataset includes over 3000 people, we can confidently say that our sample size is more than sufficient for this estimation.

**8. Hypothesis Testing**

In this section, a hypothesis test was conducted to examine whether the average number of cigarettes smoked per day is equal to 10. The goal is to determine if the actual smoking behavior in the dataset differs significantly from this assumed value.

**Hypotheses:**

* H₀ (null hypothesis): μ = 10
* H₁ (alternative hypothesis): μ ≠ 10

Why Z-test Was Used?

Since the sample size is large (n > 30) and the standard deviation was calculated from the sample, a z-test is appropriate.

**Test Results:**

The test resulted in a z-score of -4.31 and a p-value of 0.0000.

**Decision and Interpretation:**

Because the p-value is less than 0.05, we reject the null hypothesis. This means the average daily cigarette consumption in the dataset is significantly different from 10. In fact, the observed average is about 9.17, which is statistically lower than 10.

**9. Conclusion**

Based on the statistical analyses conducted, the average daily cigarette consumption was calculated as 9.17 cigarettes per day. The median value was found to be 0, indicating that a large portion of individuals do not smoke at all, and many others consume relatively few cigarettes.

The variance was calculated as 144.91, and the standard deviation was 12.04, which shows a wide variation in smoking habits among individuals. The standard error was found to be 0.1931, suggesting a reliable estimation of the mean.

According to the graphs, the histogram shows a right-skewed distribution. Most people smoke between 0 and 5 cigarettes, but some individuals consume 40 or more, creating outliers that stretch the right tail. The boxplot also highlights 50, 60, and 70 cigarettes per day as outlier values.

The hypothesis test confirmed that the population mean is significantly different from 10. This supports the conclusion that most individuals smoke fewer than 10 cigarettes per day, although the presence of heavy smokers increases the overall average.

**10. Code**

GitHub= <https://github.com/talhayilmaztr/Data/blob/main/data.py>

**11. References**

* **Dataset Source:**  
  Kaggle - <https://www.kaggle.com/datasets/jaceprater/smokers-health-data>
* **Textbooks / Course Materials:**
  + Lecture slides shared on the LMS platform (e.g., Lecture\_07\_lms.pdf, MAT19234E.pdf)
* **Python Libraries Used:**
  + pandas for data loading and preprocessing
  + math and scipy.stats for statistical calculations (mean, confidence intervals, z-tests, etc.)
  + matplotlib.pyplot for data visualization (histograms and boxplots)