## Project Description

The goal of this project was to analyze data from questionnaires presented in the Excel table "Data\_Excel," construct relevant histograms, dependency graphs for certain pairs of variables, and check the statistical significance of the impact of these parameters.

## Mathematical Methods

The following mathematical methods were used to assess the impact of the parameters and verify their statistical significance:

- \*\*Pearson Correlation:\*\* Used to assess the linear relationship between two variables.

- \*\*One-Sample t-test:\*\* Used to check if the sample mean equals a specified value (in our case, 0).

The Pearson correlation coefficient (r) is a number between -1 and 1 that shows the degree of linear relationship between two variables, measuring the strength and direction of this relationship. Examples of how to interpret the result based on the values of r are given below.

### Positive Linear Relationship (0 <= r <= 1)

- \*\*r ≈ 1:\*\* Very strong positive linear relationship.

- This means that when the value of one variable increases, the value of the other variable also increases.

- Example: When the temperature rises, the number of drinks sold also increases.

- \*\*r ≈ 0.7:\*\* Strong positive linear relationship.

- A significant portion of the variation in one variable is explained by the variation in the other variable.

- \*\*r ≈ 0.5:\*\* Moderate positive linear relationship.

- There is a positive relationship, but it is not very strong.

- \*\*r ≈ 0.3:\*\* Weak positive linear relationship.

- There is a small positive relationship between the variables.

- \*\*r ≈ 0:\*\* No linear relationship.

- The variables do not have a linear relationship. There may be other types of relationships (e.g., nonlinear).

### Negative Linear Relationship (-1 <= r <= 0)

- \*\*r ≈ -1:\*\* Very strong negative linear relationship.

- This means that when the value of one variable increases, the value of the other variable decreases.

- Example: When the number of workouts increases, body weight decreases.

- \*\*r ≈ -0.7:\*\* Strong negative linear relationship.

- A significant portion of the variation in one variable is explained by the variation in the other variable.

- \*\*r ≈ -0.5:\*\* Moderate negative linear relationship.

- There is a negative relationship, but it is not very strong.

- \*\*r ≈ -0.3:\*\* Weak negative linear relationship.

- There is a small negative relationship between the variables.

- \*\*r ≈ 0:\*\* No linear relationship.

- The variables do not have a linear relationship. There may be other types of relationships (e.g., nonlinear).

## P-Value (p-value)

The p-value is an important indicator in statistical tests that helps us assess the strength of the evidence against the null hypothesis (H0). The null hypothesis typically states that there is no effect or difference between groups.

### Interpretation of the p-value

- \*\*Large p-value (e.g., > 0.05):\*\* Indicates weak evidence against the null hypothesis, so we cannot reject the null hypothesis. This means that the observed effect is not statistically significant.

- \*\*Small p-value (e.g., <= 0.05):\*\* Indicates strong evidence against the null hypothesis, so we reject the null hypothesis. This means that the observed effect is statistically significant.

### Significance Threshold (α)

The significance threshold (α) is the value against which we compare the p-value to make a decision about the null hypothesis. Commonly used threshold values are:

- \*\*α = 0.05:\*\* Standard significance threshold in many scientific studies.

- \*\*α = 0.01:\*\* More stringent significance threshold.

- \*\*α = 0.10:\*\* Less stringent significance threshold.

### How the p-value Affects Decisions

1. \*\*If p-value <= α:\*\*

- Reject the null hypothesis. The observed data is statistically significant, indicating an effect or difference between groups.

2. \*\*If p-value > α:\*\*

- Cannot reject the null hypothesis. The observed data is not statistically significant, and we cannot conclude that there is an effect or difference.

In our project, the p-value is calculated for each pair of variables to assess the significance of the correlation between them. If the p-value is less than the chosen significance threshold (e.g., 0.05), we can consider that there is a significant correlation between the variables.

### t-test

The t-test is a statistical test used to compare means and determine if there are statistically significant differences between them. There are several types of t-tests, and a one-sample t-test was used in this project.

### One-Sample t-test

The one-sample t-test is used to test whether the mean of a sample is significantly different from a known or hypothesized population mean.

### Interpretation of Results

After calculating the t-statistic, we compare the obtained value with the critical t value (how to determine the critical value is mentioned in the "Determining the Critical t Value" section) for the corresponding significance level (α) and degrees of freedom (df = n - 1, where n is the sample size; more details in the "Determining the Critical t Value" section). If the absolute value of the t-statistic is greater than or equal to the critical t value, the null hypothesis is rejected.

We can also use the p-value returned by the t-test function to make a decision.

- \*\*If p-value <= α:\*\* Reject the null hypothesis. This means that the sample mean is significantly different from the hypothesized population mean.

- \*\*If p-value > α:\*\* Cannot reject the null hypothesis. This means that there is not enough evidence to claim that the sample mean is different from the hypothesized population mean.

## Critical t Value

The critical t value is the value that defines the threshold beyond which the result of a t-test is considered statistically significant.

### Determining the Critical t Value

The critical t value depends on:

1. \*\*Significance Level (α):\*\* This is the probability of rejecting the null hypothesis when it is true. Common values used are α = 0.05, 0.01, or 0.10.

2. \*\*Degrees of Freedom (df):\*\* This is the number of independent values in the sample that can vary. For a one-sample t-test, df = n - 1, where n is the sample size.

With these values, the critical t value is determined using the Student's t-distribution table.

In our project, the t-test is used to check whether the average difference between two variables significantly differs from zero.

## Program Description

The program consists of five main functions:

1. \*\*data\_preparation():\*\* Prepares data from the Excel file.

2. \*\*histogram\_creation():\*\* Creates histograms for each parameter in each sheet of the table.

3. \*\*get\_arrays\_of\_data():\*\* Reads data from the Excel file and converts it into Python arrays.

4. \*\*construction\_of\_graphs():\*\* Creates dependency graphs for certain pairs of variables from different sheets of the table.

5. \*\*correlation\_calculation():\*\* Checks the statistical significance of the impact of parameters using Pearson correlation and the t-test.

### data\_preparation()

This function performs the following actions:

- Reads data from the original Excel file Data\_Excel.xlsx.

- Replaces certain text values in the cells with numerical values according to the dictionary values\_to\_replace.

- If there is an empty cell in a non-empty column, it fills it with "I can't answer".

- Saves the updated Excel file as Data\_Excel(1).xlsx.

### histogram\_creation()

This function performs the following actions:

- Reads data from the prepared Excel file Data\_Excel(1).xlsx.

- Creates histograms for each parameter in each sheet of the table.

- Saves the histograms as images in the histograms folder.

- Uses a 2-second pause between constructing each histogram to avoid blocking.

### get\_arrays\_of\_data()

This function performs the following actions:

- Reads data from the prepared Excel file Data\_Excel(1).xlsx.

- Creates a dictionary all\_arrays to store data arrays from each sheet.

- Creates a dictionary all\_num\_of\_cols\_before\_empty to store the number of columns before the first empty column from each sheet (this will be used in the construction\_of\_graphs() and correlation\_calculation() functions).

- Iterates over each sheet in the file and stores the values of non-empty columns in arrays.

- Returns the result as a dictionary containing data arrays and the number of columns before the first empty column for each sheet.

### construction\_of\_graphs()

This function performs the following actions:

- Uses the results obtained by the get\_arrays\_of\_data() function.

- Uses the dictionary letters\_of\_columns to replace the headers of some columns with corresponding letters to avoid errors due to overly long file names.

- Creates dependency graphs for pairs of variables from the left and right columns from all sheets of the table (the difference between the left and right columns is determined using the all\_num\_of\_cols\_before\_empty array obtained by the get\_arrays\_of\_data() function). Depending on the number of repetitions, each point on the graph has the appropriate size, color, and label with the number of repetitions.

- Saves the graphs as images in the graphs folder.

- Uses a 3-second pause between constructing each graph to avoid blocking.