**Complete Statistical Correlation & Association Analysis of large phenotypic data and metadata**

**Shira Milo**

**2025**

**PART 1: Correlation & Association Analysis**

* 1. **Data Preparation**
* Data loading and identification: identifies categorical and numerical variables (categorical column names should include a “**CAT**” suffix).
* Reformats categorical variables to the appropriate format.
* Ensure data integrity: handles missing and infinite values.
  1. **Normality Testing**
* Numerical variables: Tests if follow a normal distribution.
* **Shapiro-Wilk** for small datasets and **Lilliefors** for larger ones.
* Saves distribution plots for visualization (in data\_dist ).
  1. **Correlation & Association Testing**
* **Numerical vs. Numerical:**
  + **Pearson correlation** (if both variables are normally distributed).
  + **Spearman correlation** (if at least one variable is not normally distributed).
* **Categorical vs. Numerical:**
  + **Mann-Whitney U test** (for binary categories).
  + **Kruskal-Wallis test** (for multi-category comparisons).
* **Categorical vs. Categorical:**
  + **Chi-square test** (for general associations).
  + **Fisher’s Exact Test** (if expected frequencies are low).
  1. **Multiple Testing Correction**
* Adjusts p-values using **False Discovery Rate (FDR)** to reduce false positives.
  1. **Visualization**
* Generates a clustered correlation heatmap for numerical comparisons.
* Generates individual boxplots for categorical vs. numerical associations.

**Output Files**

1. **Correlation results:** saved in corr\_res/correlation\_results\_with\_fdr.csv
2. **Normality test results:** saved in norm\_res/normality\_results.csv
3. **Correlation heatmap:** saved in corr\_plot folder
4. **Boxplots:** saved inboxplots folder

**PART 2: Post-hoc Analysis**

* 1. **Input & Setup**
* Loads significant correlation results from the primary analysis (corr\_res/correlation\_results\_with\_fdr.csv).
* Loads the original dataset to retrieve the categorical and numerical data.
  1. **Identification of Significant Comparisons**
* Case selection: iterates over significant associations (FDR-adjusted p-value < 0.05).
* Data evaluation: extracts independent (metadata) and dependant (phenotype) variables from each comparison.
* Data clean up: removes missing values.
  1. **Post-hoc Tests**
* **Pairwise Chi-square Test (Categorical vs. Categorical)**
  + Performed when a significant **Chi-square test** association was observed in the main analysis.
  + **Contingency tables** are created for each pair of categories.
  + **Pairwise Chi-square tests** determines which groups differ significantly.
* **Pairwise Mann-Whitney U Test (Binary Categorical vs. Numerical)**
  + Applied when the main test was **Mann-Whitney U**.
  + Compares two groups using a non-parametric rank-based test.
* **Dunn’s Test (Multi-group Categorical vs. Numerical)**
  + Performed when a **Kruskal-Wallis test** was significant.
  + Performs pairwise comparisons between all categories using Bonferroni correction.
  1. **Results and Outputs**
* Stores **all post-hoc results** in posthoc\_res/posthoc\_results.csv.

**PART 3: Post-hoc Interpretation**

* 1. **Input & Setup**
* Loads post-hoc test results from posthoc\_res/posthoc\_results.csv.
* Filters for significant (p < 0.05) post-hoc comparisons.
  1. **Processing Pairwise Comparisons**
* Extracts the two compared groups (e.g., "True vs. False").
* Identifies which group has higher or lower phenotype values.
* Determines whether a group is overrepresented or underrepresented in a phenotype.
  1. **Output**
* Outputs processed results in posthoc\_res/posthoc\_interpretation.csv.