□ What is Inferential Analysis and Why It Matters

When a lot of people respond to a survey, there are usually patterns hidden inside the answers. **Inferential analysis** is the name we give to the set of methods that look for those patterns, relationships, and trends so that we can make conclusions about groups of people—even if we didn't ask everyone.

Think of it like this:

- You have a big room full of people who've answered your questions.
- Some people answer similarly. Others don't.
- Some answers are connected to a person's age, gender, or other opinions.

This kind of analysis helps us find out things like:

- Do younger people think differently than older people?
- Are some questions strongly connected to others?
- Can we guess someone's age or gender based on how they answered?
- Can we group people into clusters based on their views?

This report does all of that—**automatically**—and saves the results for you to explore. The results can then be used in academic writing, especially for dissertations that aim to uncover social trends, compare demographics, or propose interventions.

□ What the Script Does – In Simple Language

Here's what the analysis script does step-by-step:

1. **Reads the Data**: It opens your database (`output.db`) and loads all the answers people gave in the survey.
2. **Cleans the Data**: It gets rid of messy or incomplete entries—for example:
    - Responses with missing answers
    - Duplicate IDs (if the same person submitted more than once)
    - Non-numeric answers where numbers were expected
3. **Performs Many Analyses**: It runs **10 types of statistical and machine learning techniques** to look at relationships, trends, and patterns between answers.
4. **Saves the Results in CSV Format**: Everything is exported into `.csv` files that can be opened in Excel or Google Sheets. These are grouped into a folder called `results/`.
5. **Each Method Has a Purpose**: Some methods find out what answers predict someone's age. Others group similar people together. Some look at relationships between answers.

These methods aren't random—they're based on well-established research principles in social science and data science. Each one contributes to building a deep understanding of your respondents.

□ What Are These Methods? (Complete Plain-Language Definitions)

1. **Chi-Square Test of Independence + Cramér's V**

- **What it is**: A way to check if two questions are statistically related.
- **Used for**: Categorical data—like Yes/No, Gender, etc.
- **Why it matters**: Helps you see if answers to two questions influence each other. For example, whether men and women respond differently to the same question.
- **Cramér's V**: Adds a score (between 0 and 1) showing the strength of the connection.

2. **Correlation Analysis**

- **What it is**: A way to check if numbers go up and down together.
- **Used for**: Numeric answers like age or Likert scale responses (e.g., 1 to 5).
- **Why it matters**: Tells you which responses tend to increase or decrease together. Great for identifying trends.

3. **Bootstrapped Age Differences**

- **What it is**: A modern statistical method to check if age differences between groups are real.
- **Used for**: Comparing average ages between people who answered Yes vs. No.
- **Why it matters**: Gives confidence that observed age differences aren't just luck or noise in the data.

4. **OLS Regression (Ordinary Least Squares)**

- **What it is**: A method that tries to predict a numeric outcome (like age) based on multiple answers.
- **Used for**: Finding out which answers predict someone's age.
- **Why it matters**: Lets you build profiles of different age groups based on their answers.

5. **Logistic Regression**

- **What it is**: A method that predicts the probability of a Yes/No outcome based on other answers.
- **Used for**: Predicting things like gender or whether someone is aware of an issue.
- **Why it matters**: Shows you what combinations of answers make certain responses more likely.

6. **Random Forest Models**

- **What it is**: An advanced machine learning method that ranks which answers are the most important in predicting outcomes.
- **Used for**: Both Yes/No questions and numeric ones (like age).
- **Why it matters**: Extremely powerful for identifying key variables. The "go-to" method when you want to know what matters most.

7. **PCA (Principal Components Analysis)**

- **What it is**: A method that simplifies lots of questions into a few broad themes.
- **Used for**: Reducing complex data sets to fewer dimensions.
- **Why it matters**: Helps you understand underlying structures—like whether your questions group into themes like "awareness" or "personal experience."

8. **Clustering (KMeans and Hierarchical)**

- **What it is**: A method that groups people based on how similarly they answered.
- **Used for**: Creating "types" or "profiles" of respondents.
- **Why it matters**: Crucial for audience segmentation. Helps you know who your respondents are—not just what they said.

Each method answers different research questions. Used together, they create a rich map of your respondents' minds, demographics, and shared behaviors.

□ File-by-File Explanation of Output Data

---

1. `age_by_category.csv`

What it contains: The average age of respondents grouped by different answers to categorical questions (e.g., gender, yes/no questions).

Why it matters: This helps you understand whether people of different ages answered differently. It's useful in identifying generational differences in attitude or awareness.

How to use it:

- In your dissertation: "The mean age of respondents who reported awareness of X was significantly lower than those who did not, suggesting a generational divide."

---

2. `bootstrap_age_diff_ci.csv`

What it contains: Confidence intervals for age differences between groups, generated using bootstrap simulations.

Why it matters: This tells you whether the age differences in `age_by_category.csv` are statistically meaningful.

How to use it:

- If the interval does not include 0, the age difference is likely real.
- "Bootstrapped confidence intervals show that age differences between respondents aware of X and those unaware are statistically significant."

---

3. `chi2_pairwise.csv`

What it contains: Results of Chi-Square tests between every pair of categorical questions.

Why it matters: Shows you which questions are statistically related.

How to use it:

- Use columns `p_value` and `significant` to know which relationships to write about.
- "A chi-square test revealed a significant relationship between gender and perceptions of victim credibility ($p < 0.05$)."

---

4. `cramers_v_matrix.csv`

What it contains: A table showing the strength of association between pairs of categorical questions.

Why it matters: Tells you how strongly two variables are related, even if the chi-square was significant.

How to use it:

- Values closer to 1 mean strong relationships.
- "Cramér's V score between gender and support for intervention policies was 0.42, suggesting a moderate relationship."

---

5. `ols_age_fullmodel.csv`

What it contains: Linear regression results showing which variables predict age.

Why it matters: Tells you which beliefs or responses are common in different age groups.

How to use it:

- "Regression results indicate that belief in stereotype X decreases with age ($\beta = -2.1$, $p < 0.01$)."

---

6. `logistic_coef_what_gender_do_you_identify_as.csv`

What it contains: Logistic regression showing what predicts someone identifying as a particular gender.

Why it matters: Can reveal biases or trends across genders.

How to use it:

- Focus on statistically significant coefficients.
- "Responses to questions A and B significantly predicted gender identity ($p < 0.05$)."

---

7. `logistic_coef_have_you_ever_heard_of_cases_where_men_were_sexually_abused.csv`

What it contains: Logistic regression predicting awareness of male sexual abuse cases.

Why it matters: Tells you what factors are associated with awareness.

How to use it:

- Helps support arguments for targeted awareness campaigns.

---

8. `rf_classifier_importances_what_gender_do_you_identify_as.csv`

What it contains: Variable importance rankings for predicting gender.

Why it matters: Shows you what questions are most useful for classifying respondents by gender.

How to use it:

- "Random Forest classifier revealed that responses to question X were the most influential in determining gender identity."

---

9. `rf_classifier_importances_have_you_ever_heard_of_cases_where_men_were_sexually_abused.csv`

What it contains: Same as above, but predicting awareness.

Why it matters: Points out what beliefs or experiences are most linked to awareness.

How to use it:

- Useful for building intervention strategies or awareness campaigns.

---

10. `rf_regressor_importances_age.csv`

What it contains: Importance scores for which answers best predict age.

Why it matters: Helps you understand what younger and older people think differently about.

How to use it:

- “The belief that X was a strong predictor of age, with younger respondents more likely to agree.”

---

11. `pca_components.csv`

What it contains: Principal components, which represent themes (e.g., attitudes or beliefs) in the data.

Why it matters: Simplifies complex data into underlying patterns.

How to use it:

- “The first component, representing traditional views, explained 24% of the variation in responses.”

---

12. `pca_loadings.csv`

What it contains: How much each question contributes to each PCA component.

Why it matters: Helps you label the components meaningfully.

How to use it:

- “High loadings on component 1 were seen in questions about gender norms, suggesting this component represents traditional beliefs.”

---

13. `pca_variance_ratio.csv`

What it contains: The percentage of variance explained by each PCA component.

Why it matters: Tells you how useful each component is.

How to use it:

- “Together, the first 3 components explained over 60% of all variability in responses.”

---

14. `clusters_kmeans_k3.csv`

What it contains: Cluster assignments for each respondent using KMeans with 3 groups.

Why it matters: Groups people with similar answers.

How to use it:

- Describe each group: “Cluster 1 consists of younger respondents with high awareness; Cluster 2 shows lower support for reforms.”

---

15. `hierarchical_clusters_k2.csv`, `k3.csv`, `k4.csv`

What it contains: Similar to above, but using a different clustering technique (Hierarchical).

Why it matters: Gives a different view on natural groupings in the data.

How to use it:

- Helps in understanding alternative groupings for segmentation.

---

16. `kmeans_silhouette.csv`

What it contains: Silhouette scores for cluster numbers, which help choose the best number of clusters.

Why it matters: Helps justify how many groups you divided people into.

How to use it:

- “Silhouette analysis suggested that 3 clusters best separated the data, maximizing group distinctiveness.”