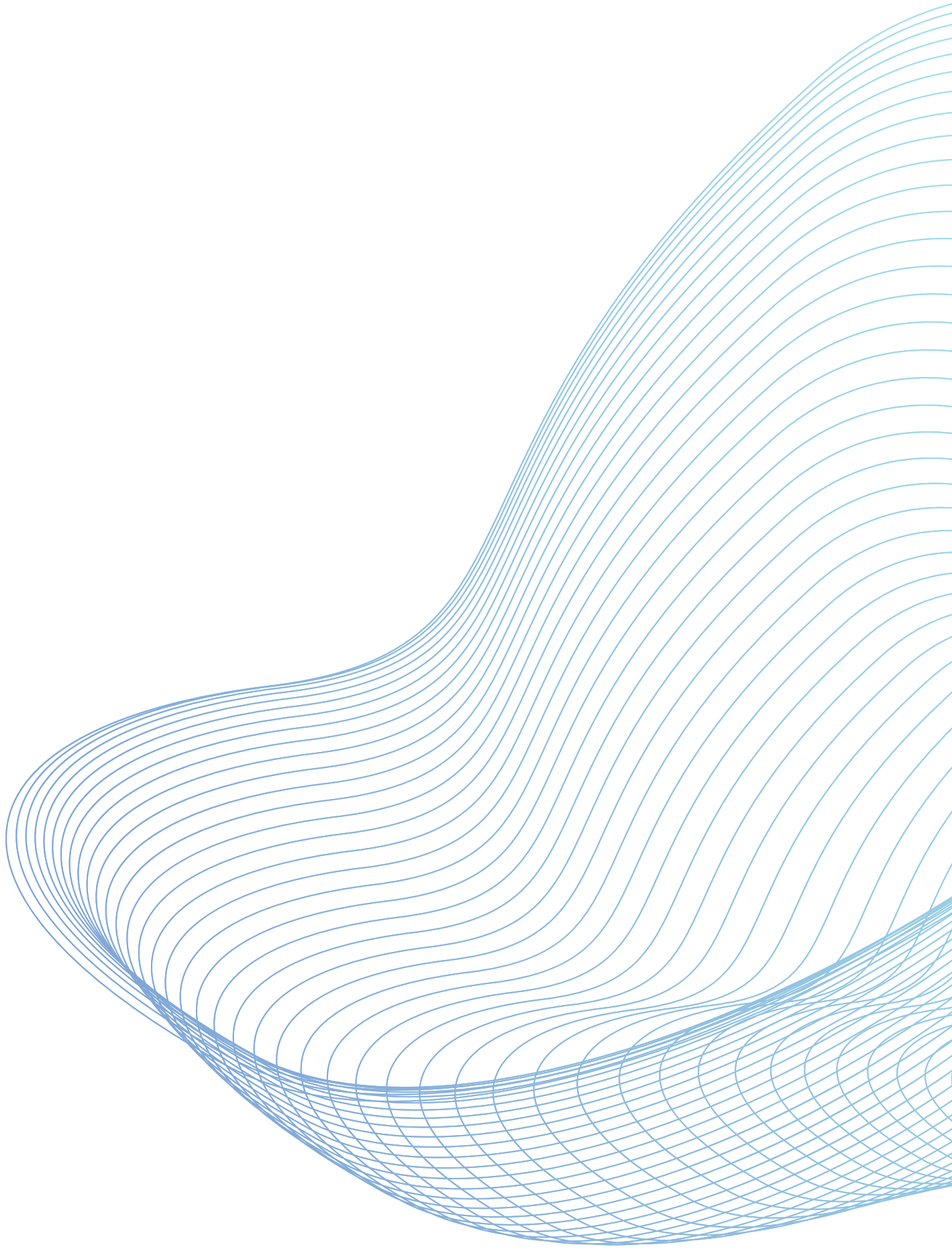
EXAMINING HYPOTHESIS TESTING IN HEALTHCARE

OVERVIEW

Pharmaceutical drugs are essential to our health, so it's crucial they are safe and have minimal adverse effects. Recently, a pharmaceutical company, GlobalXYZ, completed a randomized controlled drug trial. The main interest is in the drug's adverse reactions and whether these reactions are significant.

The project is producing independent reports on various drugs, separate from the manufacturers. Using Python, conduct several hypothesis tests to assess whether a hypothetical drug has significant adverse reactions and examine whether factors like age play a role in these reactions.



PROBLEM STATEMENT

This project aims to assess a hypothetical drug's safety using a randomized controlled trial dataset. Specifically, the researchers want to determine:

- If there is a significant difference in the proportion of adverse effects between the Drug and Placebo groups.
- Whether the number of adverse effects experienced by individuals is independent of the treatment groups.
- If there's a significant difference in age between the Drug and Placebo groups.

GOALS

01

Analyze a drug safety dataset to investigate potential adverse effects.

02

Employ statistical tests to compare the proportions of adverse effects between treatment groups.

03

Evaluate the relationship between the number of adverse effects and treatment groups.

04

Assess the age distribution between the Drug and Placebo groups.

DATASET AND SOURCE

The dataset used in this project is a modified version of the **`drug_safety.csv`** dataset, courtesy of Vanderbilt University's Department of Biostatistics, and includes data on five adverse effects: headache, abdominal pain, dyspepsia, upper respiratory infection, and chronic obstructive airway disease (COAD), along with demographic data, vital signs, and lab measures. The data has a drug-to-placebo observation ratio of 2:1.
The original dataset can be found here: https://hbiostat.org/data/repo/safety.rda

The modified dataset includes the following columns:

- **sex:** The gender of the individual
- **age:** The age of the individual
- **week:** The week of the drug testing
- **trx:** The treatment (Drug) and control (Placebo) groups
- **wbc:** The count of white blood cells
- **rbc:** The count of red blood cells
- **adverse_effects:** The presence of at least a single adverse effect (Yes/No)
- **num_effects:** The number of adverse effects experienced by a single individual

OUTPUTS

1. Determine if there's a significant difference in the proportion of adverse effects between the Drug and Placebo groups, saving the p-value in a variable named **`two_sample_p_value`**.

2. Next, assess whether the number of adverse effects is independent of the treatment (Drug) and control (Placebo) groups, storing the resulting p-value in a variable called **`num_effects_p_value`**.

3. Finally, check if there's a significant age difference between the Drug and Placebo groups, and save the p-value from this test in a variable named **`age_group_effects_p_value`**.

DATA FINDINGS

The code performs several hypothesis tests and stores the resulting p-values in designated variables:

1. **two_sample_p_value:** This variable stores the p-value from a two-proportion z-test, which assesses if the proportion of adverse effects differs significantly between the Drug and Placebo groups.

2. **num_effects_p_value:** This variable stores the p-value from a chi-square test of independence, which determines if the number of adverse effects is independent of the treatment groups.

3. **age_group_effects_p_value:** This variable stores the p-value from a Mann-Whitney U test, which evaluates if there's a significant difference in age between the Drug and Placebo groups (due to non-normal data distribution).

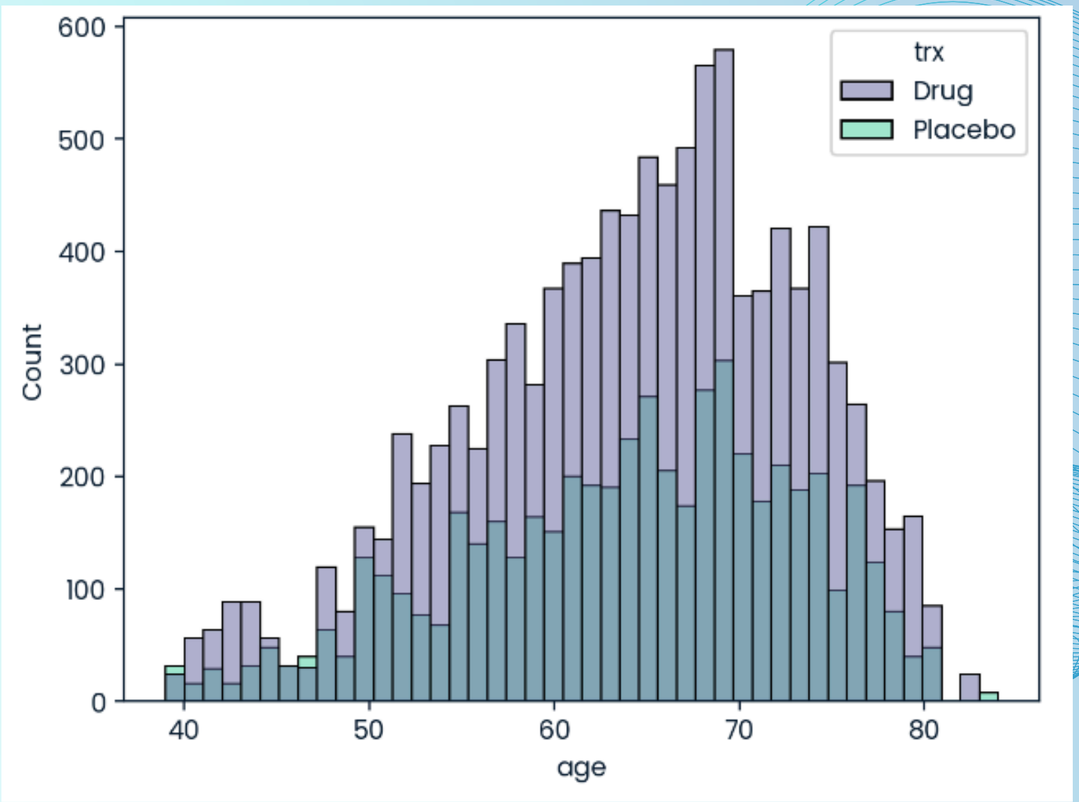
The histogram visualizes the distribution of ages within the Drug and Placebo groups. It helps us understand whether the two groups are comparable in terms of age distribution.

Key Observations:

- **Similar Distribution:** Both groups exhibit a similar distribution pattern, with a peak around 60-70 years old.
- **Overlap:** The distributions of the two groups overlap significantly, indicating that there are no major differences in age between the Drug and Placebo groups.

Significance:

- **Representative Sample:** The similar age distribution suggests that the study has a representative sample of the target population.
- **Reduced Bias:** A balanced age distribution helps minimize potential biases in the study results.
- **Valid Comparison:** The similarity in age distribution allows for a more valid comparison of adverse effects between the two groups.



In conclusion, the histogram indicates that the age distribution of the Drug and Placebo groups is comparable, which strengthens the validity of the study's findings.

INSIGHTS

1. By comparing the proportions of adverse effects and the number of effects between the Drug and Placebo groups, researchers can gain insights into the drug's safety profile.

2. Investigating the relationship between adverse effects and treatment groups can help identify potential risk factors.

3. Understanding the age distribution of participants in each group is crucial for ensuring the generalizability of the findings.