Notes

Chi-Square

* The Chi-Square Test is a statistical test used to determine if there is a significant relationship between two categorical variables.
* It helps us understand whether the differences between groups in our data are due to chance or a real effect.

When Do We Use the Chi-Square Test?

* The Chi-square Test is used when:
  + We have two categorical variables (e.g., Gender vs. Product Preference, A/B Testing Results).
  + We want to check if the categories are independent or if there’s a relationship.
  + Our data is presented in a contingency table (a table showing frequency counts).

Example

* Imagine we run an A/B to check if a website button color (Blue vs. Red) affects the purchase behavior (Bought vs. Not Bought).

|  |  |  |  |
| --- | --- | --- | --- |
| Button Color | Bought | Not Bought | Total |
| Blue | 50 | 950 | 1000 |
| Red | 80 | 920 | 1000 |

What we want to know:

* Is the button color is affecting the buying behavior, or is this difference just random?
* To find out, we can use the Chi-Square Test
* How it works:
  + Create a Contingency Table
    - Count how many observations fall into each category.
  + Calculate the Expected Counts
    - If the two groups were truly independent, how many people should be in each category?
  + Compute the Chi-Squared Statistics
    - Compare the observed values (actual data) with the expected values.
  + Find the P-value
    - If p < 0.05, the difference is statistically significant, meaning the button color affects purchases.

Example:

* Chi-Square test for Stock Prediction Model Performance
  + Background
    - A financial technology firm is developing AI models to predict stock prices.
    - Currently, they have a traditional machine learning model (Model A) that relies on historical stock prices.
    - They are now testing a Generative AI-enhanced model (Model B), which incorporates financial news sentiment analysis in addition to historical price data.
    - The company wants to determine if the new Generative AI Model significantly improves prediction quality compared to the traditional model.
  + Objective
    - To compare the prediction accuracy of the two models using a Chi-Square Testa and determine if the difference in performance is statistically significant.
  + Solution Approach
    - Define Categories for Model Performance
      * Since stock prediction errors are continuous values, we categorize predictions into two groups:
      * Good Prediction 🡪 If the Mean Absolute Error (MAE) is below a threshold (e.g., 2.0)
      * Bad Prediction 🡪 If the MAE is above 2.0
      * This allows us to analyze model performance using categorical data, which is required for a Chi-Square Test.
    - Data Collection
      * The firm runs both models on 50 different stocks over a 30-day period, collecting daily predictions.
      * For each stock, we calculate the average MAE over 30 days and classify the prediction quality as either Good or Bad.
      * A contingency table is then created to compare the performance of both models:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Good Predictions | Bad Predictions | Total |
| A (ML Only) | X1 | Y1 | 50 |
| B (GenAI Enhanced) | X2 | Y2 | 50 |

* + - Chi-Square Test Implementation
      * The Chi-Square Test implementation is applied to check if the distribution of Good vs. Bad Prediction is significantly different between the two models.
      * Null Hypothesis (H0): There is no significant difference in prediction accuracy between Model A and Model B.
      * Alternate Hypothesis (H1): There is significant difference in prediction accuracy between Model A and Model B.
      * A low p-value would indicate that Model B provides significantly more Good Predictions than Model A, proving its effectiveness.
    - Results & Interpretation
      * The test results show:
        + Chi-Square Statistic: 96.04
        + P-value: 1.13e-22
      * Since the p-value is much lower than 0.05, we reject the Null Hypothesis.
    - Conclusion
      * The Generative AI Model significantly improves stock prediction accuracy, providing more Good Predictions than the traditional ML Model.