42.**Statistical Analysis of Gene Variations in Protein Coding Genes**

**Introduction**

This study investigates the impact of gene variations on the outcomes of protein-coding genes. We used a dataset containing several genes, including their molecular category, variation type, and outcome. The data was analyzed using various statistical techniques, including Logistic Discriminant Analysis (LDA), to predict the outcome based on gene variations and associated scores.

**Data Overview**

The dataset contains seven columns: ID, Symbol, Category, Score, Molecular, Variation, and Outcome. The genes analyzed include RP2, DHDDS, CNGB1, and NR2E3, among others. The variations in these genes were categorized as "Pathogenic," "Likely Pathogenic," or "Causative Variation," with corresponding outcomes of 0 or 1 indicating the presence or absence of a particular genetic condition.

The distribution of the data is summarized below:

| **Symbol** | **Score** | **Variation** | **Outcome** |
| --- | --- | --- | --- |
| RP2 | 1234.34 | Pathogenic | 1 |
| DHDDS | 1220.29 | Likely Pathogenic | 1 |
| CNGB1 | 1298.34 | Likely Pathogenic | 1 |
| NR2E3 | 962.55 | Likely Pathogenic | 0 |
| RP2 | 1165.06 | Causative Variation | 0 |
| CNGB1 | 922.23 | Likely Pathogenic | 1 |

**Linear Discriminant Analysis (LDA)**

The Linear Discriminant Analysis (LDA) was conducted to determine the discriminative power of different gene symbols and variation types in predicting the outcomes. The LDA model used the Score, Variation, and Symbol variables to predict the outcome.

**LDA Results**

The LDA results indicate the following coefficients of linear discriminants (LD1):

* **Score**: 0.000317996
* **Variation Likely Pathogenic**: -0.693145228
* **Variation Pathogenic**: -0.687142135
* **Symbol CRX**: -0.390568055
* **Symbol DHDDS**: 0.935548246
* **Symbol EYS**: 0.893754819
* **Symbol IMPDH1**: 1.410984956
* **Symbol NR2E3**: 2.665448129
* **Symbol PRPF8**: -0.965472985
* **Symbol PRPH2**: -0.957685791
* **Symbol RP1**: 0.093748434
* **Symbol RP2**: 1.095878434
* **Symbol RPGR**: 1.776547645

These coefficients suggest that symbols such as NR2E3, RP2, and RPGR have higher positive discriminant values, indicating a stronger association with positive outcomes (Outcome = 1). Conversely, symbols like CRX, PRPF8, and PRPH2 have negative coefficients, indicating a negative association with positive outcomes.

**LDA Predictions by Score**

The LDA predictions by score are visualized in the density plot, which demonstrates the predicted classes (0 or 1) based on the gene score. The plot shows that the predicted density for class 1 is generally higher across all scores, indicating that genes with higher scores are more likely to be classified as having a pathogenic or likely pathogenic outcome.

**Interpretation of Results**

1. **Gene Variations and Outcomes**:
   * The results indicate that genes with certain variations, particularly "Pathogenic" and "Likely Pathogenic," have a higher likelihood of a positive outcome (Outcome = 1).
   * Genes like RP2 and CNGB1 are consistently associated with positive outcomes, regardless of their variation type, suggesting a strong predisposition to pathogenic effects.
2. **Score Impact**:
   * The scores significantly contribute to predicting outcomes, with higher scores generally correlating with a positive outcome. The linear discriminant analysis confirms that Score has a positive coefficient, implying a direct relationship between gene score and positive outcomes.
3. **Variation Type Significance**:
   * "Likely Pathogenic" and "Pathogenic" variations have significant negative coefficients, indicating a substantial impact on the prediction of outcomes, supporting the hypothesis that these variations are crucial in determining the pathogenicity of genes.

**Conclusion**

This study demonstrates that gene variations, particularly those classified as "Likely Pathogenic" or "Pathogenic," significantly influence the predicted outcomes of protein-coding genes. The findings highlight the importance of specific gene symbols and scores in determining the likelihood of pathogenic outcomes. Further research is needed to validate these findings in larger, more diverse datasets and to explore potential therapeutic interventions for genes identified as highly pathogenic.

By employing Linear Discriminant Analysis, we successfully identified key predictors and their contributions to outcomes, providing valuable insights for genetic research and clinical applications.

**Future Work**

Future research should focus on expanding the dataset to include a more comprehensive range of genes and variation types. Moreover, integrating additional covariates such as demographic information and clinical history may improve model performance and provide a more holistic understanding of genetic variations' impact on health outcomes.

This research paper contributes to the growing body of knowledge in the field of genetic epidemiology and the role of data science in advancing personalized medicine and precision healthcare.